NORGES HANDELSHØYSKOLE
Bergen, Spring 2012

Are Individual Forecasters Rational?

A study of inflation expectations using forecasts from the Survey of Professional Forecasters

Karen Oftedal Eikill

Advisor: Krisztina Molnár

Master thesis: Major in Financial Economics

NORGES HANDELSHØYSKOLE

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

How is the forecast behaviour of professional individuals? Are they accurate and efficient, and how are their performances compared to the consensus' performance? Do their forecasts differ in the special episodes of the Volcker disinflation and in the recent financial crisis? And are individuals employed in certain industries outperforming individuals employed in other industries? This thesis examines these issues, using survey data of the one-year ahead inflation rate in the United States, derived from the Survey of Professional Forecasters.

Several aspects of the forecasting behaviour of individuals are highlighted. The consensus mean and median forecasts and most individuals are unbiased. They also pass some efficiency tests, even though they are not strong-form rational. The performance of consensus forecasts is better than the performance of the majority of individuals, though several individuals make accurate forecasts. Even though individual differences exist, there are few differences between the forecasters employed in different industry categories. The forecasters performed were worse during the Volcker disinflation, though not as bad as we might expect. And during the recent financial crisis, the performances of forecasters have not worsened. Additionally, the forecasts seem to have improved over time.

Preface

This paper is the final thesis of my master degree in Financial Economics at the Norwegian School of Economics. It is written as a part of the "macrogroup" in the research program "Crisis, restructuring and growth" (the "KOV- project") at NHH. I am grateful for the opportunity to be a part of this program, and I would like to thank both professors and students involved in the group for helpful comments and guidance.

The inflation is very important for the economy. For central banks conducting monetary policy, it is decisive to control the inflation and therefore to have knowledge of individuals' inflation expectations. Inflation expectations have been debated and studied a lot in modern macroeconomics. Different results are found, thus no final conclusions are drawn. With this thesis I wanted to contribute to the existing literature by looking at professional survey respondents' forecasts in a detailed manner. The dataset I have been using contains individual data, and enables me to look at the forecasts on an individual level. However, the numbers of individuals have been limited, and there are several problems with the dataset that needs consideration. In addition several findings are explained by intuition, reasoning and comparison with previous results. Hence, the findings should not be viewed as hard evidence.

The process of writing this thesis has been a learning and at times a challenging experience, which I could not have done without help and guidance by certain peopled. I would especially like to thank my advisor, Krisztina Molnár, for valuable advice and help during the writing of this thesis. I would also like to thank my family and my friends Øyvind and Katrine for helpful notes and comments.

All errors and inconsistencies in this thesis are my own responsibility.

Norges Handelshøyskole

Bergen, June 18th, 2012

_____

Karen Oftedal Eikill

Contents

1. Introduction

Inflation expectations are debated and studied a lot in modern macroeconomics. Many economic agents base their real decisions on inflation expectations. Hence, their expectations are important for the economy. Among those are policymakers conducting fiscal policy, firms setting prices and management and labour negotiating on wages. For central banks, the control of inflation is decisive in their goal of pursuing good monetary policy. Because inflation expectations influence the actual inflation they also influence the conduction of monetary policy performed by the central bank (Bernanke, 2007). Macroeconomic models also emphasize inflation expectations and argue that they are crucial. Forecasts can provide important information about inflation expectations, and have in a comprehensive study by Ang et al. (2007) been found to forecast the inflation better than other possible methods. Almost all central banks with inflation targeting study and evaluate surveys with inflation expectations (Kershoff & Smit, 2002). Hence, such surveys are considered valuable and are naturally often studied and examined.

Many macroeconomic models assume that the rational expectations hypothesis holds (Mankiw, et al., 2003). The hypothesis has been an object of a lot of studies, and different conclusions have been drawn. Because the monetary policy implications of rational expectations are very different from the implications of other, more backward-looking models, studies of the hypothesis continue. In this thesis we examine the forecast behaviour of professional forecasters, investigating if they are accurate and rational. Using the Survey of Professional Forecasters (SPF), we study the one-year ahead inflation expectations of individual respondents. Even though examining the rationality of the forecasts in the SPF has been performed by previous studies, relatively few have examined rationality on an individual level. To truly understand the nature of forecasters it is important to look at how individuals perform and whether there are differences between them. Because most previous literature and economic models do not account for individual differences, we find analysing the subject both interesting and valuable. Together with the fact that our data sample is new, containing forecasts of the recent financial crisis, our detailed discussion of the rationality and accuracy of individuals is a contribution to the existing literature.

We also add to the literature an analysis of the industry variable containing in the survey. We compare the industries to find if differences exist. In addition we examine the effects of the

Volcker disinflation and whether the forecast performance of individuals has altered in the recent financial crisis. Our paper also documents problems with the SPF. No previous papers have, to our knowledge, examined all these problems.

The questions we want to answer in this paper are thus; how rational are individual forecasters? How do they perform compared to consensus forecasts and do we observe any patterns among them? Do the employment of individuals matter for their forecast performance? And have the rationality of individuals been affected by the Volcker disinflation and the financial crisis?

When analysing the whole sample, we find that the accuracy and the rationality of individuals vary a lot. Both the consensus and the majority of individuals are unbiased. With the majority of individuals passing less tests of efficiency than the consensus, the performance of individuals can be claimed worse than the consensus. But even though the majority are "less" rational than the consensus, there are many individuals whose performances are relatively good. Examining the rationality of individuals employed in different industries leaves us with no particular distinctions. A strategic incentive of for example media attention, is, however, more likely to exist among the individuals employed in the nonfinancial sector.

Results regarding the Volcker disinflation indicate quite accurate forecasts, even in this decreasing inflation period. Even though the majority of individuals are biased, there are many individuals for whom we cannot claim biasedness. A quite surprising result also emerges when we analyse the rationality of forecasters during the recent financial crisis. Both consensus forecasts and individuals performed better during the financial crisis than in the whole sample. Even when we compare with a more recent sample starting in the second quarter of 1990, this result holds.

Hence, the individuals are quite accurate, but not strong-form rational. This holds for almost all tests performed with our data. The forecasts also seem to improve over the surveyed years. Several results point in this direction; the best respondents are located in the end of the survey period, the sample that starts in the second quarter of 1990 performs a bit better than the whole sample and the forecasts made during the recent financial crisis are relatively good. Both these results are in accordance with previous literature (Croushore, 2006; Gerberding, 2006).

In the following, we start in section two with a presentation of what the inflation is and the importance of it, together with some theory about expectations. A presentation of both the survey data and the actual data follows in section three. Section four contains theory about how to examine and test accuracy and rationality of forecasts, and section five presents our dataset together with some problems that we had to deal with (and the chosen solution for those). Section six presents our analysis part. When analysing, we start examining the whole sample, before analysing the industry variables, the Volcker disinflation and the financial crisis.

2. Inflation expectations

The annual inflation is the yearly increase in the price level in an economy. The inflation makes money less worth, thus decreasing the purchasing power. If the inflation is negative the price level decreases and deflation is present.

The inflation is very important for the economy. Decreasing the value of money, it makes the value of wages and the value of loans smaller. The inflation is important for both the rulers of a country and its inhabitants. Keeping the price level stable, thus having a low inflation rate over time, will promote growth, efficiency and stability. This will, all else equal, support a maximum sustainable employment. For central banks conducting monetary policy, controlling the inflation is decisive (Bernanke, 2007). Several countries have inflation targeting as their monetary regime (Bernanke & Mishkin, 1997). The main goal of the monetary policy of the central banks in these countries is to keep the inflation stable.

We begin presenting inflation expectations and a presentation of expectations theory in section 2.1. In 2.2 we present and discuss briefly some previous literature that discusses inflation expectations.

2.1 Theory and importance of inflation expectations

The aim of this section is to present inflation expectations. Theory about expectations in general is presented in 2.2.1. We continue debating the importance of inflation expectations in section 2.2.2, before discussing how to measure them in 2.2.3.

2.1.1 Expectations theory

How expectations are formed is very important. The most popular theory about the formation of expectations is probably the rationality expectations hypothesis, with a popular alternative being adaptive expectations (Mankiw, et al., 2003). If expectations about the inflation are formed adaptively, one expects the next year inflation to be equal to the inflation over the past year (Mankiw, et al., 2003). If true, the expected inflation would contain no new information. Making an effort to gather those would then be a waste of time. However, according to several previous studies, for example the mentioned by Ang et al. (2007), expectations can provide valuable new information (others who claim this are Thomas (1999) and Gerberding (2006)). Thus, the backward-looking hypothesis of adaptive expectations fails, and finding a way to measure and trace the expectations is desirable.

The rational expectations hypothesis assumes that a sufficiently large number of people know "how the world works", making rational predictions based on the information they have available at any time (Zarnowitz, 1992). As defined by Muth (1996 cited in Gerberding, 2006, p.316); "Expectations, since they are informed predictions of future events, are essentially the same as the predictions of relevant economic theory." Hence, Muth assumes that the subjective expectations of economic agents match the predictions of the relevant economic theory, and therefore do not make systematic mistakes (Gerberding, 2006).

If expectations are rational they should be both unbiased and efficient. If unbiased, forecast errors are zero on average, and if efficient individuals use all relevant information when they form their expectations. To exploit this information, individuals have to do a lot of research and they have to keep updated on previous values of the economic variable that they are going to forecast (Gerberding, 2006).

Efficiency could be both weak-form and strong-form (Thomas, 1999). Weak-form efficiency requires that individuals adequately consider information they have in past values of the variables they are forecasting. This criterion is based on the notion that while historical information about the variable itself can be viewed as costless, other information is costly. Therefore, individuals cannot be required to account for all other information.

If individuals are strong-form efficient they exploit all information available where the marginal benefit exceeds the marginal cost of gathering, learning and utilizing this information when they predict the inflation (Thomas, 1999). Because different individuals have different marginal costs and benefits, defining the exact level of available information that individuals should utilize to be defined as strong-form efficient is difficult.[1]

Because the implications for the conduction of monetary policy are different if expectations are formed rational compared to adaptive, it is of importance for politicians and central banks to study how expectations are formed (Bullard & Mitra, 2002). Studies find that observed inflation expectations are not consistent with either adaptive or rational expectations

[1] It could also be questioned if the criterion of strong-form rationality, if expressed as individuals exploring all available information, is too strict (Gerberding, 2006). This because the amount of knowledge required is large and it is time-consuming to keep updated. However, if one considers the marginal cost of utilizing the information smaller than the marginal benefit, one should demand individuals to update themselves on this type of information.

(Roberts, 1998; Mankiw, et al., 2003). To know how individuals form their expectations is difficult, and further research on this topic is thus important.

2.1.2 The importance of inflation expectations

Inflation expectations are important for those who make decisions about the future. Policymakers conducting fiscal policy, firms setting prices and making decisions about investments, investors who are hedging the risk of nominal assets, management and labour negotiating on wages and central banks and politicians who are conducting monetary policy, all base their decision on their expectations about future inflation (Ang, et al., 2007). Because they affect real agents' decisions, the inflation expectations have a true effect on the real economy. Many macroeconomic models involving the inflation emphasize inflation expectations and argue that they are crucial (Mankiw, et al., 2003). Thus inflation expectations are important also for economic research. Several OECD countries base their monetary policy on inflation targeting. For those the inflation expectations are especially important (Diebold, et al., 1997; Thomas, 1999).[2] Naturally, the important inflation expectations have been an object of many studies (Gerberding, 2006).

Changing inflation expectations and the factors that create these changes are also important. If an increase in the inflation is expected, decision makers will change their behaviour. Workers will demand higher wages, and central banks will change their monetary policy by setting a higher rate to try to lower the inflation, given that the new expected inflation is higher than their "targeted value." New information often changes the inflation expectations of economic agents. Hence, newly published values for macroeconomic variables will be important, because agents will adjust their forecasts if the new values differ from the expected ones.

2.1.3 Measuring inflation expectations

Expectations are variables that cannot be observed. Different approaches of finding a proxy for these variables exist. It is possible to build economic models, derive measures from financial asset prices or to use time series models (Ang, et al., 2007). Another alternative is to conduct surveys. Surveys question market participants directly about their expectations of the

[2] In the United States, one of the duties of the Federal Reserve is to conduct the country's monetary policy. This is done in a pursuit of maximum employment, stable prices and moderate long-term interest rates. In a press release from January 25, 2012, the FOMC states that they judge an inflation rate of two percent to be most consistent over the longer run. Hence, they communicate an inflation goal to anchor inflation expectations, meaning that also the United States has some degree of inflation targeting (Board of Governors of the Federal Reserve System, 2012b).

desired variable over a certain time horizon (Gerberding, 2006). An advantage of surveys is that they do not depend on other assumptions, for example how the level and structure of ex ante interest rates are. If depending on other assumptions, the forecasted variable can never be better than the theory and assumptions they rely on. A comprehensive study by Ang et al. (2007) finds that surveys forecast inflation better than the other measure they consider.

Many economists have used survey data to test hypothesis about the formation of inflation expectations (Keane & Runkle, 1990). Survey participants form their expectations and report those in the survey questionnaires. Almost all central banks that have inflation targeting, study inflation expectations surveys (Kershoff & Smit, 2002). They use the surveys to forecast the inflation and to evaluate the credibility of policies that involves inflation.

2.2 Previous literature

This section briefly presents some of the previous literature that tests survey data against actual data. Some of these studies will be mentioned in more detail in the analysis section, when we compare our results with previous results.

Victor Zarnowitz has done some extensive work in terms of examining the Survey of Professional Forecasts, which is the survey data we will be using (a presentation of this survey is presented in 3.1.2).[3] In a study of rational expectations, Zarnowitz found that the null hypothesis of unbiasedness is rejected for inflation forecasts when using OLS regression estimates. However, the error terms were serially correlated, which could lead to falsely rejecting the null (Zarnowitz, 1985). He also found that the "consensus,"[4] was on average more accurate than most of the individual respondents' predictions over time (Zarnowitz, 1992). Together with Braun, Zarnowitz made a very comprehensive study of the survey in 1993, analysing a lot of the surveys' variables. Some of the results they found were, again, that the consensus forecasts are better than most individual forecasts in terms of average errors, and that the survey performs well when comparing it with other econometric and time series models (Zarnowitz & Braun, 1993).

[3] Zarnowitz was also involved in tabulating, analysing and evaluating the results when it was conducted by the ASA/NBER, and he has done a lot of research studying the survey (Croushore, 1993).
[4] Finding the consensus is done by averaging all predictions in a survey for a given variable and time period, resulting in a time series of group mean forecasts. This could also be done with the median (Zarnowitz and Braun, 1993). From now on we will refer to these as mean and median consensus forecasts.

Other studies also examine some of the issues Zarnowitz´ looked at. Among those is the mentioned paper by Ang et al. (2007). Croushore have also written several articles examining survey forecasts (Croushore, 1993; Croushore, 2006). While many studies use consensus data, Keane and Runkle (1990) tested whether or not the individual forecasters in the SPF were rational or not. They concluded that the forecasts were consistent with rational expectations.[5]

Studies examining the rationality of survey forecasts in the United States often use the SPF, the Livingston Survey of professional economists and the Michigan survey of households (Thomas, 1999; Mankiw, et al., 2003; Ang, et al., 2007). When examining and comparing accuracy measures Ang et al. (2007) find that surveys outperform other prediction models, with the SPF and the Livingston survey performing very well, and better than the Michigan survey.

A recent study examining some of the rationality issues are Mankiw et al. (2003). They argue that individuals are different, creating disagreement between the forecasters when predicting the inflation. This disagreement is something most economic models and research do not account for. Instead rationality of survey forecasts is often assumed.

[5] The mentioned study by Zarnowitz and Braun (1993) also discuss the forecast performance of individuals.

3. Choosing data

To answer the fundamental questions in this thesis, we have to analyse data. This section presents the data that we will be using. We need values for inflation expectations, actual values of the inflation, and we need data about variables that we will be using in the analysis part. We start finding forecasted values suitable for our analysis in section 3.1, before turning to the actual values of the inflation in section 3.2. In 3.2 we also present actual data of the economic variables that we are going to use in our analysis.

3.1 Forecasted values - The Survey of Professional Forecasters

Because inflation expectations are found to forecast the inflation better than many other methods of measuring, we choose to evaluate inflation expectations by using survey measures (Ang, et al., 2007). Important for our choice is the fact that surveys do not rely on other assumptions, as many other alternative measures do (Gerberding, 2006).

We will be using data from the Survey of Professional Forecasters (SPF). This survey has been conducted since the fourth quarter of 1968. The Federal Reserve Bank of Philadelphia has been providing the survey from the second quarter of 1990. Before this the responsibility of the survey was shared between the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER) (Federal Reserve Bank of Philadelphia, 2008).

There are a lot of economic variables included in this survey. Examples are employment and unemployment forecasts, inflation forecasts and production forecasts. Our focus will be on the inflation measure. The forecasted inflation is calculated using forecasted levels of pgdp, which is the level of the GDP price index (how this is done is presented in section 5.1). Even though we have survey responses from the fourth quarter of 1968, the survey has only been collecting the levels of the GDP price index since 1996. From 1968 to 1991, the forecasts were of the GNP deflator and between 1992 and 1995 the GDP implicit deflator.[6] Because these behave quite similar, and there does not seem to be any breaks in the inflation series

[6] The GDP price index is the change in the relative price on a fixed basket of goods produced (Statistics Norway, 2012). The GDP deflator is not based on a fixed basked of goods and services; it is the nominal GDP divided by the real GDP times 100, and vice versa for the GNP deflator (Bureau of Economic Analysis, 2011a). While the GDP contains the goods produced domestically from year to year, the GNP focuses on the produced goods that are owned by the respective country.

generated from the three different measures in the years where the measure was changed, we will not problematize this further (Diebold, et al., 1997; Croushore, 2006).

The survey contains individual pgdp data, which makes it possible to analyse the questions on an individual level in addition to an aggregate level. The survey also contains both point forecasts as well as probability distribution forecasts of this variable. In a probability forecasts, the respondents answers to the probability of the inflation falling into different categories the next periods. Due to the time limit of this paper, we have chosen to focus on the point forecasts, but doing the same analysis and checking whether or not the probability distributions gives conclusions consistent with the point forecasts could be an interesting topic for a further research.[7]

Even though survey measures are considered to be good forecasts, there are several issues to keep in mind when examining surveys. We will first discuss the use of individual or consensus forecasts, section 3.1.1, and why we have chosen a professional survey in section 3.1.2. A more elaborate explanation of the data, how to transform the survey data into a comparable measure as well as a discussion of some problems with the dataset that we had to handle follow in section 5.

3.1.1 Analysing on an individual or a consensus level

Because the SPF contains individual data it is possible to analyse and perform tests on both an individual and an aggregated level. Studies vary regarding their approach to this issue, and they have different arguments regarding the level they choose to focus on. Our main focus is to study forecasts on an individual level, a focus relatively few papers have had before. Because we also compare these with the consensus forecasts, the consensus is also analysed.

Several studies argue that it is better to use consensus forecasts of the individual data. The reasoning is that individual forecasts can be biased because of behavioural biases (Batchelor & Dua, 1995).[8] These biases can be eliminated, or offset when aggregating forecasts from several forecasters, for example by using the mean or the median. As mentioned, Zarnowitz

[7] There are already several papers studying the probability forecasts in the SPF, examples are Clements (2008b; c) and Diebold et al. (2008).

[8] A behavioural bias is when someone behaves irrationally, for example using behavioural heuristics to make choices that leads to sub-optimal investment choices (Goetzmann & Massa, 2003). In forecasting a thought example could be if one forecaster is more optimistic than others, hence his heuristics will lead to more optimistic forecasts then the others.

(1984) found that on average mean forecasts will perform better and be more accurate than those of individual forecasters.

Individual data can also contain important information. One example is that some individuals may be much better than others in terms of forecasting behaviour. To use only consensus forecasts can also involve some problems (Keane & Runkle, 1990). One example is that the different information sets that individuals have are not accounted for. This can cause consensus forecasts to have a serious specification bias. The reasoning is that when we test the rationality of forecasters, their forecasts may differ just because they have different information sets.[9] Hence, some might seem to be rational and others not. Not knowing the information set of the individuals it is impossible for us to state who that are truly rational. Another problem is that we will not be able to see individual deviations from rationality when only using consensus values. Individual deviations might be of great importance, for example if one wants to test if a group is rational and rationality results appear just because negative biases hide positive biases. We consider individual data to contain a lot of important information, and hope that studying the individuals on a more detailed level can give us new and valuable information.

3.1.2 The forecasters

The forecasters in the survey are, as the name suggests, professionals. They are largely from the business world; from banks, economic consulting firms, university research centres, other economic firms and from Wall Street (Croushore, 1993).

The forecasters in the SPF are professionals who are close to important economic decision makers (Giordani & Söderlind, 2002). This is believed to be a strength of the survey, because it makes it more likely that the survey reflects the beliefs that are affecting important investment and pricing decisions. In addition the survey has a careful screening of candidates, which is supposed to secure the survey against "nonsense" answers. Being professionals the forecasters use different tools to determine their forecasts. Examples are other people's forecasts, leading indicators and other surveys.

Respondents should have an incentive to report their expectations correctly. Therefore, some argue that the respondents should be those who also sell their forecasts on the market. At the

[9] When finding the mean or median of many individual rational forecasts, each conditional on a private information set, it is not said that the forecast itself will be a rational forecast on any particular information set.

same time, respondents should not have any strategic incentives to not report their true beliefs (Gerberding, 2006), and when respondents also sell their forecasts on the marked, strategic motives might be present. Examples are that they could be afraid to respond their true beliefs, because of a fear of being the only one making a mistake. A strategic motive could also be to make forecasts that do stand out from others' to get media attention and publicity (Laster, et al., 1999). Making the respondents anonymous could solve this problem. At the same time they will not be punished for mistakes nor awarded for good forecasts if they are anonymous.

All the individual forecasters that we will be working with have one confidential identification number each, and are therefore anonymous. Due to the lack of strategic incentives of anonymous forecasts this is often seen as strength to the survey (Giordani & Söderlind, 2002). However, the forecasters of the survey are often the same as those reporting forecasts for the public, implying that strategic incentives could be present. On the positive side, this makes the forecasts to some degree secured.

3.2 The actual values

The main issue of this paper involves comparing forecasts with actual values. We need reliable actual data that corresponds with the forecasted data. In this section we start considering which measure we should use and from which source we should acquire it from, in section 3.2.1. We also present other actual values that we need in our analysis, section 3.2.3.

3.2.1 Source and measure

It is important to use actual data for the same, or a very similar, variable as the one the survey asked for. In our analysis the actual value that we use is the implicit price deflator, the IPD, of the GDP in the United States. The IPD of GDP is the ratio of the current-dollar value of the GDP to its corresponding chained-dollar value, multiplied by 100 (Bureau of Economic Analysis, 2011a). The IPD is at present not the exact same value as the one the SPF participants predicts (which is the level of the gross domestic product (GDP)). It is, however, the measure the survey asked for between 1992 and 1995. At the same time the series of the IPD is very similar to the level of the GDP price index (as discussed in 3.1). Therefore we consider it a good measure to compare the survey data with and we calculate the actual inflation from this IPD of the GDP.

We use the IPD collected from the Bureau of Economic Analysis (BEA).[10] The BEA is an agency of the Department of Commerce in the United States, and is a part of the Department's Economics and Statistics Administration (Bureau of Economic Analysis, 2011b). They are one of the world's leading statistical agencies, producing a lot of economic accounts statistics that helps to promote a better understanding of the United States economy for different agents and decision makers, such as the government and the public. Their vision is to be the world's most respected producer of economic accounts, and they should therefore be a very reliable source. Some of their produced statistics are of the most closely watched economic statistics, such as the national income and product accounts (NIPAs). "Our" measure, estimates of the GDP is a very important NIPA variable.

When comparing survey data and actual data it is important to choose between revised or vintage actual data. The fully revised data is the newest value of the variable in question. If choosing vintage data, there are different sets to choose from, being the first one published or others published sometime after the first publications. Previous literature has discussed whether to use revised or vintage data with different conclusions (Keane & Runkle, 1990; Croushore & Stark, 1999; Zarnowitz & Braun, 1993). The most common choice in forecasting literature is to analyse based on the latest variables, thus revised, data (Croushore, 2006). The reasoning behind is that it is the final actual data that the forecasters are trying to predict, not some preliminary data. We emphasize this thought and choose to follow the "mainstream," using revised data as actuals for comparison. However, it is important to consider which values the individuals should have knowledge about when predicting the inflation. A more elaborate discussion regarding the use of revised or vintage data as well as what previous literature state regarding this issue is presented in appendix 2.1.

3.2.2 Economic variables needed for analysis

In our analyses we also need actual data of other variables. These are actual economic variables that we expect the professional forecasts to have accounted for when making their forecasts. Examples of such are the unemployment rate and the short-term interest rate. According to economic theory, there is a relationship between the unemployment rate and the inflation; with a high unemployment rate indicating a low inflation and vice versa. This is expressed by the Philips curve (Gärtner, 2006). The interest rate has a close relationship with the inflation rate as well, especially in countries where the conduction of monetary policy is

[10] All actual values are extracted from the Thomson Reuters Datastream.

based on an inflation target. One example is that a high interest rate today indicates contractionary monetary policy (by the central bank), which can signal a lower inflation in the next period (Mankiw, et al., 2003). We choose to use both the unemployment level and the short-term interest rate in our analysis.

The unemployment rate in the United States is the number of unemployed individuals as a percentage of the labour force. If categorized as an unemployed one have to be in the age of 16-65 and available for work. Additionally, one should not have been working during the survey week, and at the same time have made an effort to find a job within the previous four weeks (Bureau of Labor Statistics, 2012b). We use data from "The Bureau of Labor Statistics" (BLS) of the U.S. Department of Labour. This is a Federal government agency responsible for measuring the labour market activity, working conditions and price changes in the economy, and is thus a reliable source (Bureau of Labor Statistics, 2012a). The unemployment rate is generally subject to only small revisions, which makes it preferable for testing (Mankiw, et al., 2003).

For the short-term interest rate we use the federal funds rate of the United States. The tools that the Federal Reserve controls; the discount rate, the reserve requirements and the open market operations, alter this short-term interest rate. By using these three tools the Federal Reserve influences the demand for, and supply of, balances that depository institutions hold at Federal Reserve Banks. This is what influences and alters the federal funds rate, the interest rate which depository institutions lend balances at the Federal Reserve overnight to other depository institutions (Board of Governors of the Federal Reserve System, 2012a). It is thus a key benchmark for the interest rates in the short-term money market in the United States. The source of the data is Reuters Ecowin. Ecowin gets its data directly from the primary sources, with the most major economic indicators reflected only minutes after they have been released (Thomson Reuters, 2012). The federal funds rate and the unemployment rate in the United States are presented in figure 6.14, section 6.3.1.

4. Evaluating and testing forecasts

It is important to examine differences between survey forecasts and real values. This section presents accuracy measures and tests we can use to investigate such differences. One can examine how accurate the forecasts are by comparing actuals and forecasts using different accuracy measures presented in section 4.1. To find out whether forecasts and actuals differ significantly, hence if forecasts are rational, we can perform tests presented in section 4.2. In the presentation of these measures and tests we talk about actuals and forecasts in general, but in some examples we refer to the inflation forecasts and the SPF specifically.

4.1 Evaluating forecast accuracy

To investigate how accurate and useful a survey is, we examine the forecast accuracy.[11] There are several measures of forecast accuracy that we can use. All the accuracy measures that we present involve a comparison of the mean forecasted errors and the actual values. The forecast error is given by subtracting the forecasted inflation of a period (t), F_t, from the actual inflation in that same period, A_t:

$$Forecast\ error = A_t - F_t$$

We will focus on four different measures presented in different sections; the mean error in section 4.1.1, the mean absolute error in 4.1.2, the root-mean-squared error in 4.1.3 and the mean normalized squared error in section 4.1.4.

4.1.1 The Mean Error (ME)

The first measure is the mean error; the average difference between the actual value and its forecasted values:

$$ME = \frac{\sum_{t=0}^{N}(A_t - F_t)}{N}$$

A_t is the actual values and F_t is the forecasts, N is the number of observations and time is denoted by t. For a forecast to be unbiased, the ME should be close to zero over time. Because the sign of the error is taken into account, a positive error can offset a negative one. A positive value for the bias indicates that on average the actual values has been

[11] See Batchelor (2000), Mankiw et al. (2003).

underestimated and vice versa (Batchelor, 2000). Being the average forecast bias, the ME can be used analysing the unbiasedness of forecasts as well as the forecast accuracy.

4.1.2 Mean absolute error (MAE)

The mean absolute error (MAE) is calculated as:

$$MAE = \frac{\sum_{t=0}^{N} |(A_t - F_t)|}{N}$$

MAE is the average of all forecast errors; the differences between actual values and mean forecasts (Batchelor, 2000). The sign of the error is disregarded, so a negative error does not offset a positive error. MAE is more accurate the closer it gets to zero.

4.1.3 Root-mean-squared error (RMSE)

This statistic is calculated by squaring all the errors, thus disregarding their signs, and then averaging them by dividing on the number of observations, finding the mean squared error (MSE) (Batchelor, 2000). The RMSE is the square root of this MSE:

$$RMSE = \sqrt{\frac{\sum_{t=0}^{N} (A_t - F_t)^2}{N}}$$

This RMSE penalizes forecasters who make a large errors heavily compared to forecasters who make many small errors, thus assuming that the seriousness of an error increases sharply with square of the size of the error.[12] The closer the RMSE gets to zero, the better is the forecast accuracy.

4.1.4 Mean normalized squared error (MNSE)

We want to use an accuracy measure that accounts for the variation in the actual value. If the variation in a variable (the actual value) is large, forecasting can be more difficult than if the dispersion is small. We thus calculate the mean normalized squared error (MNSE):

$$MNSE = \sqrt{\frac{\sum_{t=0}^{N} \frac{(A_t - F_t)^2}{\sigma_p^2}}{N}}$$

[12] An error of for example ±2 % is treated as four times as important as an error of ±1 % in the RMSE. MAE assumes that the seriousness of the errors depends of the size of the errors directly. This means that an error of ±2 % is twice as "serious" as one of ±1 %.

By dividing the squared error by the standard deviation of the actual values in a period p, σ_p^2, we adjust the prediction error for volatility that can be present in the actual values. Also in terms of MNSE the forecasts accuracy is better the closer it gets to zero.

4.2 Rationality tests

When testing the rationality hypothesis, we examine whether the made forecasts exhibit systematic mistakes or not. It is common to divide the tests in two requirements necessary for rationality; unbiasedness, presented in section 4.2.1 and efficiency, presented in section 4.2.2. Bonham and Dacy (1991) present a hierarchy of rationality tests. "Weak" rationality implies that forecasts are unbiased and meet tests of week-form efficiency. "Strong" rationality demands the forecasts to be weekly rational, in addition to the forecast error being uncorrelated with any variable in the respondents information set available at the time of the predictions (Bonham & Dacy, 1991; Stekler, 2002).

4.2.1 Test of bias

When testing for bias, we find whether the survey respondents' forecasted values are correct on average. This implies testing if the average forecast error is zero. To test this we regress the actual values of a variable at a time, A_t, on a constant, α, and the corresponding forecasts for the same time period, F_t (Stekler, 2002):

$$A_t = \alpha + \beta F_t + \varepsilon_t$$

The test involves testing the joint null hypothesis that $\alpha = 0$ and $\beta = 1$. If the null hypothesis cannot be rejected, we cannot claim the forecasts biased. Even though it is not completely correct statistically to claim them unbiased if we cannot reject the null hypothesis, we will sometimes use the word "unbiased" if this is true.[13]

Holden and Peal (1990) states that even if the null of unbiasedness is rejected using this regression, there is still a possibility of the forecasts being unbiased. Thus, rejecting the null is not sufficient for stating that the forecasts are biased. We can use a test that is both

[13] This goes for the efficiency tests as well. When not rejecting the null of efficiency, we will sometimes say they are efficient even though the most correct thing statistically is to say that we cannot claim theme not efficient. This issue is also highlighted in the analysis section.

necessary and sufficient for unbiasedess. This involves regress the forecast errors on a constant (Stekler, 2002)[14]:

$$A_t - F_t = \alpha + \varepsilon_t,$$

and test if the constant can be restricted to zero with the null hypothesis $\alpha = 0$. We will use this last form of the test, being the one both necessary and sufficient for unbiasedness to hold.

4.2.2 Tests of efficiency

For a forecasters to be rational, his or hers forecast errors must be uncorrelated with the entire information set this forecaster has available when making the predictions. It is hard to define the exact information that these sets should contain. We can, however, test whether or not the forecast errors are correlated with important information that the forecasters should have and utilize when making their forecasts (Stekler, 2002). We use different tests regarding such information. The tests we use are: to add lagged values of the actual value, section 4.2.2.1, to add forecasts, 4.2.2.2, to add lagged forecast errors, 4.2.2.3, and to add the full information set, section 4.2.2.3.[15] Tests presented in 4.2.2.1, 4.2.2.2 and 4.2.2.3 are weak-form efficiency tests, while the test in section 4.2.2.4 is a strong- form efficiency test. For the forecasts to be truly rational, they have to pass the test of unbiasedness discussed in 4.2.1 and these efficiency tests.

4.2.2.1 Efficiency test 1: Adding lagged actual values

One test implying weak-form efficiency if not rejected is to add lagged values of the actual variable as independent variables. If efficient, the coefficients of these should be zero (Lovell, 1986). The thought is that if the forecasts are rational, the prediction errors should be uncorrelated with historical values of the forecasted value. We add the lagged inflation, running the regression:

$$A_t - F_t = \alpha + \beta_1 A_{t-4} + \varepsilon_t$$

[14] The actual values will in our paper be the calculated actual inflation in period t, while the forecasted value is the calculated one-year ahead inflation forecast made one year before t.
[15] In all tests we will use the Holden and Peel (1990) version of the tests, by regressing the forecast errors on the abovementioned variables.

Where α and β_1 should not differ significantly from zero if the forecast is rational. The joint null hypothesis is thus α=β_1=0. If the joint null is not rejected the forecast is weakly rational based on this test.

It is common to include the most present realized value of the actuals that is known to the forecasters. But, when looking at quarterly levels, it could be that the realized quarterly values contain some seasonally noise. If the forecasting period of the forecaster is the next year, thus the next four quarters, the actual value that we should include should be the one calculated for the last four quarters, being A_{t-4}. The first report for the quarterly NIPA values is released in the end of the first month in the next quarter. With the first release of the actual inflation, A_{t-4}, being about three quarters ago, the forecasters of the SPF should have knowledge about this actual value (Federal Reserve Bank of Philadelphia, 2008).

4.2.2.2 Efficiency test 2: Adding forecasts

Another weak-form efficiency test is to include forecasts on the right-hand side of the equation to examine if there is information in the forecasts themselves that can predict forecast errors (Mankiw, et al., 2003). We test this by running the regression:

$$A_t - F_t = \alpha + \beta F_t + \varepsilon_t,$$

And test the joint null hypothesis, α=β_1=0. If the joint null is not rejected, the forecasters are efficient and weakly rational.

4.2.2.3 Efficiency test 3: Adding forecast errors

We can also test if forecast errors are persistent or not. We regress the forecast error on the previous year forecast error, to see if information in these previous values has any predictive power for the forecast error. If they do, then the forecast errors are persistent, and the forecasts can improve if knowing the last years' forecast error.[16] We regress the forecast error on the previous year's forecast error:

$$A_t - F_t = \alpha + \beta(A_{t-4} - F_{t-4}) + \varepsilon_t$$

[16] Testing this on an individual level requires that previous forecasts of the same individual are available, and therefore consecutive periods of information for different individuals.

When testing if the error made a year ago is still persistent, we test if autocorrelation exists. If the joint null hypothesis, α=β_1=0, cannot be rejected, we cannot claim the forecasts not efficient, hence the forecasters are efficient. The coefficient, β_1, tells us to which degree the errors made a year ago are still present in today's forecasts.

4.2.2.4 Efficiency test 4: Adding an information set- relevant available information

To test strong-form efficiency we need the information set available to individuals when they make the forecasts. To know exactly which variables to include is difficult, and we have to make some assumptions. The rule is that the information set should include all variables that would be contained in a sophisticated economic model of the variable being analysed. Adding those variables, we test if these are significantly correlated with the forecast errors. If they are, then the agents have not taken sufficiently account of this information in their forecasting (Thomas, 1999). Hence, they are not strong-form rational.

Regarding which variables to include, we assume that they have to be publicly available. One example is to run the regression:

$$A_t - F_t = \alpha + \beta_1 F_t + \beta_2 A_{t-4} + \beta_3 i_t + \beta_4 U_t + \varepsilon_t$$

Performing the tests for the inflation forecasts, we include the forecast itself, F_t, the last actual inflation known at the time and not seasonally affected, A_{t-4}, as well as the current unemployment rate, U_t, and the current interest rate, i_t. To be sure to expect the forecasters to have knowledge about these values, it is important that the data we use are not subject to great revisions. We test if the individuals take sufficiently account of the information about these known variables when they respond to the survey. Hence, we test if α and the β values can be restricted to zero. If we cannot reject the null hypothesis of rationality strong-form rationality can be stated.

5. Working with the survey data

The survey of professional forecasters (SPF) is a large database, and there are several potential problems that we should look into. In this section we present the data thoroughly and discuss different problems we need to consider when working with the dataset. Even though there have been a lot of studies working with the SPF, there have, to our knowledge, not been a lot of focus on examining the problems with the data set in previous literature. We find examining and documenting these issues interesting, and we will therefore present and document those in a thoroughly manner.

In the following we start explaining how to transform the data into comparable measures, section 5.1, before we take a preliminary look at the data in section 5.2. In 5.4 we discuss the industry variable included in the dataset and in section 5.5 we deal with problems that the data set contains.

5.1 Transforming survey data into a comparable measure

There are both quarterly and annual point forecasts of the pgdp levels in the survey, but the survey did not ask for annual levels before the third quarter of 1981. We want a measure of the forecasted annual inflation for the whole time period. By using the quarterly forecasted pgdp levels in the current quarter (pgdp2) and the forecasted level a year from now (pgdp6), we find a measure of the expected one-year ahead inflation:

$$INFPGDP1YR_t = [\left(\frac{PGDP6_t}{PGDP2_t}\right) - 1]$$

This calculated inflation is the measure we use for the forecasted one-year ahead inflation. When analysing and comparing with the actual data, we calculate actual values the same way, only using the IPD of the GDP instead of the pgdp levels.

5.2 A preliminary look at the data

In this sub-section we take a first look at the dataset. We look at the forecasted pgdp levels, the forecasted inflation of the individuals and the mean and median inflation forecasts of those each quarter.

Figure 5.1: The number of forecasted pgdp2 levels for each individual.

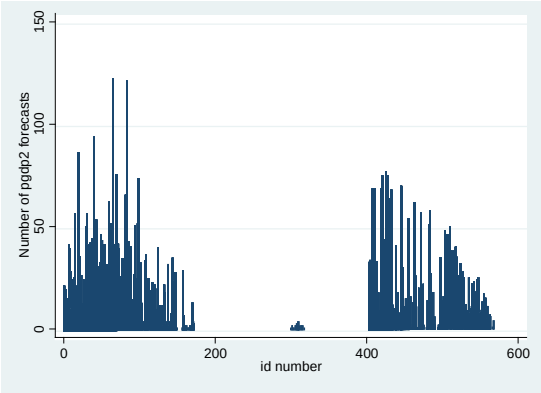
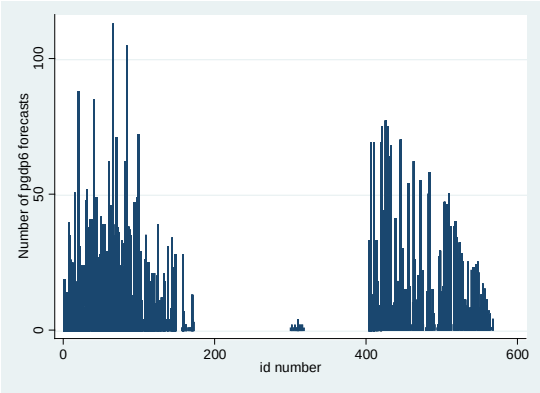


Figure 5.2: The number of forecasted pgdp6 levels for each individual.

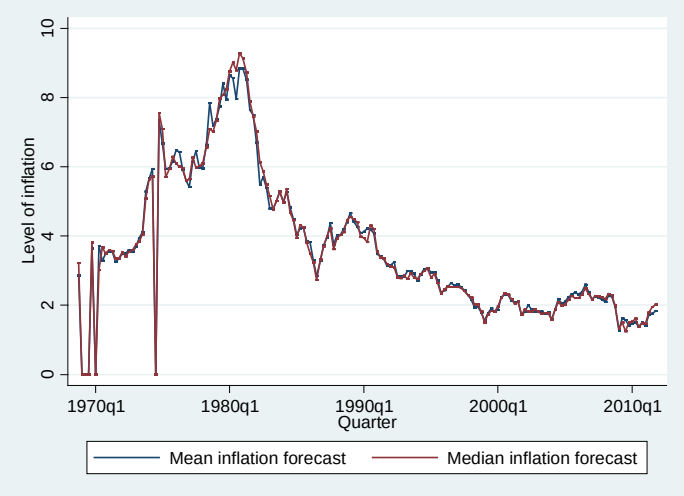


We start by presenting the numbers of forecasted levels of pgdp2 and pgdp6 for each respondent, in figure 5.1 and 5.2. The pattern is almost the same for the other forecasted pgdp levels, presented in appendix 2.2. There are large differences between the individuals in terms of these forecasted levels. Some respondents did not forecast either levels any quarters, and for those we will not be able to calculate the forecasted inflation.

The calculated mean and median one-year ahead inflation forecast of the data before doing anything with the sample is shown in figure 5.3. We see that for some quarters we were not able to calculate either the mean or the median forecast, because of the abovementioned problem of no individuals responding to either pgdp2 or pgdp6 these quarters.

The figures 5.4 and 5.5 show us that there are large irregularities in terms of number of individuals responding to the survey. The number of participants has varied a lot over the years. In 1968 the number was around 60. During the 1970s and 1980s this number decreased, being as low as 14 in 1990. When the survey was taken over by the Federal Reserve Bank of Philadelphia the number increased again, and stabilized at around 30 (Giordani & Söderlind, 2002). The number of respondents will, naturally, matter for the strength of the analysis.

Figure 5.3: The mean and median inflation forecast each quarter. A forecasted value of zero indicates that there are no forecasted inflation forecasts for any individual that quarter. The forecasted value each quarter presents the one-year ahead inflation forecast given in that quarter.



The fact that some quarters have missing one-year ahead inflation forecasts, is also visible through figure 5.5. [17] The number of responses each quarter is presented in figure 5.4, while figure 5.5 displays the number of inflation forecasts each quarter. In figure 5.5 a response involves only that the respondent has received the questionnaire. Therefore, the number of responses is different from the number of responded forecasted pgdp levels (and off course also different from the number of inflation forecasts). Because we will analyse the inflation forecasts, it is number of inflation forecasts that are of most relevance to us.

To show the dispersion in the data, we present the highest and lowest inflation forecast each quarter, presented in figure 5.6. The dispersion is also visible by plotting the standard

Figure 5.4: The number of responses to the survey each quarter. A response involves that a survey questionnaire have been sent to the individual.

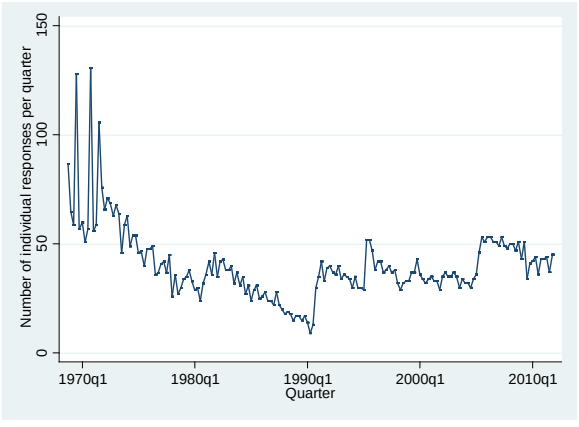
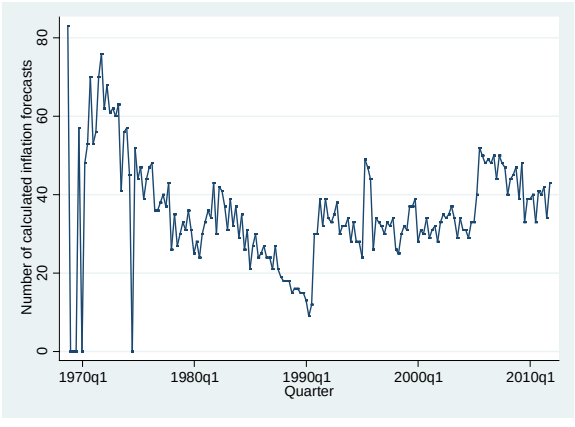


Figure 5.5: The number of inflation forecasts each quarter. An inflation forecast demands that both the pgdp2 and the pgdp6 level have been forecasted.



[17] In the rest of the paper we will often talk about the inflation forecast, meaning the one-year ahead inflation forecasts, without this being specified.

Figure 5.6: The highest and lowest inflation forecast given each quarter. The forecasted value each quarter presents the one-year ahead inflation forecast given in that quarter.

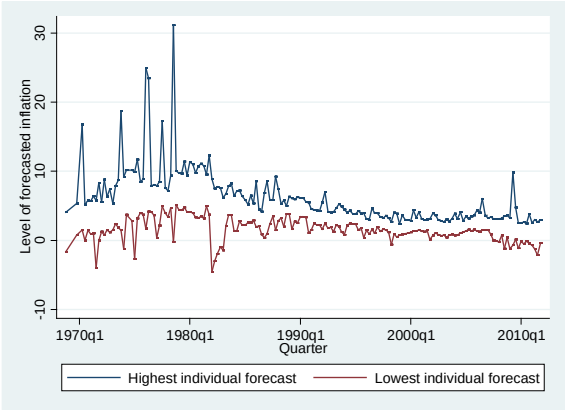
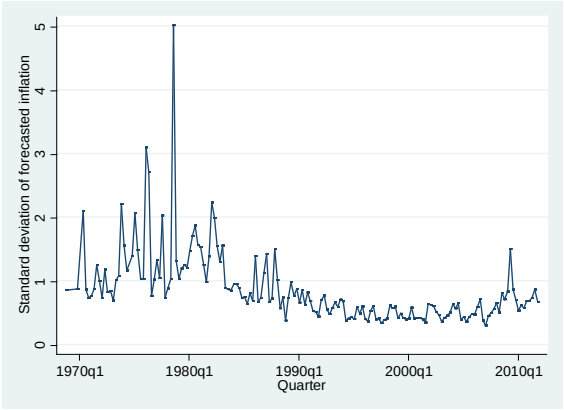


Figure 5.7: The standard deviation of the inflation forecasts given each quarter.



deviation of the calculated inflation forecasts each quarter, figure 5.7. Looking at the two figures we see that there are large timely differences in the variation of the forecasted values. Both the standard deviations and the differences between the highest and the lowest forecast are larger in the beginning of the survey than in the end.

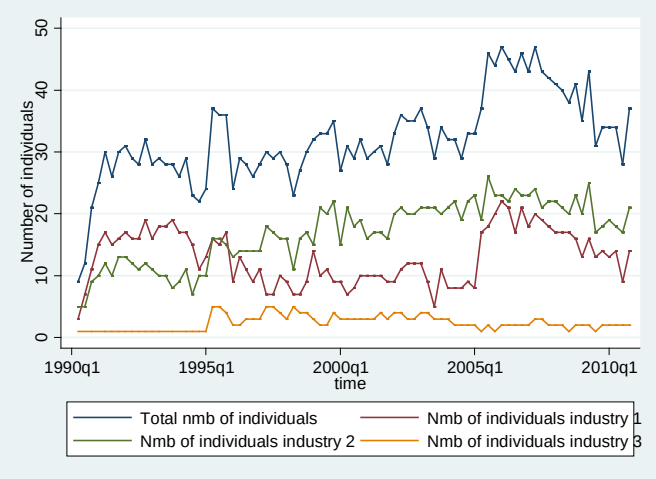
5.3 The industry variable

In addition to the anonymous individual number, the individual data includes an industry classification of the individual respondents (Federal Reserve Bank of Philadelphia, 2008). These were released in May 2008 for the responses after the Philadelphia Fed took over the survey, that is, from the second quarter of 1990. For surveys before 1990 it is not possible to provide industry classification because of lacking hard- copy historical records.

Each forecaster is divided in one out of three industry categories. An industry variable with a value of one means that the respondent is employed in a firm characterized as a financial service provider and a value of two means that the respondent is employed in a nonfinancial service provider firm.[18] If the forecaster is classified with an industry variable of three, they have not been able to classify the industry of the firm where the respondent is employed. The industry classification is conservative, meaning that an industry variable is only assigned to a respondent if they are certain of the respondent's employment and the classification of the firm where he or she is employed. Some might think that including such an industry variable

[18] Being a financial service provider is a firm involved in insurance, investment banking, commercial banking, payment services, hedge and mutual funds, asset management or in association of financial service providers. If employed in a nonfinancial service provider, one is employed in a university, a manufacturing firm, forecasting firm, investment advisor firm, a research firm or a consultant firm (Chew & Price, 2008).A more elaborate discussion of the industry variables is discussed in the analysis, section 6.4.

Figure 5.8: The number of individuals being employed in firms with the different industry classifications each quarter.



may affect the important of the forecasters. However, when using a broad two-sector classification as described above this should not be a big problem.

A respondent's industry variable can change if he or she quits his or hers job and starts working in another firm. The number of participants in each category may also change because of changing composition of the panel. This leaves us with an unpredictable pattern of individuals included in each industry category over the time span. The number of individual forecasters included in every industry variable, as well as the total number of individuals is presented in figure 5.8.

The motivation behind including the industry variables is that different forecasters can have different goals, objectives and constraints, which can be related to the place of employment. One would think that the forecaster's primary objective is to make the most accurate and best forecasts. However, other incentives, for example strategic, can be present. Hence, the industry affiliation of the forecasters can be important when understanding the individual's forecasts (Stark, 1997).

5.4 Problems with the data set

This section documents the problems with the dataset. It has been said that the most important shortcoming of the survey is the high turnover of participants and large frequency gaps in the responses of those participants (Zarnowitz & Braun, 1993). These are issues that we will focus on when examining the dataset.

We start by investigating the forecasters who have only responded to the survey a few times, in section 5.4.1. Then we continue with the respondents who have some missing values in their forecasts in section 5.4.2, before discussing the problem of reallocation of id numbers in 5.4.3. We also discuss the issue of overlapping observations in 5.4.4. A discussion of changing base years are presented in appendix 2.4 and a discussion and some tests regarding the consistency of the inflation forecasts are presented in appendix 2.5.

5.4.1 Respondents with few responses

The respondents have not responded to the survey all years. Some responded in the beginning, others responded later. Some also have gaps in their quarterly responses, answering to the survey some quarters before stopping and responding later again. The analyses of the forecast behaviour of those who only responded to the survey a few times will be weak. To account for this we will restrict the sample.

Almost all previous studies restrict the sample to include only regular forecasts- those who have responded to the survey more than a certain number of times. The number of required surveys answered varies, some choosing 12 responses as their limit, others using 10 or 20 (Keane & Runkle, 1990; Zarnowitz, 1992; Zarnowitz & Braun, 1993; Clements, 2004; Clements, 2008a). We follow the same example as most of the previous studies, deleting those respondents who have 12 or fewer responses in total.[19]

In some quarters there are individuals who did not forecast any of the pgdp levels. For some respondents this goes for all quarters, leaving them with no responses at all to the survey.[20] For others these quarters will be "blank" responses in the middle of forecasted values. We consider these "responses" of both individuals who have not responded to any surveys, as well as for the individuals that have some of these "blank responses" in the middle of their forecasts as not really having responded to the survey this quarter. Hence we exclude these individuals from the data. This means that from now on all of the quarters where an individual have responded to the survey should contain at least one forecasted pgdp level.

[19] It is, however, important to be aware of the fact that even though we have registered a response from an individual, that does not guarantee that the individual have given enough information to make us able to calculate his or hers inflation forecast, hence, it is not certain that we will have this exact number of inflation forecasts per individual.

[20] Looking at the figures 5.1 and 5.2 we can see for whom this is a problem, noting that the pattern of the pgdp2 and pgdp6 responded levels are quite similar as the other pgdp level (as presented in appendix 2.2). This involves that a lot of respondents with an identification number between 200 and 400 not really having answered to the survey at all.

We want to know how restricting the sample to respondents with 12 or more responses affects the dataset. We begin examining the effect this restriction has on the number of surveys and forecasts per respondent, as well as respondents per quarter, in section 5.5.1.1. In 5.2.1.2 we show how restricting the sample affects the value of the forecasts.

5.4.1.1 The effect of restricting the sample on number of surveys and forecasts

When dropping individuals with less than 12 responses, the number of respondents per survey, as well as survey responses per individual, will change. These changes are presented in table 5.1. The number of surveys per respondent is presented in panel A (where the number of surveys per respondents is the number of quarters the respondents have answered to one or more pgdp levels). The average number of surveys increases as the irregular forecasters are removed, giving us a dataset more eligible for analysis. Naturally, the standard deviation of surveys per respondent also decreases when individuals with few responses are dropped. The highest number of surveys that an individual responded to is 123. This does, however, not mean that this was 123 consecutive responses or 123 forecasts.

Panel B in table 5.1 shows the number of respondents per survey. When eliminating irregular forecasters the total number of unique forecasters decreases along with the average number of forecasters per survey. This means that we include data from fewer respondents than we would have if we included the whole dataset. Because the changes are not very severe, it does not seem like removing irregular forecasters alter the database to a great extent.

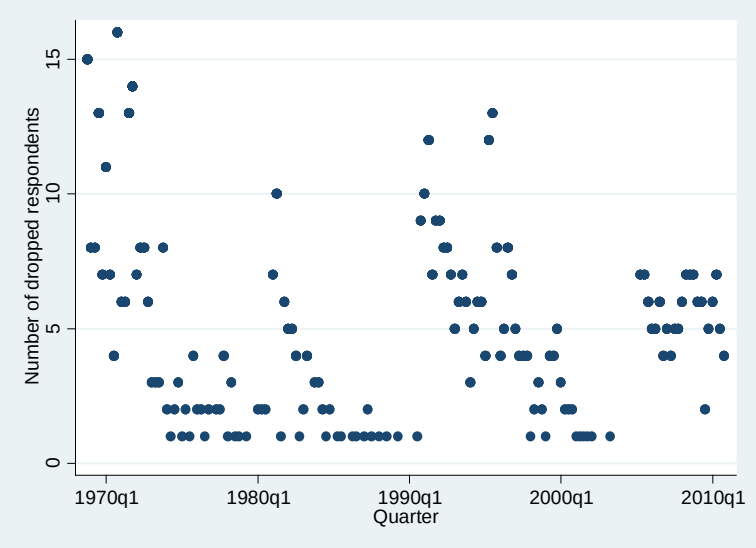
The numbers of inflation forecasts is presented in panel C in table 5.1. The average number of forecasts increases from 38.28 to 41.86, and the standard deviation ("std" in table 5.1) is decreasing. The minimum forecasts per individual have increased from zero to seven, meaning that (at least) one of the respondents who have received the survey 12 or more times, have not responded to both the pgdp2 and pgdp6 in more than seven of the survey questionnaires that he or she received.

In figure 5.9 we plot the number of dropped individuals against the time variable, quarter. The dropped respondents span the whole time period. The maximum respondents dropped in one quarter are 16 in 1970q4. Even though there are respondents dropped over the whole time span, there are not more than seven dropped in one quarter from year 2000. This indicates that the problem with few responses per individual respondent has gone to some degree down.

Table 5.1: Descriptive statistics of the numbers of surveys, respondents and forecasts before and after restricting the sample to those with 12 or more responses. A response means that they have responded to at least one pgdp level in the given quarter.

	All	<12
Deleted observations	-	651
Panel A:	Nmb of surveys per respondent	
Total nmb of surveys	169	169
Mean	40.74	44.53
Std	26.49	25.25
Min	1	12
Max	123	123
Panel B:	Nmb of respondents per survey	
Total nmb of respondents	312	174
Mean	42.24	37.51
Std	13.52	11.24
Min	9	9
Max	83	68
Panel C:	Nmb of forecasts per respondent	
Total nmb of forecasts	6408	5761
Mean	38.28	41.86
Std	24.91	23.71
Min	0	7
Max	113	113

Figure 5.9: Number of dropped respondents, being those with 12 or less observations each quarter.



5.4.1.2 The effect of restricting the sample on the descriptive statistics of the responses

The descriptive statistics of the forecasted levels of pgdp and of the forecasted inflation, changes when we drop respondents. These statistics before and after restricting the sample is presented in table 5.2. The average forecast level of both pgdp2 and pgdp6 are increasing. Thus, the one-year ahead forecasts also increases, from 3.74 % to 3.77 %. The standard deviations are a bit increasing, while the minimum and maximum values are the exact same. Hence, none of the forecasters with the highest and lowest forecasts were dropped. The dispersion in the data will therefore be very similar to the one presented in the preliminary look at the data.

Figure 5.10 show the changes in the forecasted values between the full dataset and the dataset with only those with 12 or more responses. We plot the differences, subtracting the "new" median forecast in the restricted sample from the "old" median forecast. This is done for forecasts of pgdp2, pgdp6 and the forecasted one-year ahead inflation. With the change in the inflation forecast being close to zero, the changes are not very severe.[21]

Figure 5.11 presents the same, but for the inflation forecasts only. The largest differences between the median inflation forecast from the full sample and the restricted sample are in the beginning, and they decrease over time.

Table 5.2: Statistics of the forecast pgdp2 and pgdp6 level, as well as the calculated one-year ahead inflation forecasts for the sample with all respondents and in the sample with only those with more than 12 observations.

Statistics	All			>=12		
	pgdp2	pgdp6	Inflation forecast	pgdp2	pgdp6	Inflation forecast
Observations	6403	5975	5974	5757	5391	5390
Mean	139.18	144.99	3.74	139.61	145.42	3.77
Std	32.16	36.12	2.16	32.69	36.59	2.17
Min	104.41	105.70	-4.57	104.41	105.70	-4.57
Max	235	247	31.14	235	247	31.14

[21] Looking at the figure it seems as there are two outliers in the beginning of the survey. Around 1970 there is one change in the median of pgdp2 that is one, much higher than the others. For the difference in the median forecasts there is also one difference that stands out, being about -0.7 in around the beginning of 1970. These outliers will be eliminated because of our choice of starting the sample in the fourth quarter of 1974.

Figure 5.10: Plots of the median forecast found in the SPF minus the new median after dropping those participants who responded less than 12 times to the survey for pgdp2, pgdp6 and the inflation forecast.

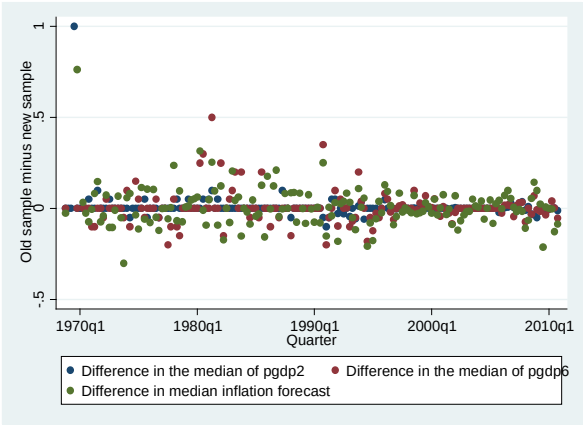
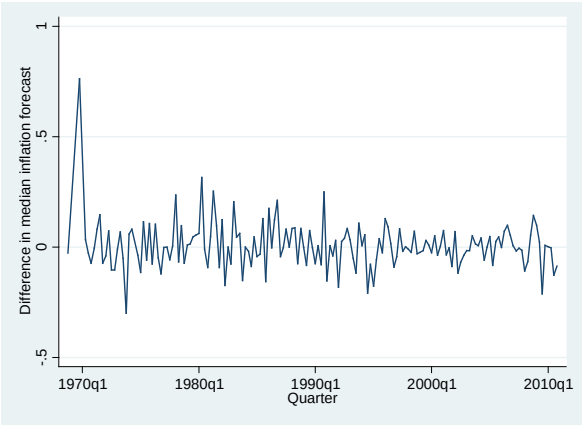


Figure 5.11: The difference between the median inflation forecasts, calculated as the median from the SPF minus the new median when dropping forecasters with 12 or less responses against time.

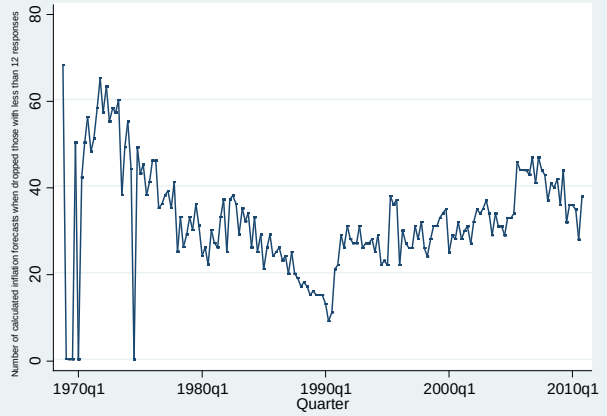


5.4.2 Individuals with some missing forecast values

The missing values in the survey affect the relationship between the number of respondents in a quarter and inflation forecasts in the same quarter (as previously shown in figure 5.4 and figure 5.5).

Figures 5.1 and 5.2 show that the patterns of forecasted values of pgdp2 and pgdp6 for given individuals are similar. However, some differences exist. Some individuals forecasted one of the levels, but not the other one in a given quarter. In some quarters there are no individuals who have forecasted both levels, leaving us without any inflation forecasts that quarter. When restricting the dataset to include only those with 12 or more responses, this problem has not been eliminated. Figure 5.12 presents the number of inflation forecasts each quarter, showing that this problem exists in the three first quarters of 1969, the first quarter of 1970 and the third quarter of 1974. This figure is similar to figure 5.5, except that it is for the restricted sample containing only those with 12 responses or more. For all those quarters, it is the forecast for pgdp6 that is missing.

Figure 5.12: The number of inflation forecasts each quarter in the sample with individuals with 12 or more responses only.



When analysing we want to have at least one inflation forecast for each quarter. Hence, it is preferable to do something about the missing forecasts problem. There are different solutions one could think of. Examples are to restrict the sample to only include forecasts after the last missing one-year ahead forecast (as presented in 5.4.2.1) and filling in an estimated measure for the missing values[22]. These two alternatives have different positive and negative consequences. While the first one does not demand us to alter the dataset, it does force us to delete a lot of observations. And while the second one does not require us to delete observations, we have to change the data. Then some of the forecasts that we analyse will not be an actual forecast.

When choosing one of the given alternatives, we emphasize the fact that changing the dataset is a task that needs a lot of consideration. Additionally we do not know of anyone else who have worked or are working with the SPF doing anything similar. Hence, given that this is "only" a master thesis paper, we do not wish to alter the data to a great extent. Therefore we choose the option of restricting the sample to include forecasts only after the third quarter of 1974.[23]

There are other aspects, not related to the missing values, which also leads us towards this conclusion. With the period from 1968 to 1973 considered to be a challenging period to forecast, leaving this period out of the sample might actually strengthen our analysis (Su &

[22] In terms of filling in an estimated value for the missing values, there are several methods one could think of. Examples are to make a liner projection if having the other pgdp levels necessary to do so, to fill in lead and lag values of pgdp2 and pgdp6, and to find out how the individual has performed compared to the mean before and then fill in a value equivalent to those for the missing value.

[23] After making this restriction we again restrict the sample to those with less then 12 responses. These will be, the ones who because of this restriction now have too few responses.

Su, 1975). The fact that in the early years, 1968, 1969 and 1970 the forecasts were rounded to their whole number, causing the forecasts in these years to be quite erratic, also suggests that the eliminating these years could be the best solution (Croushore, 2006) [24] (these issues are discussed in appendix one).

In the following we present how the dataset is altered when we exclude the quarters before the third quarter of 1974. In appendix 2.3 we present one option of filling in estimated values, and how that would have altered the dataset.

5.4.2.1 Restricting the sample to those after the last missing inflation forecast

The alternative of restricting the sample to forecasts made after the third quarter of 1974 do not demand us to change the data that are still included in the dataset. However, the mean and median values will be altered because we have less data. In addition we will be missing out on a lot of data, especially because the largest number of survey respondents is in the beginning.

In this section we present how starting the sample to in the fourth quarter of 1974 changes the data. The new number of forecasts per individual, the number of survey responses per individual and the number of responses per survey are shown in table 5.3. Also presented are the new values of the pgdp2, pgdp6 and the forecasted inflation. The corresponding values for the whole time span are previously presented in tables 5.1 and 5.2. We see that the average number of surveys per respondent is now 42.75, which is a small decline from the previous number of 44.53. The number of respondents per survey has also declined, while the number of forecasts per respondents is almost the same as before, with 41.52 now, compared to the previous 41.86. The fact that these numbers have declined could make our analysis weaker. However, because we focus on the individuals and the number of forecasts per individual is almost unchanged, this does not seem to be a severe problem.

The forecasted values of both pgdp2 and pgdp6, thus also the inflation has decreased a little. This is natural, because some of the early years with a high inflation are now deleted.

[24] The fact that the forecasts have a larger dispersion in the beginning than later (as shown in figure 5.6 and 5.7) could also be an advantage of this alternative, leaving us with a dataset where the data are more stable.

Table 5.3: Descriptive statistics for number of surveys answered and forecasted values in the data sample starting in 1974q3.

Statistics nmb of surveys answered	Only after 1974q3		
	Nmb forecasts per ind	Nmb surveys per ind	Nmb responses per survey
Mean	41.516	42.746	32.833
Std	24.209	25.704	7.466
Min	7	12	9
Max	102	118	49
Statistics forecasted values	Only after 1974q3		
	pgdp2	pgdp6	Inflation forecast
Mean	136.703	142.140	3.640
Std	35.228	38.861	2.256
Min	104.41	105.7	-4.569
Max	235	247	31.137

5.4.3 Reallocation of identification numbers

The individual forecasters in the SPF have confidential identification numbers. These are supposed to be consistent over time, and one should be able to trace a given forecaster from survey to survey (Federal Reserve Bank of Philadelphia, 2008). Even though one identification number is supposed to belong to one specific individual, this is not guaranteed true for all individuals. This section discusses whether this could involve problems when analysing the individual respondents. To our knowledge, this problem has not been debated in any previous literature.[25]

When ASA/NBER conducted the survey, from its start in 1968 until the first quarter of 1990, hard-copy historical records are missing (Federal Reserve Bank of Philadelphia, 2008). Thus, we should be careful when interpreting the results from the individuals who are forecasting in this period. The problem arises if an identification number stopped responding for a long time, before responding again later. Then the identification number could have been given to another forecaster, which can cause problems when we analyse the individuals' forecast behaviour.

A possible solution is to divide individuals who have large gaps in their responses into two or more individuals. However, the Philadelphia Fed is not sure if the individual numbers really were re-used. It could be that a person decided to stop responding to the survey for a long time, and then started again at a later stage. Nevertheless, if the gap is big enough we can argue that the respondent can have changed over the years anyhow. To call her or him a new

[25]A reason could be that many previous studies use consensus forecast where this is not a problem. Those who are conducting individual tests have not discussed the issue either, maybe because they were not aware of the problem at the time. We do not know exactly when the awareness of this problem arisen.

person would then not be terribly wrong. When deciding how large the gap should be before the identification numbers are divided up, we need to consider that some responses can be absent due to natural causes such as child birth and sick leave. The gap should therefore be large enough to consider such causes, for example 5 years (or 20 quarters) or more.

In table 5.4 we present how large the gaps are in the restricted sample containing only those with 12 or more responses. This is also presented for the shorter sample where the forecasts before the fourth quarter of 1974 are excluded. On average the gap is about 1.67 quarters in the total sample, and with one quarter gap being no gap there are several individuals do not have a gap. When only looking at the shorter sample, the average gap is smaller, 1.45 quarters. If we take a look at those with gaps only, the mean gap is about 4.10 quarters, a bit more than one year. The largest gap is at 52 quarters, or 13 years.

We want to find if many individuals have large gaps. The possibility of individuals having several large gaps also exists. When examining we did not find any individuals who have more than one gap longer than five years, thus no individuals with several longer gaps than five years either. The maximum gap of the total sample and the shorter sample together with the number of individuals having a gap longer than five, ten, fifteen and twenty years are presented in table 5.5. 42 individuals have a gap larger than five years, 17 have a gap of more than ten, but no one have a gap larger than 15 years.

Table 5.4: Statistics of the gaps between individual responses. We show statistics for observations without any gaps, for observations in the total sample, for the observations with gap as well as the shorter sample starting in 1974q4.

Gap between responses	Obs	Mean	Std	Max Gap
No gap	4381	1	0	1
All	5590	1.671	3.415	52
With gap	1209	4.103	6.812	52
From 1974q4	4327	1.449	1.937	47

Table 5.5: The maximum gap in the total sample and in the shorter sample starting in 1975q4, together with the number of individuals with a gap longer than 5, 10, 15 and 20 years.

Gap length/Number of individuals with gap		
Gap	All	From 1974q4
Max Gap (yr)	13	11.75
>5yr	42	13
>10yr	17	2
>15yr	0	0
>20yr	0	0

Figure 5.13: Number of Individuals with gaps larger than five years through time. The points indicate where the gap started.

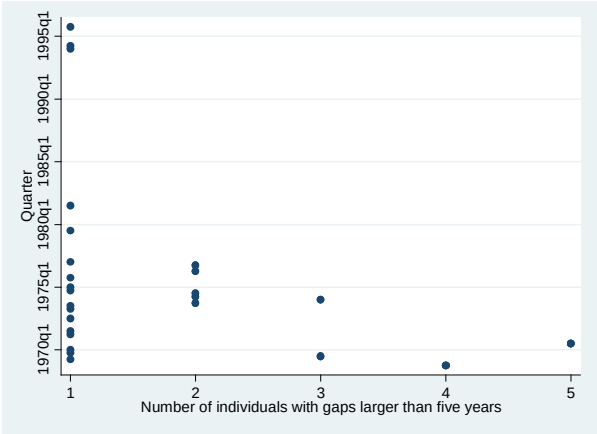
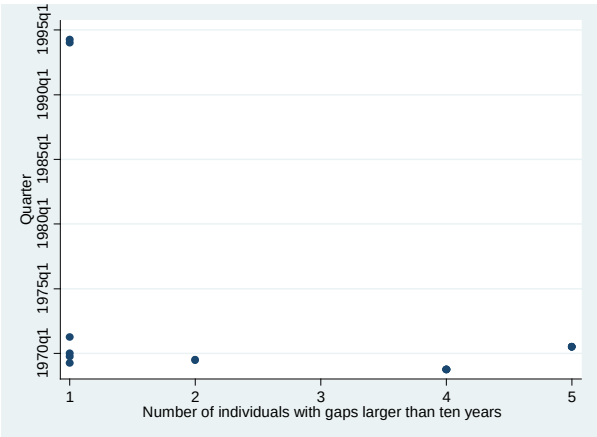


Figure 5.14: Number of individuals with gaps larger than ten years through time. The points indicate where the gap started.



We also want to locate where in the survey the gaps are located. Looking at figure 5.13 and figure 5.14 we see that the large gaps are located mainly in the beginning of the survey. Most gaps longer than five years are before 1975, the same period where the largest number of individuals with gaps is localized. In the third quarter of 1970 there are five individuals with gaps larger than five years. Also between 1975 and 1982 there are some gaps larger than five years. From 1982 there are no gaps of this size before the three localized just before 1995. With only three such large gaps coming from Philadelphia Feds period, this problem exist mostly during the ASA/NBER period. All but two of the gaps longer than ten years are located before 1975.

Because a lot of the gaps are located in the beginning of the survey, the problem decreases when we choose to use data only after the third quarter of 1974. Table 5.5 shows us that there are only 13 individuals with gaps longer than five years and only two with gaps longer than ten years in this period.

This leaves us somewhat unsecure about the severity of the problem. Together with the fact that the problem decreases when choosing to use the sample starting in the fourth quarter of 1974, we choose not to divide the uncertain individuals in more than one respondent.

The industry number introduced in the Philadelphia Fed period of the survey leads to another potential problem. If the forecasters change place of employment, the issue of whether the identification number should follow the individual or the place of employment arise (Federal Reserve Bank of Philadelphia, 2008). The survey tried to solve this by letting the identification number stay with the firm if the forecast seems more associated with the firm

than with the individual and vice versa. If the identification number is more identified with a firm than an individual, individual rationality tests will examine whether the firm is rational rather than the individual. We examine this problem, and find it not to be a big problem in our data because there are only seven individuals where this could be a problem. These are individual number 65, 407, 420, 421, 426, 448 and 463. For most of those the industry variable switches from one to two. One exception is number 65, who first has an industry variable of one, then two, before being categorized with an industry variable of three. The other exception is individual 463 where the variable changes from two to three.

5.4.4 Overlapping observations and autocorrelation

Because we use forecasts that span a four-quarter horizon, there may be some overlapping observations that can create autocorrelation. When shocks occur they will affect the actuals, and thus the forecast errors for several consecutive periods, because the forecasts span a longer period than the sampling frequency (Croushore, 2006). When a shock hit the economy, for example the oil price shock in 1973q2, all forecast errors that include this quarter will be affected. This means that the forecast errors for this given example will be correlated in the surveys conducted in 1972q2, 1972q3, 1972q4, 1973q1 and 1973q2.

A “normal” ordinary least squares regression (OLS) require the errors to by serially uncorrelated. Because the abovementioned overlapping observations will create a moving average (MA) error term, the OLS parameter estimates will not be efficient in this sample, and tests performed by OLS will therefore be biased (Hansen & Hodrick, 1980; Harri & Brorsen, 2009). A normal OLS also requires homoscedasticity, meaning that the error terms have a constant variance. If forecasting is more difficult in some periods than others, heteroskedasticity might exist. This will make the OLS regression not efficient. To be able to perform tests, we have to find a correct way of running regressions that accounts for both of the abovementioned problems.

One way of correcting for the overlapping observations problem is to use a restricted sample, thus cutting the SPF sample into five pieces (Harri & Brorsen, 2009). This will, however, limit our dataset, and we will not be able to exploit all the information that we have, using only each fifth observation. Because this would give us even fewer responses per individual than we already have, this does not seem like a good solution. Also the fact that we cannot be sure that the individuals have responded in exactly these necessary quarters is an argument against using this solution.

Another solution is to use the overlapping observations, but to account for autocorrelation when testing. Several estimators that are both heteroskedastic and autocorrelation consistent have been constructed (HAC estimators). These make hypothesis testing valid when using data with overlapping observations (Harri & Brorsen, 2009). Such estimators are favourable for our analysis, and include among others Hansen and Hodrick (Hansen & Hodrick, 1980), and Newey-West (Newey & West, 1987). By computing a weighted variance – covariance matrix that gives less weight to the errors made in the observations that are either highly-serially correlated or heteroskedastic, the Newey-West method guarantees for a positive definite covariance matrix. This consistent covariance matrix is computed by using the OLS residuals (Harri & Brorsen, 2009). The method can easily be done in Stata, by running tests with p-values based on a chi-squared test using this method.[26]

Because of the advantages of the Newey-West method, this is the method we want to use in our analysis. However, because the method exploits information in lags, some problems can arise if the data we are analysing have missing values. If m is the specified maximum lag, the method multiplies the covariance of lag j by the weight [1- j/(m+1)] (Petersen, 2009). If an individual respondent have not given responses every quarter when participating in the survey, he or she may not have the lags necessary for making the estimates. This can be a problem in our analysis, and we will mention those if they rise.

[26] The Newey-West variance HAC estimator accounts for autocorrelation up to and including a lag of m. Thus, autocorrelation at lags greater that m is ignored. We have to be aware of how many lags to use. If the overlap is 5, then the errors are MA(4). As the lag length increases, so do the standard errors estimated by the Newey-West method. Hence, the standard errors will be higher, and the t-statistics lower than when running an OLS regression.

6. Analysis

This section contains our main analysis of the data. We use the revised dataset where we restricted the sample to those with 12 or more responses. We also choose to start the sample in the fourth quarter of 1974 (as discussed in section five). Starting in section 6.1 we first take a look and the data, investigating briefly how the forecasted inflation and the actual data have evolved through time.[27] Then, in section 6.2, we examine the forecast errors calculating different accuracy measures to see how accurate the inflation forecasts of the respondents have been. The main part of our analysis involves testing if the forecasters are rational or not, section 6.3.

Previous literature has studied the accuracy and rationality of inflation expectations (Zarnowitz, 1985; Thomas, 1999; Mankiw, et al., 2003; Clements, 2004; Croushore, 2006; Gerberding, 2006). However, there have been relatively few studies analysing individual forecasts. Examples of such are Keane and Runkle (1990) and Zarnowitz (1985). With this in mind, stating what we can contribute to the literature is important. First, we have a new sample, which also includes the financial crisis. Our analysis will therefore be of current interest. Our analysis of the individual data is also more detailed than previous literature. We find how many individuals who are rational and how many who perform better than the consensus. We also find which individuals that are the best forecasters in terms of each test. These are the main topics of section 6.2 and 6.3, where we examine these issues looking at the whole data sample.

Other interesting issues regarding rationality of forecasts also arise. Examining the industry variable in the SPF and whether or not there are differences between the individual respondents employed in the different industry categories is something we do not have knowledge of other papers investigating. In this thesis we add to the previous literature by looking into this industry variable. This discussion is presented in section 6.4.

A lot of previous papers have investigated differences between different sub-periods, for example differences between expansions and contractions (for instance Su and Su (1975), McNees (1992), and Mehra (2002)). Adding to the literature, we examine how the Volcker

[27] A discussion about whether large forecast errors could have originated because of challenging time period that could have created forecasting difficulties is presented in appendix one. This discussion involves presenting what previous literature has found.

disinflation affected forecasters' performances. In a paper from 2002 Mehra excluded the Volcker disinflation from the entire sample, and found that the SPF performed better when this period was excluded. However, we do not have knowledge of literature examining this on an individual level and in a very detailed manner. Our motivation behind focusing on this period is that when we look at the differences between the actual and the forecasted inflation in figures 6.1 and 6.2, the forecast performance of the respondents seems to have been different these years. Our discussion and analysis of the Volcker disinflation's effect on the forecasts are presented in section 6.5.

Having a new data sample containing the financial crisis makes us question whether the financial crisis had an effect on the forecasts. Adding to the literature we examine the rationality during the recent crisis. We also examine the performance of the individual respondents during this special time period. This analysis is presented in section 6.6.

With our detailed analysis of individual respondents, we may find differences and disagreement between individual respondents. Such distinctions between individuals are something that most macroeconomic models do not account for (Mankiw, et al., 2003). Hence, to be able to understand how individuals form their expectations is important to look for differences and patterns among the individual respondents. This is a task we wish to contribute to in this paper.

6.1 A preliminary comparison of the forecasted inflation and the actual inflation

When comparing the calculated inflation forecasts with the actual data, we have to transform the actual values in the same manner as we did with the survey data:

$$INFPGDP1YR_t = [\left(\frac{IPD_t}{IPD_{t-4}}\right) - 1]$$

Figure 6.1: The calculated mean inflation forecast from the SPF and the calculated actual inflation for the entire sample. The forecasted value presents the one-year ahead inflation forecast given in that quarter, and the corresponding actual the next year inflation.

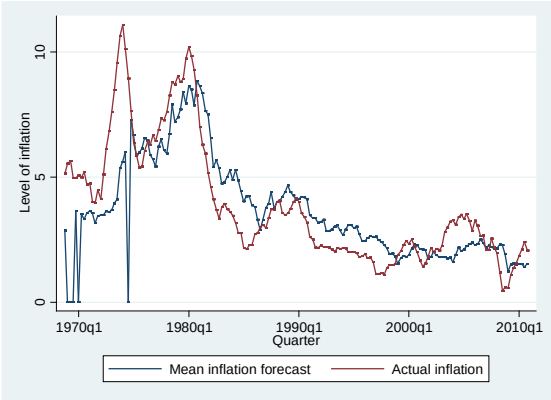


Figure 6.2: The actual inflation and the individual inflation forecasts from the SPF for the entire sample. The forecasted value presents the one-year ahead inflation forecast given in that quarter, and the corresponding actual the next year inflation.

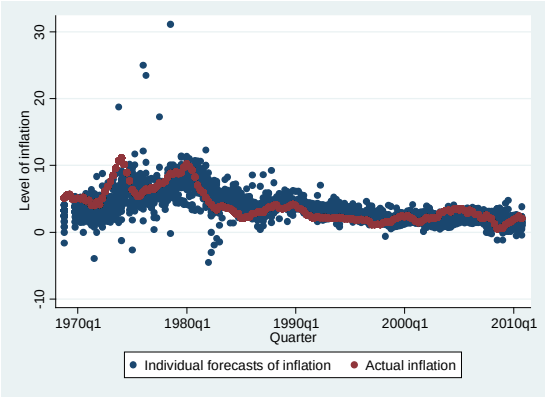


Figure 6.3: The calculated mean inflation forecast from the SPF and the calculated actual inflation for the shorter sample starting in 1974q4.

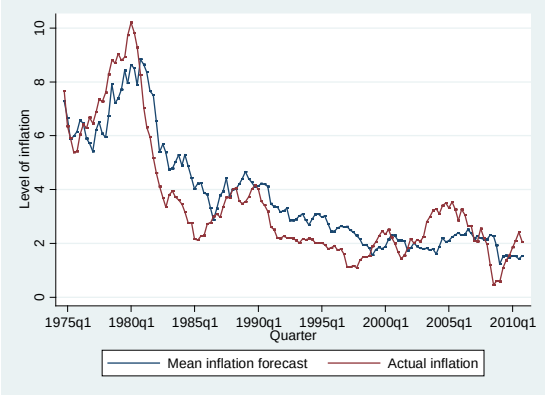
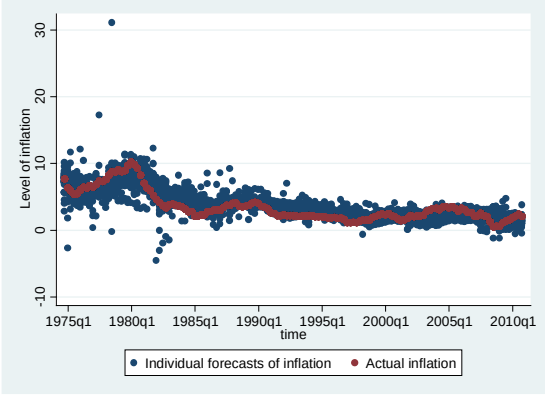


Figure 6.4: The actual inflation and the individual inflation forecasts from the SPF for the shorter sample starting in 1974q4.



In figure 6.1 we see how the mean forecasted inflation of the individuals in the SPF as well as the actual inflation have evolved through time. We use the forecasted one-year ahead inflation made each quarter and compare them with the actual inflation that same quarter. The forecast error in a given quarter in the graph is thus the calculated actual inflation over the next year, minus the calculated forecasted one-year ahead inflation made in that quarter.[28]

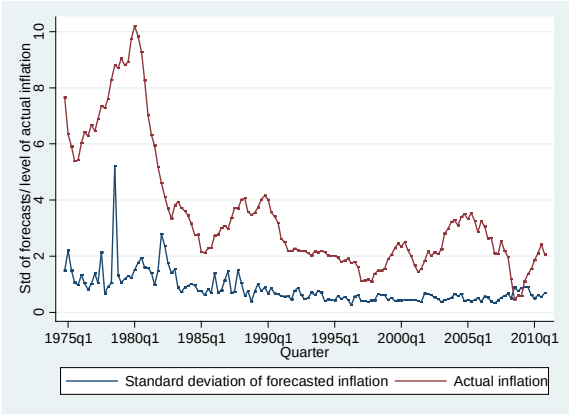
By looking at figure 6.1 we find differences, with larger forecast errors between the predicted and the actual inflation in the first years of the survey than later. This is in line with previous literature by McNees (1992) Thomas (1999), Gerberding (2006) and Croushore (2006). Some of the explanation may be that the levels of the predicted inflation, as well as the dispersion in the actual inflation were also higher these years (Mehra, 2002; Thomas, 1999). By plotting all the individual forecasts against the actual inflation in figure 6.2, we see that the pattern of the mean forecasts, naturally, reflect the individual respondents' forecasts.

[28] This is the way we will present the forecast error and the actual data in the entire paper.

Because we choose to restrict the sample to the quarters after the fourth quarter of 1974 (explained in 5.4.2), some of the large differences in the beginning will be eliminated. Figure 6.3 shows the mean inflation forecasts against the actual inflation in this new and shorter period of time, and figure 6.4 plots the individual forecasts against the actual inflation. When comparing figure 6.1 and figure 6.3 we see that the problems with missing forecasts of inflation (represented by a forecast of zero in figure 6.1) for some quarters are absent. Some other forecasts with large forecast errors made in the beginning of the survey are naturally also eliminated.

Even with our shorter dataset the forecast errors still seem to be much larger in the beginning than later. The standard deviations of the individual responses over the years show us the dispersion in the forecasts and can hint when the inflation was hardest to predict. Large standard devations involves relatively large differences between the forecasts of individuals, which again indicate diasgreement between respondents. A large disagreement may indicate that it is hard to predict the inflation.[29] The standard deviation of the whole sample is presented in figure 5.7, section 5.2. Figure 6.5 shows the standard devations of the forecasts, together with the level of the actual inflation, for our shorter sample starting in the fourth quarter of 1974. We see that the level of the standard deviation follow to some degree the same pattern as the level of the actual inflation, being higher when the level of the actual inflation is higher. This relationship seems to hold, at least before around 1995. After year 2000 the level of the actual inflation have increased without the standard deviation of the forecasts following, thus this relastionship seem to have ceased.

Figure 6.5: The standard deviations of the forecasts and the level of the actual inflation over time.



[29] We could imagine that investigating the probability forecasts of each individual would give us a more valid measure of when the uncertainty involving the inflation were the highest, thus when predicting the inflation was difficult. However, Giordani and Söderlind (2002) examined this and found that the dispersion between individual forecasters to be a good measure of uncertainty and disagreement.

6.2 Evaluating forecasts using different accuracy measures

This section examines the accuracy of the respondents' inflation forecasts. We analyse the forecast accuracy of both the consensus forecasts and the forecasts of individual respondents from the fourth quarter of 1974 until the end of 2010 (using the accuracy measures described in section 4.1). We find the individuals who are the best forecasters in terms of each accuracy measure and whether these are the same when using different accuracy measures. In addition we compare them with the consensus accuracy measures. We start presenting the accuracy measures for the consensus in section 6.2.1, before continuing with the individuals in chapter 6.2.2. Different from only calculating the accuracy measures of the consensus (as done for by for example Thomas (1999)), examining the forecasts of the individuals is to our knowledge relatively seldom.

6.2.1 Forecast accuracy of the consensus forecasts

The accuracy measures for the consensus mean and median forecasts for the entire survey period are presented in table 6.1. With the errors calculated by subtracting the forecasted value of the inflation for a given period from the actual inflation in the same given period, a positive mean error (ME) indicates underestimation, while a negative ME indicates overestimation. The ME values are negative for both the mean and median consensus, hence the forecasters overestimates the inflation on average. An overestimation of the actual value could be explained by a negative development in actual values (DeLong, 1997; Thomas, 1999; Mehra, 2002), and with the inflation level decreasing during our sample, this explanation seems to hold.[30] With the inflation measured in percentages, and the ME of the

Table 6.1: Accuracy measures for the consensus in the whole sample starting in 1974q4.

Accuracy measures for the consensus		
	Mean	Median
ME: Mean Error	-0.275	-0.281
MAE: Mean Absolute Error	0.865	0.885
RMSE: Root mean squared error	3.313	3.382
MNSE: mean normalized squared error	2.222	2.263
Standard devation	2.014	2.075
Number of forecasts per individual	41.909	42

[30] However, in the most recent years, the inflation level has been much more stable, hence the overestimation should not be as distinct in the more recent surveys. The Volcker disinflation period in the early 1980s is the period where the inflation is decreasing the most, indicating that this period might be the reason for the overestimation result. Hence, when looking at a more previous sample, as done in 6.4.2, and the sample where we have excluded the Volcker disinflation, section 6.5.1, we may expect the results to differ.

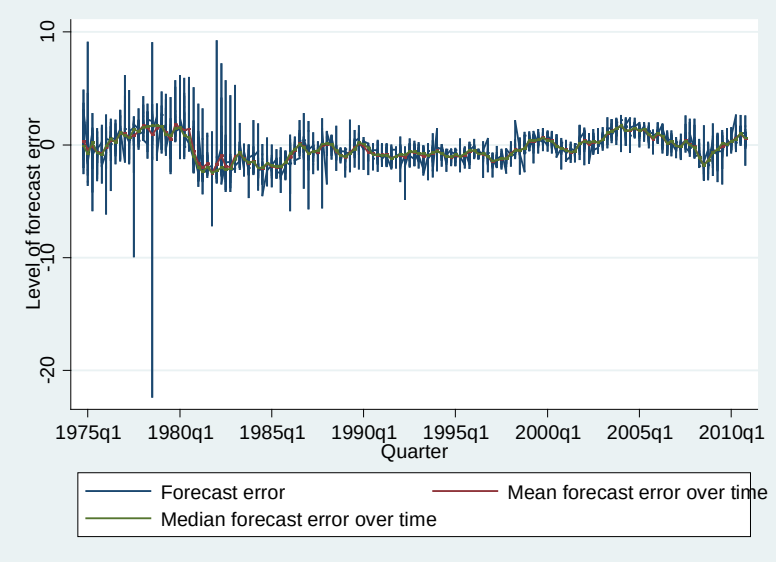
mean consensus being 0.28 (and quite similar using the median consensus), the inflation forecasts is on average 0.28 % too high. Because the ME takes the sign of the errors into account, the magnitude of this measure is not very reliable, because of a probability of positive values offsetting negative ones.

The mean absolute error (MAE), which disregards the sign of the errors are presented in the second rows of table 6.1. MAE has a value at 0.87 for the mean consensus forecasts, with the median being a bit higher. Hence, the respondents' forecast errors are a bit less than one percent on average, not giving us any information of whether they usually underestimate or overestimate the inflation. The root mean squared error (RMSE) also disregards the sign of the errors, and in addition it penalizes large errors more heavily than small errors. Also for the RMSE and the mean normalized squared error (MNSE), which accounts for the variation in the actual inflation, the mean and median consensus values are quite similar, with the median being a bit higher than the mean. We also present the standard deviations of the inflation forecasts, showing that the consensus deviation from the mean forecast is over two percent. The consensus mean and median numbers of forecasts per individual is 41.9 for the mean and 42 for the median, indicating that the individuals on average have been forecasting in about ten years, given that their forecasts are mostly consecutive.

6.2.2 Forecast accuracy of individuals

This section calculates the accuracy measures of each individual. We find the best and the worst forecasters in terms of each measure, and we examine if these best and worst exhibit any patterns. In order for us to gain a better understanding of how expectations of individuals

Figure 6.6: The inflation forecast errors and the mean and median inflation forecast errors over time. The errors are calculated as the actual inflation minus the forecasted inflation.



are formed, for example whether they are accurate as the rational expectations hypothesis presumes (Gerberding, 2006), differences between them can be important to examine.

The individual respondents' one-year ahead inflation forecasts and the actual inflation are presented in figure 6.3. The forecast errors together with the mean and median forecast error over time are presented in figure 6.6. The forecast errors were larger in the beginning of the survey period, than in the more recent surveys. This indicates that it was harder to forecast in the beginning, when the level and the dispersion of the actual inflation were higher. This is in accordance with results found by McNees (1992), Thomas (1999) and Croushore (2006), and is an issue that we will discuss when looking at the forecast accuracy of the individuals.

We continue with calculating the different accuracy measures for the individual respondents. Starting with presenting the pattern of the calculated accuracy measures of the individuals in section 6.2.2.1 we continue examining the ten most accurate and the ten least accurate forecasters in terms of each accuracy measure, in section 6.2.2.2. In 6.2.2.3 we investigate if the best and the worst forecasters in terms of the accuracy measures overlap, thus if some forecasters are better than others in terms of all measures.

6.2.2.1 A first look at the calculated accuracy measures for the individuals

The calculated accuracy measures of each individual respondent are presented in figures 6.7-6.10, and in figure 6.11 we present the standard deviation of the actual inflation in the periods where the individual respondents answered the survey. By taking a quick look at the figures, we see that there are large differences in terms of accuracy between the individual respondents.

A common pattern is found when looking at ME, MAE and RMSE in figures 6.7-6.9. For all these accuracy measures the largest values are among the first individual numbers, who are located mostly in the beginning of the survey. This indicates that the forecasters performed worse in the beginning; hence that forecasting might have been harder in the early years of the survey than later. Therefore, when allowing positive errors to offset negative ones using the ME measure, when not allowing such offsetting of values using the MAE and the RMSE measures, as well as when penalizing larger errors more than small errors in the RMSE measure, the early respondents are worse than the late respondents. These findings are in accordance with other studies (Croushore, 2006; Gerberding, 2006).

Because the time period the individuals are forecasting the inflation for seems to matter for their accuracy, finding a way to normalize for time seems relevant. We do this by finding the respondents' MNSE value. The MNSE accounts for the dispersion in the actual inflation, measured by the standard deviation in the actual inflation in the given periods (as explained in section 4.1.4). Looking at the individual respondents' MNSE values presented in figure 6.10, the pattern of the highest values being located with the lowest individual numbers is less prominent. The explanation is to some degree seen in figure 6.11, with the standard deviation of the actual inflation being higher for respondents with low individual numbers located in the beginning of the survey. Thus, when normalizing for the variation in the actual inflation in the individual respondents' forecasting period, the earliest forecasters do not seem clearly worse than the later ones. Again these results for the individuals seems to be in line with the abovementioned consensus studies, with early studies finding a poor survey performance that according to Croushore (2006) can be explained by the period of time they were examining. The 1970s and 1980s included oil price shocks and bad monetary policy (Croushore, 2006), followed by the Volcker disinflation period that made the inflation unstable and unpredictable, increasing the dispersion in the inflation and making forecasting harder.

With the ME measure allowing positive errors to offset negative ones, we should be able to discover patterns of over,- and underestimation by the individual respondents. With a negative ME value indicating overestimation, figure 6.7 seems to show a larger degree of overestimation in the beginning of the survey, where some individuals have large negative values of the ME. However, also in more recent times, from approximately individual number 400 to 500, there were a period of overestimation. Thus, making any conclusions regarding over,- and underestimation is hard.

Figure 6.7-.6.10: The calculated accuracy measures for each individual number.

Figure 6.7: ME

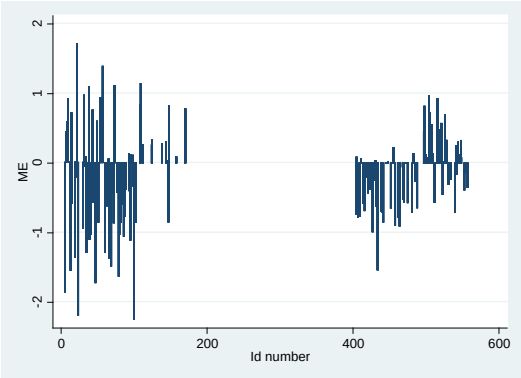


Figure 6.8: MAE

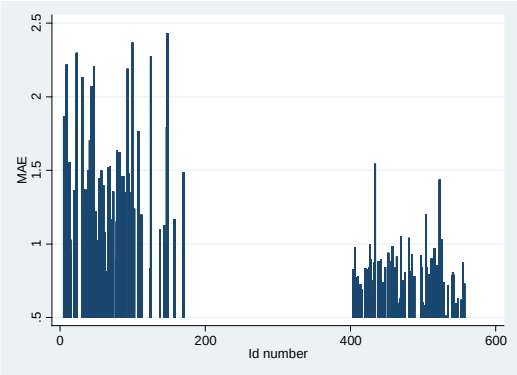


Figure 6.9: RMSE

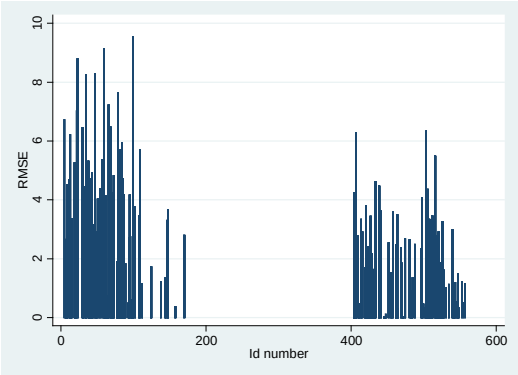


Figure 6.10: MNSE

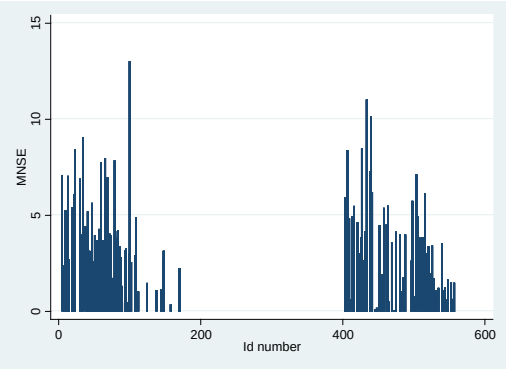
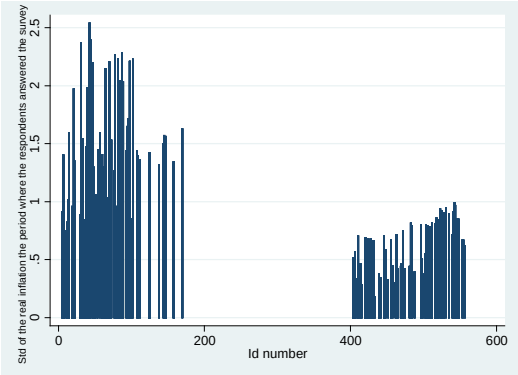


Figure 6.11: The standard deviations in the period of forecasting for each individual.



6.2.2.2 Examining the best and the worst ten individuals in terms of accuracy

In this section we first find the ten most and the ten least accurate individuals in terms of each measure, presented in tables 6.2-6.9. In addition to their rank and their accuracy measure values, we present their individual number, their time period of forecasting, their sum of errors, their number of forecasts and their standard deviations of both forecast errors and the actual inflation in their period of forecasting. We want to investigate if any patterns that can highlight why some forecasters are accurate and others are not exist.

Taking a first glance at the tables, we see that the most accurate ones with the lowest ranks, naturally have much lower accuracy values than the worst ones. With the level of the accuracy measures RMSE and MNSE being higher, the differences are naturally much higher for these measures than for the ME and the MAE. The differences are presented by the "absolute difference" in the first row in table 6.10-6.12. Because positive errors can offset negative ones when calculating ME, a larger ME value does not necessarily imply that the errors of the best respondents are small and vice versa for the worst respondents. The standard deviation of the forecast errors of the respondents, presented in the third row of table 6.10, are, however, larger for the worst ten than the best ten, pointing towards larger errors made by the worst ten respondents in terms of ME.

From the discussion in chapter 6.2.2.1 as well as previous literature, we would expect the best forecasters to be located in the beginning of the survey and the worst to be located later. This seems to hold for the most accurate and the least accurate forecasters concerning ME, MAE and RMSE. In terms of ME, the majority of the most accurate ones forecasted only after1990, while most of the least accurate ones forecasted much earlier. The respondents ranked best by RMSE have responded in the end of the survey, with an exception being individual 145. The worst respondents ranked by RMSE are located in the beginning of the survey.

As discussed in section 6.2.2.1 this timely pattern may be explained by the variation in the actual inflation in the different periods of time. However, looking in the fourth row in table 6.10, we see that the opposite is true for the ME measure. The inflation actually has a larger dispersion in the forecasting periods of the best ten compared to the periods of the worst. The pattern is more like we expect it for both MAE and RMSE, with the dispersion in the actual inflation being higher in the forecasting period of the worst ten compared to the best ten. Ranking individuals by MNSE can highlight this pattern further. The second column in table 6.8, presents the period of forecasting of the best respondents in terms of MNSE. Here we do not discover a very obvious timely pattern. However, even though some of the best respondents in terms of MNSE were located in the end of the survey, we have almost no late respondents among the ten worst, seen in the second column in table 6.9. Thus, it seems like the forecasting performance of the respondents were to some degree worse in the beginning of the survey, even when we account for the standard deviation of the inflation. With these results, is seems hard for us to make any conclusions regarding whether the dispersion in the actual inflation can account for why the worst individuals are located mostly in the beginning

of the survey. Even though the dispersion in the actual inflation itself do not make us able to make any conclusions regarding why the accuracy was worse in the beginning, some of the factors highlighted by Croushore (2006) could still have had an impact. For example the oil price shocks and the bad monetary policy in the 1970's and 1980's made the future of the economy more unpredictable, and could alone have made forecasting harder.

The sum of errors in column five in the tables 6.2-6.9 expresses whether the best and the worst forecasters seem to overestimate or underestimate the inflation. For all measures, the average of the sum of errors is positive for the ten most accurate respondents, while the ten least accurate ones have negative errors. This indicates underestimation by the best respondents and overestimation by the worst.[31] Because those ranked worst in terms of ME, MAE and RMSE forecasted in an early period when the inflation level was decreasing, and the best ranked forecasted in a more recent period with a more stable actual inflation, these results are in line with previous mentioned consensus results found by Thomas (1999) and Mehra (2002).

We also take a look at the number of forecasts per respondent. The mean number of forecasts for the worst and the best respondents are presented in the last row in tables 6.10-6.13. For ME there is a large dispersion among the best ten respondents (column six in table 6.2). Their mean of 47.9 forecasts is a bit higher than the consensus mean of 41.91. Also for RMSE and MNSE the number of forecasts among the best ones varies a lot. While the average number of forecasts for the most accurate ones in terms of both MAE and RMSE are lower than the consensus, the forecasters ranked best by MNSE have a higher average number of forecasts. Some of the best respondents in terms of MNSE have more forecasts than the mean consensus. Individual number 65 affects the average a lot, because he or she has predicted the inflation for 99 quarters (column six in table 6.8). For all accuracy measures, the average numbers of forecasts made by the least accurate forecasters are smaller than the average of the most accurate forecasters.[32] An explanation can be that worst did not gain any forecast experience, or they may be "random" respondents, not caring much about the forecasts they make.

[31] The degree of underestimating in terms of the sums of errors by the most accurate ones are naturally lower than the degree of overestimating by the leas accurate ones, because the best ones are better forecasters with lower errors, as shown by the absolute value of the sums of errors in tables 6.10-6.13.
[32] The difference is not very high for the respondents in terms of the RMSE measure.

Table 6.2: General patterns of the ten best respondents in terms of ME.

Respondents ranked by the lowest mean error (ME)	Individual number	Time period	Rank by lowest ME	ME	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
	472	1995q2-2010q4	1	-0.001	-0.071	52	-1.654	1.884	0.538	0.747
	446	1993q2-2008q1	2	0.004	0.283	66	-1.746	2.276	0.527	0.703
	448	1993q2-2008q1	3	0.020	0.020	30	-1.814	1.401	0.627	0.587
	524	2003q1-2010q4	4	0.026	0.721	28	-1.993	1.626	0.641	0.885
	431	1991q1-2010q4	5	0.028	1.699	61	-2.554	2.199	0.572	0.660
	424	1990q4-2010q4	6	-0.029	-0.721	25	-1.849	1.870	0.789	0.681
	145	1974q4-1981q2	7	0.034	0.609	18	-1.549	2.850	1.064	1.570
	65	1974q4-2007q3	8	0.046	4.598	99	-2.436	2.175	2.184	2.142
	31	1974q4-1986q2	9	-0.059	-1.990	34	-4.452	4.269	1.345	2.369
	411	1990q4-2010q4	10	0.059	3.880	66	-1.828	2.581	0.707	0.702

Table 6.3: General patterns of the ten worst respondents in terms of ME.

Respondents ranked by the highest mean error (ME)	Individual number	Time period	Rank by highest ME	ME	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	100	1983q3-1990q1	1	-2.251	-40.510	18	-5.647	0.804	1.884	0.542
	23	1981q3-1986q3	2	-2.196	-35.134	16	-3.771	0.789	1.290	1.104
	5	1981q3-1984q4	3	-1.865	-24.240	13	-2.520	-0.517	1.220	0.911
	47	1975q1-1984q1	4	-1.729	-39.762	23	-22.332	1.183	5.352	2.199
	22	1975q4-1980q4	5	1.702	28.939	17	-0.895	9.000	2.080	1.352
	79	1981q3-1987q4	6	-1.630	-35.856	22	-3.352	-0.100	1.476	0.958
	13	1981q4-1988q3	7	-1.553	-24.847	16	-7.125	-0.054	2.048	0.789
	434	1991q1-1994q3	8	-1.538	-13.845	9	-2.646	0.006	0.828	0.176
	69	1981q3-1989q4	9	-1.484	-28.201	19	-3.468	0.280	1.619	0.871
	68	1981q3-1986q2	10	-1.484	-26.704	18	-2.781	0.063	0.828	1.035

Table 6.4: General patterns of the ten best respondents in terms of MAE.

Respondents ranked by the lowest mean absolute error (MAE)	Individual number	Time period	Rank by lowest MAE	MAE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	531	2005q2-2009q4	1	0.508	-3.375	11	-1.244	0.462	0.442	0.947
	405	1990q3-2007q2	2	0.511	2.513	33	-0.926	1.503	0.455	0.495
	422	1990q4-2010q4	3	0.514	-6.964	29	-1.297	1.155	0.558	0.476
	510	1999q2-2010q4	4	0.558	6.327	46	-1.459	1.068	0.354	0.780
	502	1999q2-2005q1	5	0.579	1.253	16	-1.017	1.481	0.376	0.549
	544	2005q2-2009q2	6	0.581	4.606	15	-0.974	1.491	0.625	0.968
	507	1999q2-2010q4	7	0.585	5.450	42	-1.415	1.521	0.334	0.723
	546	2005q3-2010q4	8	0.593	2.432	21	-1.570	1.561	0.517	0.849
	465	1995q4-2003q1	9	0.595	-1.534	26	-1.482	1.045	0.485	0.423
	500	1999q3-2003q2	10	0.600	1.717	14	-0.808	1.234	0.481	0.374

Table 6.5: General patterns of the ten worst respondents in terms of MAE.

Respondents ranked by the largest mean absolute error (MAE)	Individual number	Time period	Rank by largest MAE	MAE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	148	1974q4-1981q2	1	2.429	16.324	20	-6.076	5.039	2.733	1.382
	100	1983q3-1990q1	2	2.365	-40.510	18	-5.647	0.804	1.884	0.542
	23	1981q3-1986q3	3	2.294	-35.134	16	-3.771	0.789	1.290	1.104
	125	1974q4-1981q2	4	2.271	8.910	27	-9.888	9.052	3.171	1.420
	9	1981q4-1988q4	5	2.215	22.072	24	-2.199	9.177	2.796	0.748
	47	1975q1-1984q1	6	2.203	-39.762	23	-22.332	1.183	5.352	2.199
	93	1974q4-1989q3	7	2.187	1.951	16	-5.806	6.084	2.414	1.438
	31	1974q4-1986q2	8	2.132	-1.990	34	-4.452	4.269	1.345	2.369
	43	1974q4-1988q3	9	2.068	32.297	43	-2.676	6.107	1.273	2.537
	5	1981q3-1984q4	10	1.865	-24.240	13	-2.520	-0.517	1.220	0.911

Table 6.6: General patterns of the ten best respondents in terms of RMSE.

Respondents ranked by the lowest root-mean squared error (RMSE)	Individual number	Time period	Rank by lowest RMSE	RMSE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	472	1995q2-2010q4	1	0.010	-0.071	52	-1.654	1.884	0.538	0.747
	446	1993q2-2010q4	2	0.035	0.283	66	-1.746	2.276	0.527	0.703
	448	1993q2-2008q1	3	0.107	0.588	30	-1.814	1.401	0.627	0.587
	524	2003q1-2010q4	4	0.136	0.721	28	-1.993	1.626	0.641	0.885
	145	1974q4-1981q2	5	0.143	0.609	18	-1.549	2.850	1.064	1.570
	424	1990q4-2010q4	6	0.144	-0.721	25	-1.849	1.870	0.789	0.681
	431	1991q1-2010q4	7	0.218	1.699	61	-2.554	2.199	0.572	0.660
	465	1995q4-2003q1	8	0.301	-1.534	26	-1.482	1.045	0.485	0.423
	502	1999q2-2005q1	9	0.313	1.253	16	-1.017	1.481	0.376	0.549
	549	2006q1-2010q4	10	0.320	1.321	17	-1.771	1.209	0.473	0.805

Table 6.7: General patterns of the ten worst respondents in terms of RMSE.

Respondents ranked by the highest root-mean squared error (RMSE)	Individual number	Time period	Rank by highest RMSE	RMSE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	100	1983q3-1990q1	1	9.548	-40.510	18	-5.647	0.804	1.884	0.542
	60	1974q4-1993q3	2	9.120	-64.489	50	-2.902	1.970	1.259	1.407
	23	1981q3-1986q3	3	8.783	-35.134	16	-3.771	0.789	1.290	1.104
	47	1975q1-1984q1	4	8.291	-39.762	23	-22.332	1.183	5.352	2.199
	35	1981q3-1992q2	5	8.248	-52.811	41	-2.788	0.893	1.156	0.843
	79	1981q3-1987q4	6	7.645	-35.856	22	-3.352	-0.100	1.476	0.958
	66	1981q3-1989q4	7	7.221	-38.212	28	-3.297	1.177	1.140	0.831
	22	1975q4-1980q4	8	7.019	28.939	17	-0.895	9.000	2.080	1.352
	5	1981q3-1984q4	9	6.723	-24.240	13	-2.520	-0.517	1.220	0.911
	69	1981q3-1989q4	10	6.470	-28.201	19	-3.468	0.280	1.619	0.871

Table 6.8: General patterns of the ten best respondents in terms of MNSE.

Respondents ranked by the lowest mean normalized squared error (MNSE)	Individual number	Time period	Rank by lowest MNSE	MNSE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	472	1995q2-2010q4	1	0.011	-0.071	52	-1.654	1.884	0.538	0.747
	446	1993q2-2010q4	2	0.042	0.283	66	-1.746	2.276	0.527	0.703
	145	1974q4-1981q2	3	0.115	0.609	18	-1.549	2.850	1.064	1.570
	448	1993q2-2008q1	4	0.140	0.588	30	-1.814	1.401	0.627	0.587
	524	2003q1-2010q4	5	0.145	0.721	28	-1.993	1.626	0.641	0.885
	424	1990q4-2010q4	6	0.175	-0.721	25	-1.849	1.870	0.789	0.681
	31	1974q4-1986q2	7	0.222	-1.990	34	-4.452	4.269	1.345	2.369
	431	1991q1-2010q4	8	0.268	1.699	61	-2.554	2.199	0.572	0.660
	65	1974q4-2007q3	9	0.316	4.598	99	-2.436	2.175	2.184	2.142
	158	1974q4-1981q2	10	0.322	1.712	21	-2.315	2.543	1.655	1.342

Table 6.9: General patterns of the ten worst respondents in terms of MNSE.

Respondents ranked by the highest mean normalized squared error (MNSE)	Individual number	Time period	Rank by highest MNSE	MNSE	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std real inflation
	100	1983q3-1990q1	1	12.975	-40.510	18	-5.647	0.804	1.884	0.542
	434	1991q1-1994q3	2	10.986	-13.845	9	-2.646	0.006	0.828	0.176
	440	1991q3-1994q2	3	10.114	-8.549	12	-1.088	-0.430	0.209	0.060
	35	1981q3-1992q2	4	8.982	-52.811	41	-2.788	0.893	1.156	0.843
	427	1991q1-1993q4	5	8.459	-11.909	12	-1.385	-0.508	0.299	0.165
	23	1981q3-1986q3	6	8.361	-35.134	16	-3.771	0.789	1.290	1.104
	407	1990q3-2010q4	7	8.353	-50.628	65	-2.823	1.040	0.712	0.565
	66	1981q3-1989q4	8	7.920	-38.212	28	-3.297	1.177	1.140	0.831
	79	1981q3-1987q4	9	7.811	-35.856	22	-3.352	-0.100	1.476	0.958
	60	1974q4-1993q3	10	7.689	-64.489	50	-2.902	1.970	1.259	1.407

Table 6.10-6.13: An overview of the ten best (most accurate) and ten worst (least accurate) respondents in terms of each accuracy measure.

Table 6.10: Overview ME.

Mean values	Best ten	Worst ten	Absolute difference
ME	0.013	-1.403	1.390
Sum of errors	0.903	-24.016	24.919
Std of errors	0.899	1.863	0.963
Std of real inflation	1.105	0.994	0.111
Nmb of forecasts	47.9	17.1	30.80

Table 6.11: Overview MAE.

Mean values	Best ten	Worst ten	Absolute difference
MAE	0.562	2.203	1.640
Sum of errors	1.242	-6.008	7.251
Std of errors	0.463	2.348	1.885
Std of real inflation	0.658	1.465	0.806
Nmb of forecasts	25.3	23.4	1.90

Table 6.12: Overview RMSE.

Mean values	Best ten	Worst ten	Absolute difference
RMSE	0.173	7.907	7.734
Sum of errors	0.415	-33.027	33.442
Std of errors	0.609	1.848	1.238
Std of real inflation	0.761	1.102	0.341
Nmb of forecasts	33.9	24.7	9.20

Table 6.13: Overview MNSE.

Mean values	Best ten	Worst ten	Absolute difference
MNSE	0.175	9.165	8.990
Sum of errors	0.743	-35.194	35.937
Std of errors	0.994	1.025	0.031
Std of real inflation	1.169	0.665	0.504
Nmb of forecasts	43.4	27.3	16.10

6.2.2.3 Some of the best and the worst forecasters overlap

We want to investigate whether the best and the worst forecasters in terms of the different accuracy measures overlap. This could be done by comparing the tables 6.2-6.9. We also summarized these tables by listing the best and the worst ten forecasters in terms of each accuracy measure. These tables are presented in appendix 3.1, table A3.1 and A3.2.

Some respondents overlap, with some being among the best for several accuracy measures, and others being among the worst for several accuracy measures. This pattern diminishes when taking the variation in the actual inflation into account. However, also for the MNSE there are some respondents who overlap with the other measures.

6.2.3 Concluding remarks regarding forecast accuracy

Summarizing our accuracy findings we see that the accuracy measures of the individual forecasters vary a lot.[33] For both the mean and the median consensus forecasts' and for the individual respondents' ME, MAE and RMSE there seems to be a timely pattern. The best ten respondents are located in the end of the survey and the worst are located in the beginning of the survey. This pattern is weaker when we normalize with the standard deviation of the actual inflation using the MNSE measure. Thus, the dispersion in the actual inflation seems to affect the forecast accuracy. Together with the level of the actual inflation, this dispersion is higher in the beginning than later. This made it harder to forecast the inflation, thus the

[33] Some summarizing tables, one with the mean and the median accuracy measure for the consensus mean and median forecasts and for the best ten respondents and the worst ten respondents are presented in the appendix 3.2, table A3.3 and A3.4.

forecasts of the respondents are worse in this early period. Hence, both consensus and the individual forecasters seem to be in accordance with previous research (Croushore, 2006). For the ten most accurate and the least accurate respondents this pattern is weaker, with no late respondents among the ten worst ranked by the MNSE.

The least accurate respondents seem to overestimate the inflation with negative sums of errors. Previous research states overestimation by survey respondents when the inflation level is decreasing (Thomas, 1999; Mehra, 2002). Because the least accurate respondents in our sample are located in the beginning of the survey when the level of inflation was decreasing, this relationship seems to hold also in our data.

In terms of number of forecasts per individual for the best ten and the worst ten, there is a prominent tendency of the worst respondents having responded fewer times than both the best respondents and the consensus. This makes us believe that it is hard to forecast accurately when having made few forecasts, thus having little experience. In appendix 3.2, table A3.4 we present an overview of the number of individual forecasts of the best and the worst ten survey participants.

6.3 The rationality of the inflation forecasts

After examining the accuracy measures of the individuals in chapter 6.2 we have established that individual accuracy differences exist. In this section we go one step further, testing the rationality of the calculated one-year ahead inflation forecasts. We begin testing the rationality of the consensus mean and median forecasts in 6.3.1, continuing with the rationality of individuals in section 6.3.2. We use different rationality tests, we compare the individuals with the consensus and we find which individuals who are the best forecasters in terms of rationality. Again it is important for us to examine the differences between individuals, because we want to understand how the forecasts of individuals are formed in more detail. This is important to see if the assumption of rational expectations made by many macroeconomic models (Mehra, 2002) can be defended.

6.3.1 The consensus forecasts are unbiased, though not strong-form rational

We start using the test of bias described in section 4.2.1, to test if the mean and the median consensus forecasts are biased or not. We regress the forecast error on a constant using the Newey-West method in Stata. The Newey-West variance HAC estimator handles autocorrelation up to and including a lag of m, ignoring autocorrelation in lags larger than m. Because the number of overlapping quarters is five, the errors are MA(4), and we tell the Newey-West estimator to handle autocorrelation in lags up to four quarters (this is explained in more detail in section 5.4.4).

When testing for bias, we test if the forecast errors are zero on average. In other words, we test how accurate the forecasts are. Thus, we will find similar results as in section 6.2. Previous literature presents different results. Early studies give poor results, indicating biasedness that gave a bad reputation to the survey measures (as found by for example Mincer and Zarnowitz (1969), Pearce (1979) and Zarnowitz (1985) and later discussed by Croushore (2006)). However, more recent studies find that the forecasts of survey respondents are unbiased (Gerberding, 2006; Croushore, 2006), indicating that episodes in the early years can be responsible for the earlier bad results. The results of the bias test are shown in table 6.14. We present the p-values of the tests as well as the estimated constant and coefficients for the different variables. With the null of unbiasedness rejected at low p-values, we see that the mean and median consensus forecasts are not biased with p-values over 10 %. Because the forecasts cannot be claimed biased, we call them unbiased, even though the

correct thing statistically would be to say that we could not claim them biased.[34] Hence, the first criterion of rationality is fulfilled, a result in line with the abovementioned recent studies. The constant is negative for both the mean and median consensus forecasts; meaning that the forecast error is negative. The forecasters therefore seem to overestimate the inflation. This result is in line with the accuracy measures of the consensus presented in 6.2.1, and can again be explained by the negative development in actual values, as suggested by for example DeLong (1997) and Thomas (1999)).

We continue with the efficiency tests presented in 4.2.2. Previous literature has found varying results using different efficiency tests, with survey respondents passing some efficiency tests (Mankiw, et al., 2003; Croushore, 2006; Gerberding, 2006), but not tests of strong-form efficiency (Roberts, 1998; Thomas, 1999; Gerberding, 2006).

The first efficiency test, described in 4.2.2.1, tests if historical actual values can help us explain the forecast error by adding lagged actual inflation as an independent variable. Figure 6.12 plots the lagged actual inflation and the consensus forecasts' forecast error against time. They seem to follow a similar pattern, but when performing the test neither the constant nor the lagged inflation are significant (as presented in table 6.14). Thus the mean and median consensus forecasts are weak- form efficient in terms of this test. Hence, they are, as some of the abovementioned literature suggest, to some degree efficient.

Figure 6.12: The lagged actual inflation and the consensus forecast error over time.

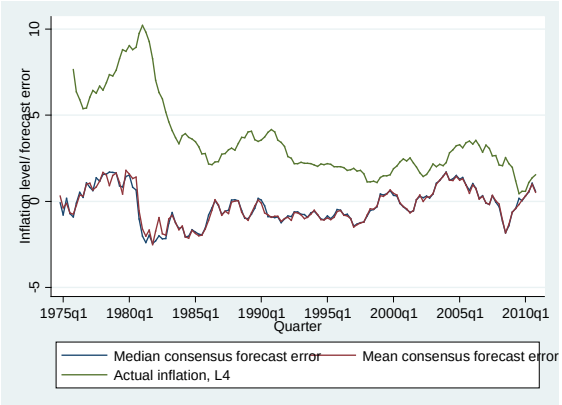
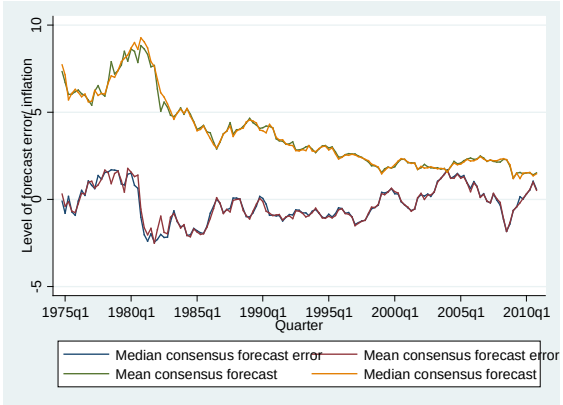


Figure 6.13: The mean and median forecast error and the forecast itself over time. The errors are calculated as the actual inflation minus the forecasted inflation.



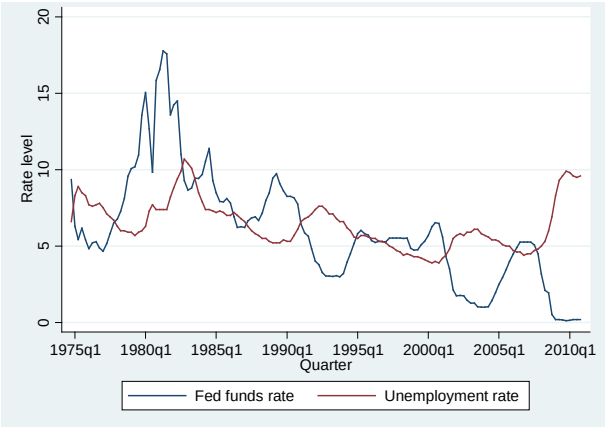
[34] When testing for efficiency, we will in the same manner sometimes say that they are efficient if the null hypothesis is not rejected, even though the most statistically correct thing to do would be to just claim them not unefficient. This is because always using the last option is a bit inconvenient when discussing these issues as much as we do in this paper.

Efficiency test number two (described in section 4.2.2.2) examines if there is information in the forecasts themselves that can explain the forecast error. The mean and median forecast error and the forecast are plotted in figure 6.13. No values in the test results presented in table 6.14 are significant, thus the consensus forecasts are weak-form efficient also according to this test.

Efficiency test three demands the previous forecast errors not to be persistent. Looking at table 6.14, we can reject the joint null hypothesis that the constant and the coefficient of the previous forecast error together are zero. Thus, knowing previous forecast errors would have improved the forecasts of today, and because the forecasters should have knowledge of their previous errors this test fails the rationality criterion. We should, however, keep in mind the fact that the actuals that we use are revised data (a problem discussed in 3.2.1). Hence, the respondents may not know the exact forecast errors when they make their new predictions (Croushore, 2006), and the demand that the respondents should know their previous forecast error could therefore be a bit loosened (Ball & Croushore, 2003).[35]

For us to be able to claim the forecasts of the consensus strong-form rational, their forecasts need to pass efficiency test four. This test demands that the entire information set that the individuals should have knowledge of when predicting, do not correlate with the forecast error. We use the lagged actual inflation and the forecast itself and in addition we include the federal funds interest rate and the unemployment rate in the United States (as discussed in 4.2.2). This will not be the complete information set that all individuals should know of, but we regard it as a good approximation. Figure 6.14, plots the federal funds rate and the

Figure 6.14: The federal funds rate (source: Ecowin) and the unemployment rate (source: Bureau of Labor Statistics) over time.



[35] This argument applies in our later analyses when performing efficiency test three, even though not always mentioned.

unemployment rate over time.

When demanding all coefficients to be zero at the same time, the null of rationality is rejected with a p-value of zero, presented in the last row of table 6.14. Hence, the consensus is not strong-form rational because they do not seem to account for all the information they should have available when predicting the inflation. This result is in line with research by Mankiw et al. (2003) and Gerberding (2006). We also take a look at the estimated coefficients of the regression, interpreting using economic theory. This is done in the same manner as in Mankiw et al. (2003). For the lagged inflation, the coefficient has a positive value. With the thought that a high inflation in one period should be followed by a high inflation in the next period, this positive value indicates that the survey respondents reacted too little to the recent inflation news. Turning to the estimated federal funds rate, the coefficient is negative. A high nominal interest rate indicates contractionary monetary policy by the Federal Reserve, leading to the conclusion that a high nominal interest rate today could signal a lower inflation tomorrow. With the estimated coefficient being negative, high interest rates lead forecasters to predict a too high inflation, hence make negative forecast errors, and again the survey respondents seems to be under- reacting to the new they receive. The estimated coefficient for the unemployment rate is also negative, indicating that the respondents are overestimating the inflation when the unemployment is high. This is because a period of higher unemployment is usually followed by a lower inflation (Gärtner, 2006). Hence, survey participants seems to underestimate the inflation based on recent news of other macroeconomic variables, a result in line with Mankiw et al. (2003) and Ball and Croushore (2003).

We conclude that the consensus is unbiased and, when adding the lagged actual inflation as well as the forecast itself, also weak-form efficient. However, when adding the lagged forecast error and the other actual values that they should be aware of and account for, they do not pass the efficiency test, hence they are not strong-form rational. In accordance with other abovementioned studies, they are therefore quite accurate, even though not strong-form rational.

Table 6.14: Results if the rationality tests of the consensus mean and median forecast for the entire sample.

Rationality tests for consensus			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.275	-0.281
	α=0	0.105	0.113
Efficiency test 1: Lagged actual values	α (constant)	-0.458	-0.353
	β (lagged infl.)	0.051	0.021
	α=β=0	0.120	0.173
Efficiency test 2: Forecasted inflation	α (constant)	-0.247	-0.116
	β (forecasted infl.)	-0.007	-0.044
	α=β=0	0.231	0.268
Efficiency test 3: Lagged forecast error	α (constant)	-0.096	-0.090
	β (forecast error)	0.603	0.617
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	1.296	1.601
	β1 (lagged infl.)	0.433	0.336
	β2 (forecasted infl.)	-0.024	0.115
	β3 (fed funds)	-0.256	-0.293
	β4 (unemployment)	-0.232	-0.272
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

6.3.2 Testing rationality of the individuals

This section tests how rational the inflation forecasts of the individual respondents are. Again we start with unbiasedness, in section 6.3.2.1, before turning to the efficiency tests, section 6.3.2.2. Individual forecasts from the SPF (the ASA/NBER study at the time) have been examined by early studies by Zarnowitz (1985) and Keane and Runkle (1990). While Zarnowitz (1985) found the individual respondents not rational and also biased, Keane and Runkle (1990) found that the individuals were rational. Keane and Runkle (1990) argued that the results against rationality of other literature at the time could be due to the fact of consensus testing. With this result, together with more recent consensus studies finding the consensus unbiased and to some degree efficient, we expect to find that the individual respondents make relatively good forecasts. In addition to test individuals' performances and show how many of them who are rational in terms of each test, we also find the best individuals in terms of each test. These are the ones with the highest probability of being rational. We hope that investigating these best forecasters can give us valuable and perhaps new information about the individuals.

6.3.2.1 Individuals are unbiased

We start by performing the test of bias using Newey-West regressions. When running Newey-West regressions it is important to be aware that they exploit information in lags (as explained in 5.4.4 and 6.3.1). For some individuals the sufficient amount of lags necessary for the Newey-West method to be able to create a consistent covariance matrix is not available. For these individuals we cannot perform the Newey-West regression, thus we will not be able to test their rationality. When running the test of bias there is only one respondent, number 21, who has this problem. When we have excluded this individual we are left with 141 individuals to test biasedness for.

Table 6.15 presents the number of individuals who are unbiased and the part they make up of the total number of individuals. Also presented is the number of individuals better than the consensus.[36] Due to the time and space limit of this paper, we do not present the coefficients from the individual regressions.[37]

When demanding a 1 % significance level to reject the null hypothesis 73.0 % of the respondents cannot be rejected as unbiased forecasters, presented in the first column of table 6.15. Using the "normal" significance level of 5 % this has sunk to 61.0 %, and on a 10 % level, not more than almost half the individuals make forecasts that cannot be claimed biased. The respondents with a p-value larger than the consensus p-value of 10.5 % (presented in table 6.14), thus, the amount of individuals who perform better than the consensus is also a bit more than 50 %. Due to the fact that many former studies have claimed that consensus forecasts are more accurate than most individuals (Zarnowitz, 1984), with one article actually claiming the consensus better than almost all individual forecasters (McNees, 1987), this result is surprising. Hence, individual respondents seem to make good predictions.

With over 60 % of the forecasters passing the test of unbiasedness at a 5% significance level, we conclude that most of the respondents forecasts cannot be rejected unbiased. This result is in line with recent literature of consensus as well as Keane and Runkle's (1990) test of the individuals.

[36] The consensus value we use in this analyse, is the consensus calculated as the mean of all respondents. This goes for the efficiency test analysis as well.
[37] This applies in our later individual analyses, in section 6.4, 6.5 and 6.6 as well.

Table 6.15: An overview of the test of bias of the individual respondents for the sample. "All" is the total number of individuals who we have performed the test on, while the values under "the unbiased" are the number of individuals passing the test/better than the consensus mean forecast's p-value.

Overview unbiasedness:	Number of individuals:		Part being unbiased
	All	The unbiased	
	141		
1% significance		103	0.730
5%significance		86	0.610
10% significance		72	0.511
Better than consensus		72	0.511

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

We have ranked the individuals based on their p-values. The individuals with the highest p-values are the "best" forecasters in terms of unbiasedness, because they have the largest probability of unbiased forecasts. We start examining the accuracy of the best ten individuals in terms of unbiasedness. The accuracy measures of these best forecasters and their ranking in terms of the accuracy measures are presented in table 6.16. Because having unbiased forecasts on average mean that the average forecasts are quite accurate, we expect some

Table 6.16: An overview of the accuracy measures for the ten individuals with the highest p-value of being unbiased.

Bias test: ten best respondents accuracy measures and rankings in terms of accuracy measures	Individual number	ME	Rank ME	MAE	Rank MAE	RMSE	Rank RMSE	MNSE	Rank MNSE
	472	-0.001	1	0.748	29	0.010	1	0.011	1
	446	0.004	2	0.735	25	0.035	2	0.042	2
	448	0.020	3	0.836	51	0.107	3	0.140	4
	145	0.034	7	0.994	75	0.143	5	0.115	3
	31	-0.059	9	2.132	134	0.341	11	0.222	7
	524	0.026	4	0.666	17	0.136	4	0.145	5
	424	-0.029	6	0.766	32	0.144	6	0.175	6
	93	0.122	22	2.187	135	0.488	19	0.407	15
	431	0.028	5	0.630	15	0.218	7	0.268	8
	158	0.082	16	1.164	93	0.374	12	0.322	10

Table 6.17: An overview of general patterns concerning the ten individuals with the highest p-value of being unbiased.

	Individual number	Time period	Rank by highest p	p- value	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
Bias test: Respon-dents ranked by the highest p-value	472	1995q2-2010q4	1	0.995	-0.071	52	-1.654	1.884	0.888	0.747
	446	1993q2-2010q4	2	0.985	0.283	66	-1.746	2.276	0.931	0.703
	448	1993q2-2008q1	3	0.956	0.588	30	-1.814	1.401	0.995	0.587
	145	1974q4-1981q2	4	0.942	0.609	18	-1.549	2.850	1.280	1.570
	31	1974q4-1986q2	5	0.935	-1.990	34	-4.452	4.269	2.483	2.369
	524	2003q1-2010q4	6	0.928	0.721	28	-1.993	1.626	0.846	0.885
	424	1990q4-2010q4	7	0.919	-0.721	25	-1.849	1.870	0.916	0.681
	93	1974q4-1989q3	8	0.874	1.951	16	-5.806	6.084	2.952	1.438
	431	1991q1-2010q4	9	0.872	1.699	61	-2.554	2.199	0.848	0.660
	158	1974q4-1981q2	10	0.853	1.712	21	-2.315	2.543	1.403	1.342

overlapping between these individuals and the best in terms of the accuracy measures. This holds for ME, RMSE and MNSE. For MAE the pattern is not as clear. Because the best ten in terms of unbiasedness are overlapping many of the best ten for most of the accuracy measures, we make the expected conclusion that the individuals with the highest probabilities of having unbiased forecasts are also some of the most accurate forecasters.

In table 6.17 we present some general patterns among those ten best individuals, presented in the same manner as we presented the ten best and worst in terms of the accuracy measures in section 6.2. They have high p-values, all being over 85 %, and when looking at their time periods of forecasting, we see that most of them are late forecasters. This implies that the respondents' forecasts have improved over the years, in line with results presented in 6.2.2, as well as results in previous literature (for example Croushore (2006)). The early respondents among these ten best have large standard deviations of both the actual inflation and of the forecast errors. This indicates that a large variation in the actual inflation tends to give a large variation in the errors, and at the same time a large variation in errors indicates difficulties when making forecasts.

The sum of errors is for most individuals positive, thus the most accurate individuals seem to have been underestimating the actual inflation. This a different conclusion then the consensus conclusion for the individual respondents (section 6.3.1), but in line with the results of the best ten respondents in terms of the different accuracy measures, presented in 6.2.2. The sums of errors are, however, not that large, so this implication is not very strong. The number of inflation forecasts ranges from 16 to 66, with no clear pattern. This leaves it difficult to conclude that gaining forecasting experience by predicting the inflation many times is an advantage for making accurate forecasts.

These best individuals are also ranked in terms of the other rationality tests, presented in a table in appendix 4, table A4.1. There is not much overlapping between these, implying that the ones who have the highest probability of being unbiased not necessarily have a higher probability of being efficient.

6.3.2.2 Efficiency tests

In the previous section we saw that most of the respondents are unbiased. Because of the relatively good recent rationality results of the consensus (Gerberding, 2006; Croushore, 2006) and Keane and Runkle's (1990) rationality results of individual respondents we expect

relatively good results for the efficiency tests also. This section performs efficiency tests on the individual respondents. Again the four tests discussed in 4.2.2 are used.

The results are divided in different issues. First we present the amount of individuals with sufficient observations for us to be able to perform the tests (1), before (2) testing if we can conclude with efficiency or not. Finally we take a look at the ten best individuals in terms of each test (3).

Individuals with sufficient observations for testing

Because of the problem of not having the sufficient amount of observations to be able to run the Newey-West regressions (as explained in 5.4.4), there are some individuals for whom we cannot perform the efficiency tests. This problem arises for individual number 21 when adding lagged actual values, the forecast itself and when adding the information set. For these tests we are left with 141 individuals. For efficiency test three, adding the lagged forecast error, there are more individuals without sufficient observations. This is because we need sufficient observations of the lagged forecast error to be able to run the regression. We have to exclude 26 respondents, leaving us with a sample of 115 respondents for this test.

The majority of individuals do not pass any of the efficiency tests

In this sub-section we present the rationality results of the individual respondents' efficiency tests. Table 6.18 presents the individuals who are efficient based on the different significance levels for all efficiency tests. The p-value of the mean consensus forecast and the fraction of the individuals that performs better than the consensus is also presented.

Looking at table 6.18, we see that efficiency test one, adding the lagged actual inflation, has a mean consensus p-value of 12.0 %. The number of individuals with a p-value larger than 12.0% is 51. In terms of percentages, 36.2 % of all individuals have performed better than the consensus. This is quite a large percentage given that the previous claim that the consensus is much better than almost all individuals (McNees, 1987). At a 5 % significance level the number of respondents that cannot be claimed efficient in this first efficiency test is 68, a bit under 50 % of the entire sample. Hence, the majority of the respondents are not weakly efficient and thus not rational in terms of this test. However, being able to claim almost half the individuals efficient, tells us that the individuals are quite good forecasters.

Adding the forecast itself, in efficiency two presented in table 6.18, we see that 12.8 % of the forecasters have a p- value larger than the consensus value of 23.1 %. At a 5 % significance level, there are only 39 respondents where the joint null is rejected. Hence, only 27.7 % of the individuals are efficient in terms of the forecast itself not explaining some of the forecast errors. The forecast itself is definitely something the forecasters have information about when making forecasts. This conclusion seems to point towards most of the respondents not being efficient and rational.

Turning to the test where we add the lagged forecast error, efficiency test number three (described in section 6.2.2.3) table 6.18 shows that the consensus p-value is zero. When finding the part of the individuals who have performed better than the consensus we want to exclude the individuals who also have p-values very close to zero. We therefore set the p-value limit that they should be better than to 0.0001. Doing this leave us with almost 75 % of the individual respondents being better than the mean consensus forecast, a result tending towards the individuals performing very well compared to the consensus. However, the made assumption that we can set the limit at 0.0001 might give us more individuals better than the consensus than what is really true, contributing to this very high percentages of individuals performing better than the consensus.

For efficiency three 36.5 % of all respondents cannot be rejected efficient on a 5 % significance level. Thus, the lagged forecast error can explain some of the current forecast error for most respondents, and most of the forecasters do not account for their previous forecast errors as we expect them to. The majority can therefore not be claimed weakly rational based on this efficiency test. Again we should keep in mind that the actuals that we use are revised data, and that the exact forecast error therefore cannot be known to the respondents (Croushore, 2006). Even though the majority is not weakly rational, there are many forecasters who seems to be performing well.

For the respondents to be strong-form efficient, and thus strong-form rational, they have to pass the fourth efficiency test, where we add the information set (as described in section 4.2.2.4). The results of this test are also presented in table 6.18. When testing individual respondents the individuals should optimally have one information set each, based on the information they have available and base their forecasts on when responding to the survey.

Table 6.18: The number of individuals who are efficient and rational based on the different efficiency tests, and the part that they make of all respondents. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview efficiency tests for the individual respondents in the whole sample				
	All sample	1 %	5 %	10 %
Efficiency test 1: α=β=0				
Total nmb	141	88	68	60
Part of all		0.624	0.482	0.426
Better than the mean consensus:				
Consensus p-value	0.120			
Nmb	51			
Part of all	0.362			
Efficiency test 2: α=β=0				
Total nmb	141	48	39	25
Part of all		0.340	0.277	0.177
Better than the mean consensus:				
Consensus p-value	0.231			
Nmb	18			
Part of all	0.128			
Efficiency test 3: α=β=0				
Total nmb	115	62	42	32
Part of all		0.539	0.365	0.278
Better than the mean consensus:				
Consensus p-value	0.000			
Nmb	86			
Part of all	0.748			
Efficiency test 4: α=β1=β2=β3=β4=0				
Total nmb	141	11	4	1
Part of all		0.078	0.028	0.007
Better than the mean consensus:				
Consensus p-value	0.000			
Nmb	41			
Part of all	0.291			

Eff. test 1: $A_t - F_t = \alpha + \beta A_{t-4} + \varepsilon_t$

Eff. test 2: $A_t - F_t = \alpha + \beta F_t + \varepsilon_t$

Eff. test 3: $A_t - F_t = \alpha + \beta(A_{t-4} - F_{t-4}) + \varepsilon_t$

Eff. test 4: $A_t - F_t = \alpha + \beta_1 F_t + \beta_2 A_{t-4} + \beta_3 U_t + \beta_4 i_t + \varepsilon_t$

This is, however, hard to retrieve, because we do not have any specific information about the respondents, other than their forecasts and their industry variable (the industry variable is analysed in detail in section 6.4). Therefore, the information set that we use here and expect the respondents to have knowledge about is, except for their own forecasts, the same for all individuals. The new information added are the economic variables discussed in the consensus section, the federal funds interest rate and the unemployment rate in the United States.

Also this test has a consensus p-value of zero, and we demand the ones being better than the consensus to have a p-value higher than 0.0001. The fraction of individuals better than the consensus is here 29.1 %. At a 5% significance level, only 4 out of 141 respondents can be claimed efficient or strong-form rational.

From our results we see that the majority of the individuals do not pass any of the efficiency tests, but in some tests a large fraction of them does. This is in line with literature stating that most strong efficiency tests reject rationality of inflation forecasts, but that many of them still are quite accurate (Gerberding, 2006).

The ten best forecasters in terms of each test

We have ranked the ten respondents with the highest p-values in terms of the joint null hypothesis for each test. In this section we present these individuals. We first take a look at their calculated accuracy measures, presented in tables 6.19-6.22, as well as their rankings in terms of these accuracy measures (1). Then we present some general patterns or the lack of such among these ten best (2). Their rankings in terms of the other efficiency tests are also discussed. In appendix 4.2 we take a look at the four strong-form rational respondents.

The "most" rational ones are not necessarily the most accurate ones

As for the test of bias, we examine the accuracy measures of the ten best individuals in terms of each efficiency test. These are presented in the tables 6.19-6.22.[38]

The calculated accuracy measures for ME, RMSE and MNSE are, for most of the ten best in the different efficiency tests, smaller than the corresponding consensus values, presented in table 6.1, section 6.2.1. This implies that the forecasters are more accurate than the consensus, and is true for all the weak-form efficiency tests. Given that these are the best respondents, this result may not be surprising, but it is still a result against the mentioned claim by McNees (1987) that consensus forecasts are better than almost all individuals. However, when examining MAE we do not find any clear pattern of small or large values for any of three weak efficiency tests. For the strong-form efficiency test, presented in table 6.22, we do not find any clear relationship between the values of the accuracy measures when comparing them with the consensus values. This makes it is hard for us to conclude that the

[38] Again, it seems natural with some overlapping in terms of the best having low rankings of the accuracy measures, because having an average forecast error of zero is one of the criterions in all the efficiency tests. This implication is though not as strong as for the test of bias because we now demand also other variables not being able to explain the error.

best ones in terms of strong-form efficiency, make more accurate forecasts than the other respondents.

General patterns among the best ten are hard to find

Turning to looking for general patterns among the best ten in terms of each efficiency test, we present the individuals' time period of forecasting, the number of forecasts they have made, their sum of errors, etc. These are presented in tables 6.23-6.26.

We expect the p-values of the ten best to be high if a lot of the best individual respondents are efficient. This pattern holds for efficiency test one and three as presented in the p-value column of table 6.23 and 6.26. However, when looking at efficiency test two the p-values of the best ten have a large range, from 53 % to 96 %. Hence there are not many respondents with a very large (if by large we think of p-values above 50 %) probability of being efficient when performing this efficiency test. Looking at efficiency test four there are no high p-values, and only one individual who can be claimed rational based on all significance levels, and four at "normal" 5 % significance level. This is a natural result because the majority of the best ten are not efficient based on this test. The four strong-form rational ones are presented in appendix 4.2.

Examining the best respondents' time period of forecasting, we do not find any clear patterns for most of the tests. For efficiency test number one, table 6.23, there is a pattern of the best respondents being located late in the survey, for test two there are both early and late respondents and for test three there, three out of the four "best" are early respondents. Test four, presented in table 6.26, shows a small tendency of the best forecasters being located in the end of the survey. Looking at all tests together, we cannot conclude that the best forecasters in terms of efficiency are located in any particular time span of the survey. This result is a bit different from our results in terms of accuracy (as discussed in 6.2.2.2), and also a bit different than we expected, because previous literature have found that forecast performance have improved (Croushore, 2006). However, with efficiency test one and four tending towards the best ones being late respondents, there is also in this data, a small tendency towards the expected result being true.

Table 6.19: The accuracy measures of the ten best respondents in terms of efficiency test one, and the rankings of those in terms of the different tests.

Efficiency test 1: ten best respondents accuracy measures and rankings in terms of accuracy measures	Individual number	ME	Rank ME	MAE	Rank MAE	RMSE	Rank RMSE	MNSE	Rank MNSE
	429	-0.063	12	0.637	16	0.447	15	0.577	20
	34	0.089	17	1.032	80	0.418	13	0.337	11
	424	-0.029	6	0.766	32	0.144	6	0.175	6
	502	0.078	15	0.579	5	0.313	9	0.423	16
	543	0.234	35	0.745	28	0.992	31	0.997	32
	541	0.254	40	0.804	41	1.017	32	1.065	36
	65	0.046	8	0.813	44	0.462	17	0.316	9
	528	0.327	51	0.739	26	1.600	47	1.687	46
	524	0.026	4	0.666	17	0.136	4	0.145	5
	527	0.286	46	0.679	18	1.402	44	1.491	44

Table 6.20: The accuracy measures of the ten best respondents in terms of efficiency test two, and the rankings of those in terms of the different tests.

Efficiency test 2: ten best respondents accuracy measures and rankings in terms of accuracy measures	Individual number	ME	Rank ME	MAE	Rank MAE	RMSE	Rank RMSE	MNSE	Rank MNSE
	98	0.104	18	1.248	103	0.579	23	0.390	14
	145	0.034	7	0.994	75	0.143	5	0.115	3
	124	0.254	39	0.831	49	0.949	30	0.824	28
	78	-0.274	43	1.147	91	1.129	36	0.750	26
	546	0.116	21	0.593	8	0.531	22	0.576	19
	535	-0.240	36	0.715	23	1.124	35	1.188	39
	549	0.078	14	0.612	12	0.320	10	0.357	12
	507	0.130	25	0.585	7	0.841	27	0.989	31
	34	0.089	17	1.032	80	0.418	13	0.337	11
	510	0.138	27	0.558	4	0.933	35	1.056	35

Table 6.21: The accuracy measures of the ten best respondents in terms of efficiency test three, and the rankings of those in terms of the different tests.

Efficiency test 3: ten best respondents accuracy measures and rankings in terms of accuracy measures	Individual number	ME	Rank ME	MAE	Rank MAE	RMSE	Rank RMSE	MNSE	Rank MNSE
	144	0.298	47	1.124	89	1.334	42	1.089	38
	34	0.089	17	1.032	80	0.418	13	0.337	11
	465	-0.059	11	0.595	9	0.301	8	0.463	17
	125	0.330	52	2.271	138	1.715	51	1.439	43
	543	0.234	35	0.745	28	0.992	31	0.997	32
	549	0.078	14	0.612	12	0.320	10	0.357	12
	527	0.286	46	0.679	18	1.402	44	1.491	44
	548	0.326	50	0.625	13	1.495	45	1.622	45
	500	0.123	23	0.600	10	0.459	16	0.750	27
	485	-0.259	41	0.924	68	1.347	43	1.733	48

Table 6.22: The accuracy measures of the ten best respondents in terms of efficiency test four, and the rankings of those in terms of the different tests.

Efficiency test 4: ten best respondents accuracy measures and rankings in terms of accuracy measures	Individual number	ME	Rank ME	MAE	Rank MAE	RMSE	Rank RMSE	MNSE	Rank MNSE
	488	-0.642	83	0.776	36	2.487	68	3.969	97
	462	-0.780	99	0.840	54	2.467	66	4.499	106
	432	-0.619	82	0.752	30	1.637	48	4.127	100
	502	0.078	15	0.579	5	0.313	9	0.423	16
	60	-1.290	128	1.395	110	9.120	140	7.689	132
	39	-1.107	124	1.498	120	3.990	100	3.231	79
	42	-0.274	45	0.858	57	1.025	34	0.717	25
	498	0.815	100	0.840	55	4.076	102	5.715	120
	520	0.345	54	0.785	38	1.888	56	2.059	57
	145	0.034	7	0.994	75	0.143	5	0.115	3

Wanting to check if the best forecasters are overestimating or underestimating, we take a look at the sum of errors, the fifth column of tables 6.23-6.26. As for the test of bias, the sums of errors are positive for efficiency test one and three, implying some degree of underestimation by the best forecasters. However, when performing efficiency test four, the sum of errors are actually mostly negative, implying overestimation. For efficiency test two the pattern is unclear, leaving us with an overall inconclusive pattern.

We investigate whether there seems to be a positive effect from answering to the survey several times. Such a pattern could imply that the forecasters "learn" and improve if they forecast the inflation many times. However, for most of the tests the opposite is true. For test one and two, the differences in the numbers of responded forecasts are large, but the majority has fewer forecasts than the consensus, presented in table 6.1, section 6.2.1. For efficiency test three and four, the highest number of forecasts of the best ten is far below the consensus value of about 42. These numbers implies that it does not seem to be an advantage or a "learning- effect" coming from forecasting several times. Making a lot of forecasts could rather seem to be a disadvantage, making it hard get among the "top ten" respondents.[39] A previous study by Lamont (2002) might be able to explain this relationship. Lamont (2002) finds that older and more established forecasters tend to give more radical forecasts, and are therefore less accurate, maybe because of a wish of getting a reputation. This finding is, however, not completely in line with our result in 6.2.2.2, where the more accurate forecasters had made inflation forecasts several times more than the least accurate ones.

We also rank the best individuals of the different efficiency tests in terms of the other rationality tests, finding not much overlapping. Thus, there does not seem to be a clear pattern in terms of the ones ranked best by one test also being among the best ones in the other tests. These tables are presented in the appendix 4.1, in table A4.2-A4.5.

[39] It could be that some of these best forecasters simply did not respond to the survey during the most challenging times and episodes, leaving them with better and more accurate forecasts. This could be an interesting topic to dig into, but with the time limit and space issues of this paper, we do not examine this issue further.

Table 6.23: General patterns of the ten best respondents in terms of efficiency test one.

Efficiency test 1: Respon-dents ranked by the highest p-value in terms of Ho: α=β=0	Individual number	Time period	Rank by highest p	p- value	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
	429	1991q1-2008q4	1	0.997	-3.164	50	-1.497	1.678	0.798	0.658
	34	1974q4-1981q3	2	0.988	1.962	22	-2.514	2.024	1.269	1.428
	424	1990q4-2010q4	3	0.973	-0.721	25	-1.849	1.870	0.916	0.675
	502	1999q2-2005q1	4	0.971	1.253	16	-1.017	1.481	0.744	0.617
	543	2005q2-2009q3	5	0.939	4.210	18	-1.721	1.634	0.909	0.990
	541	2005q2-2010q4	6	0.922	4.069	16	-1.846	1.721	0.980	0.882
	65	1974q4-2007q3	7	0.890	4.598	99	-2.436	2.175	0.967	2.237
	528	2005q1-2010q4	8	0.871	7.839	24	-3.070	1.769	0.954	0.900
	524	2003q1-2010q4	9	0.868	0.721	28	-1.993	1.626	0.846	0.902
	527	2004q3-2010q4	10	0.846	6.867	24	-2.867	1.333	0.927	0.937

Table 6.24: General patterns of the ten best respondents in terms of efficiency test two.

Efficiency test 2: Respon-dents ranked by the highest p-value in terms of Ho: α=β=0	Individual number	Time period	Rank by highest p	p- value	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
	98	1974q4-1986q1	1	0.962	3.226	31	-2.534	2.429	1.481	2.413
	145	1975q5-1981q2	2	0.863	0.609	18	-1.549	2.850	1.280	1.420
	124	1974q4-1979q4	3	0.822	3.551	14	-0.989	1.902	0.967	1.303
	78	1974q4-1989q1	4	0.811	-4.653	17	-2.940	1.965	1.399	2.372
	546	2005q3-2010q4	5	0.750	2.432	21	-1.570	1.561	0.789	0.841
	535	2005q2-2010q4	6	0.710	-5.271	22	-2.070	1.432	0.894	0.882
	549	2006q1-2010q4	7	0.680	1.321	17	-1.771	1.209	0.777	0.805
	507	1999q2-2010q4	8	0.639	5.450	42	-1.415	1.521	0.710	0.774
	34	1974q4-1981q3	9	0.617	1.962	22	-2.514	2.024	1.269	1.428
	510	1999q2-2010q4	10	0.530	6.327	46	-1.459	1.068	0.629	0.774

Table 6.25: General patterns of the ten best respondents in terms of efficiency test three.

Efficiency test 3: Respon-dents ranked by the highest p-value in terms of Ho: α=β=0	Individual number	Time period	Rank by highest p	p- value	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
	144	1975q1-1981q2	1	0.972	5.965	20	-3.043	2.099	1.397	1.448
	34	1974q3-1981q3	2	0.971	1.962	22	-2.514	2.024	1.269	1.428
	465	1995q4-2003q1	3	0.956	-1.534	26	-1.482	1.045	0.710	0.407
	125	1974q4-1981q2	4	0.951	8.910	27	-9.888	9.052	3.353	1.420
	543	2005q2-2009q3	5	0.943	4.210	18	-1.721	1.634	0.909	0.990
	549	2006q1-2010q4	6	0.933	1.321	17	-1.771	1.209	0.777	0.805
	527	2004q3-2010q4	7	0.919	6.867	24	-2.867	1.333	0.927	0.937
	548	2005q3-2010q4	8	0.882	6.850	21	-1.688	1.190	0.703	0.841
	500	1999q3-2003q2	9	0.878	1.717	14	-0.808	1.234	0.657	0.361
	485	1995q2-2004q1	10	0.875	-6.997	27	-2.109	1.705	1.078	0.552

Table 6.26: General patterns of the ten best respondents in terms of efficiency test four.

Efficiency test 4: Respon-dents ranked by the highest p-value in terms of Ho: α=β1=β2=β3=β4=β5=0	Individual number	Time period	Rank by highest p	p- value	Sum of errors	Number of forecasts	Minimum error	Maximum error	Std of errors	Std of real inflation
	488	1995q2-2001q4	1	0.293	-9.632	15	-1.269	0.553	0.542	0.402
	462	1995q2-1999q1	2	0,095*	-7.802	10	-1.613	0.300	0.604	0.306
	432	1991q1-1994q3	3	0,058*	-4.332	7	-1.522	0.464	0.618	0.157
	502	1999q2-2005q1	4	0,056*	1.253	16	-1.017	1.481	0.744	0.617
	60	1974q4-1993q3	5	0,049**	-64.489	50	-2.902	1.970	0.878	2.370
	39	1974q4-1984q2	6	0,041**	-14.385	13	-2.887	1.390	1.330	2.077
	42	1975q4-1984q2	7	0,034**	-3.835	14	-2.595	1.588	1.173	2.235
	498	1998q4-2010q4	8	0,030**	20.382	25	-0.235	2.000	0.500	0.766
	520	2002q1-2010q4	9	0,020**	10.343	30	-1.720	1.951	0.942	0.853
	145	1974q4-1981q2	10	0,016**	0.609	18	-1.549	2.850	1.280	1.420

6.3.3 Concluding remarks on the rationality of forecasts

Different conclusions can be drawn from the performed tests. For the consensus mean and median forecasts we cannot reject the null of rationality in the test of bias and in efficiency test one and two. This indicates unbiasedness and weak-form efficiency, thus weak-form rationality for the consensus. However, in terms of efficiency test three and four, the null of efficiency is rejected, and we can therefore not claim the forecasters rational. This is in accordance with other previously mentioned literature (Gerberding (2006), stating that survey participants are quite accurate, though not strong-form rational.

Turning to the individual respondents, the conclusions are a bit altered. At a significance level of 5 %, over 50 % of the respondents do not have biased forecasts, thus the first criterion of rationality is fulfilled. However, turning to the efficiency tests, neither of the weak-form efficiency tests nor the strong-form test gives us that more than half of the respondents are efficient. When demanding strong-form rationality, only 2.8 % of all individuals pass the test. Even for the best ten individuals the null of strong-form rationality is rejected at a 10 % and a 5 % significance level for most of these best respondents, and we conclude with most of the respondents not being strong-form rational. This result seems to coincide with Zarnowitz (1984) and McNees (1987) who states that the consensus is better than the individual forecasters. However, looking at the fraction of individual respondents who are rational, efficient and better than the consensus in each test, we find many respondents who have a relatively good forecasting performance.

The ten respondents with the largest p-values for all tests are not overlapping much in terms of being among the best ten in several tests. This implies different individual conclusions in the different tests. Drawing any conclusions regarding whether some respondents always are better than others is therefore difficult. However, for most rationality tests, except the fourth, the best respondents tend to have low accuracy measures. This indicates that the efficient ones are also accurate, as we would expect. However, because this does not hold for the strong-form rationality test it is hard for us to be too sturdy stating this as a fact.

Other patterns worth mentioning are that for the test of bias and the efficiency test of adding the lagged actual forecast (efficiency test one) most of the best respondents are located in the end of the survey, a result coinciding with the accuracy results in 6.2.2.2. This pattern is not as clear in the other tests. The standard deviation of the actual inflation in the forecasting

period of those individuals can explain some of this, because the variation in the inflation level is higher for the forecasters in the beginning of the survey.

The pattern of underestimating the inflation among the best also seems valid for most tests. However, this pattern is reversed for the strong-form rationality test, leaving us with difficulties drawing conclusions. Another pattern is that the best respondents have relatively few numbers of forecasts. Hence, gaining experience by answering to the survey many times does not seem to improve the forecasts. This finding is supported by Lamont (2002).

It is hard to make strong conclusions about the individuals' forecasting performance. The fact that the different tests give us different "best" individuals, and that the patterns of these are not clear, shows us that individual differences exists. Such differences can be important to document, because of the stated fact that many economic models do not take differences and disagreement between individuals into account, assuming that individuals make rational forecasts (Mehra, 2002; Mankiw, et al., 2003).

6.4 Examining differences between industries

So far we have not used the information regarding the industry the individuals are employed in. We have information about the industry variables from the second quarter of 1990 (Chew & Price, 2008). In this section we investigate if there are differences between the individuals employed in the different industries using data from the SPF. To our knowledge this has not been done before.

Most empirical studies examining professional survey participants assume that forecasters will produce and present their best estimates, because they are paid to do so. Because consensus forecasts, which are stated better than individual respondents by many are available to the public, it seems strange that firms still produces rather inaccurate forecasts (Laster, et al., 1999). Ehrbeck and Waldmann (1996) suggest that this is because the forecasters are overconfident, and therefore put too little weight on the available consensus forecasts. Others (Laster, et al., 1999; Lamont, 2002; Gerberding, 2006) suggest that some forecasters pursue other goals than accuracy when making predictions. One example is to have a strategic goal of getting a reputation, and therefore publish rather extreme forecasts. Examining if there are differences between the performances of individuals with different industry variables in the SPF may give us a hint of whether such differences exist between the different firms in this survey.

A study by Laster et al. (1999) examines differences between survey participants employed in different firms, investigating the professional survey participants in the "Blue Chip Economic Indicators." They suggest that some survey respondents deliberately bias their forecasts, for example because of a wish of getting publicity. They sort the respondents in six different industry categories, finding significant differences between them. The thought is that the firms who are expected to make forecast accuracy a high priority should produce forecasts closest to the consensus. In their findings industrial corporations were closest to the consensus. Such firms value accuracy more than publicity because they use forecasts extensively for internal planning. They also found that individual forecaster firms together with consulting and advisory firms seem to make forecasts far from the consensus, probably because they value publicity and media attention in addition to accuracy (a result also supported by Lamont (1999)). Banks and econometric forecasting firms have more intermediate results compared to the consensus. Even though one might expect these to value accuracy highly, econometric forecasting firms may have an additional pressure to outshine

competitors, and banks may be able to attract new clients with favourable publicity. Hence, these two types of firms may value publicity in addition to accuracy.

An important feature of the SPF is that the forecasters are anonymous. Strategic incentives should therefore be non-existing. However, many of the individuals responding to the SPF also publish forecasts to the public, and the thought of them giving different forecasts to the public and the survey seems unlikely. Hence, strategic incentives can exist for the SPF forecasters as well. However, it is likely that some of the firms do not publish their forecasts. Therefore the strategic incentives of the forecasters in the SPF may be smaller than in other surveys.

There are three different industry variables in the SPF. Individuals with industry variable one are employed in a financial service provider firm, while individuals with industry variable two are employed in nonfinancial service firms. The third industry variable contains those where they cannot decide whether the individual is employed in industry one or two. A financial service provider is involved in insurance, investment and commercial banking, payment services, hedge and mutual funds, asset management or in association of financial service providers. If employed in a nonfinancial service provider, one is employed at a university, a manufacturing firm, a forecasting firm, an investment advisor firm, a research firm or a consultant firm (Chew & Price, 2008; Federal Reserve Bank of Philadelphia, 2008).

Wanting to find out if Laster et al.'s (1999) results hold in the SPF, we have to compare their different industries with the industry variables of the SPF. With banks and asset management firms being located in the first category in the SPF, we expect industry one to have relatively accurate results, which not deviate too much from the consensus. Industry two contains both manufacturing firms who should be relatively accurate, and consultant firms who are the least accurate in Laster et al.'s study. Hence, it seems hard to make any conclusions regarding what result to expect for industry two. However, with consultant and investment advisor firms, as well as forecasting firms being located in this category, we believe that the strategic incentives of those may be stronger than the accuracy incentives of the manufacturing firms. Hence, we conclude with believing that strategic incentives of publicity and attention are stronger for industry two.

When dealing with the industry variables it is important to be aware that some of the respondents may have changed their employment over time, implying that the industry

variable can have changed (Chew & Price, 2008). However, in section 5.4.3 we find that this is not a big problem in our data. When performing tests, the individuals with different industry variables will at each time belong to the industry where they at that time are classified.[40]

Before performing tests, we start presenting how the forecasts of the respondents in the different industries and the number of individuals in each industry classification have evolved through time. We also present the consensus mean and median forecasts in this period after the second quarter of 1990. The consensus mean and median forecasts are presented in figure 6.15, while figure 6.16 presents the mean forecasts of the individuals in the different industries and the actual inflation. The deviations from the actual inflation follow to some degree similar patterns for the different industries, but there are some distinctions. For industry three it looks like we have one (or more) outlier in the beginning of the 1990s. Because industry three includes fewer forecasters than the other industries (as seen in figure 6.17), this outlier could exist because of one single respondents mistake.[41] Figure 6.17 shows the number of survey participants categorized in the different industries each quarter. In the beginning the majority of individuals were employed in financial service provider firms, labelled with industry variable number one. Later the individuals employed in nonfinancial servicer provider firms, have outnumbered this "industry".[42] When analysing it is important to be aware that we do not have very many individuals employed in each industry to examine. This makes the statistical tests weaker, and makes it hard to draw very sturdy conclusions.

We show the dispersion in the forecasts, presented by the standard deviation of the forecasts, in the three industries in figure 6.19. We see that there are differences in the forecasts for the different industries when comparing them with eachother as well as when comparing them with the whole sample. The highest dispersion in the data is for industry three. The dispersion in this industry is zero up until about 1995, because there was only one individual located in this industry before 1995. If only comparing the standard deviation of forecasts for industry

[40] We should also be aware of the fact that when the panel changes over time, the amount of individuals in the different categories might also change; meaning that if we want to look at different sub-periods, there can be a lot of differences in the number of individuals available for analyses in each industry.
[41] Because of the unknown nature of the firms of individuals with individual number three, as well as the number of respondents in this category being very low, we will not discuss this industry as much as the other two.
[42] From here on we will sometimes call the financial service provider firms industry one, the nonfinancial service provider firms industry two and the unknown ones industry three.

one and two, presented in table 6.18, we see that there are timely differences between them. It is, however, hard to say that one of the industries have more dispersed forecasts than the other one.

Figure 6.15: The consensus mean and median forecasts and the actual inflation from 1990q2. The values in a given quarter are the forecast given of the next year inflation that quarter and the actual inflation for the next year.

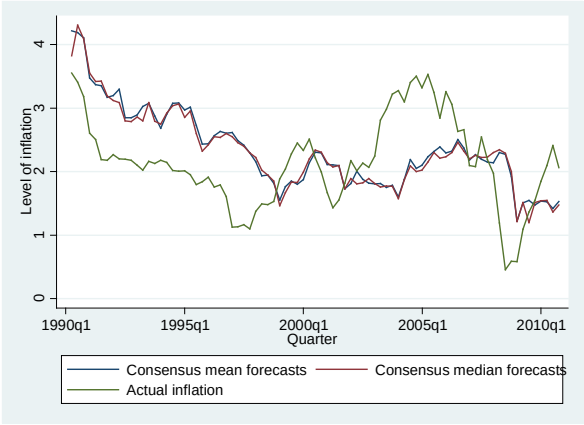


Figure 6.16: The mean forecast of the individuals in each industry and the actual inflation over time. The values in a given quarter are the forecast given of the next year inflation that quarter and the actual inflation for the next year.

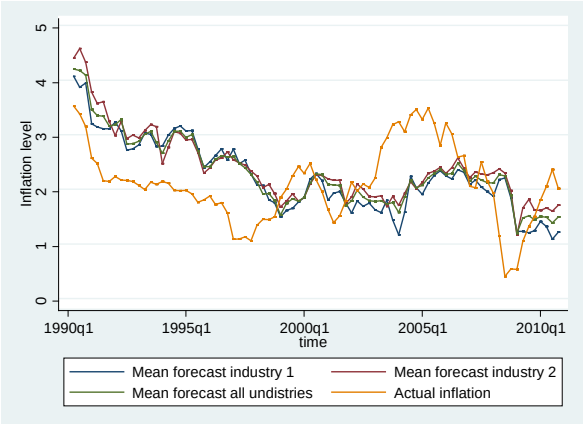


Figure 6.17: The number of individuals employed in each industry over time.

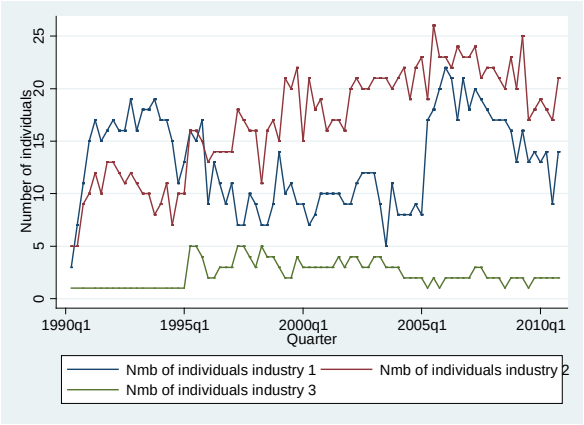


Figure 6.18: The standard deviation of the forecasts in industry one and two over time.

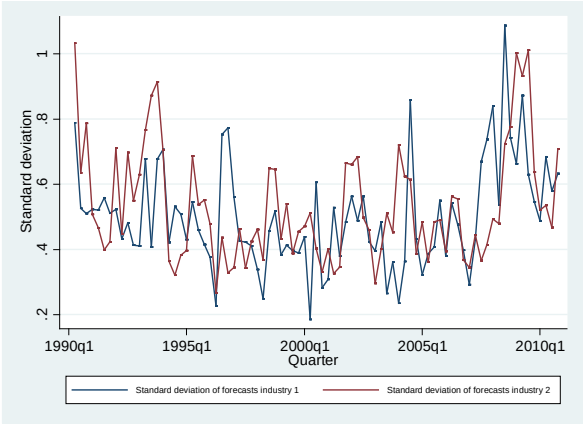
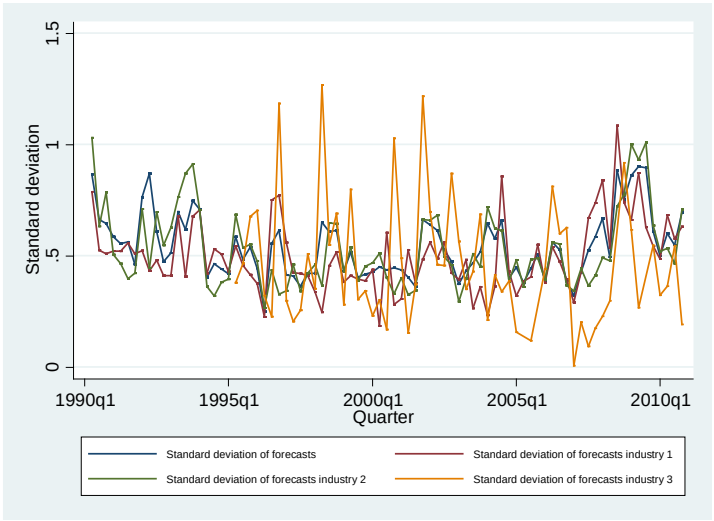


Figure 6.19: The standard deviation of the forecasts in each industry over time.



When examining the forecast performance of the individuals employed in the different industries, we start comparing the accuracy measures in section 6.4.1. When testing the rationality of the individuals, we divide the sample in three by their industry variables. Then we perform the tests of rationality on these three samples' consensus forecasts as well as the individuals.[43] Because the sample starts in the second quarter of 1990, the high inflation period in the beginning of the survey (presented in figure 6.1), and the period of the Volcker disinflation, is now excluded. To be able to compare the different industries with the performance of the total sample in the same period properly, we first examine the total sample of forecasts in this period, presented in 6.4.2. Afterwards we test individuals employed in industry number one in section 6.4.3, continuing with the other industry variables in 6.4.4 and 6.4.5.

6.4.1 Comparing accuracy measures in the different industries

We calculate the accuracy measures in the different industries, looking for differences in the forecasting performance of these individuals that we should be aware of.[44] The accuracy measures for the consensus of the sample starting in the second quarter of 1990 as well as for the different industries are presented in table 6.27. Looking at ME we see that the pattern of overestimating the inflation level is present also in this sample.[45] This pattern is also valid for the different industries. One observed difference is that the consensus of industry one has a lower ME value than the others as well as the total sample in this period. This means that they are more accurate than the other industries based on this measure.

In terms of MAE the different industries have very similar values, though industry two and three have somewhat smaller values than industry one. Thus, the different industries have very similar average errors when we do not allow for positive errors to offset negative ones. For RMSE, which punishes larger errors more than errors, the consensus of industry one has lower values than the others. This indicates that these are more accurate in terms of not having many large errors. The difference between industry one and two is quite large,

[43] However, we do not examine the individuals as thoroughly as we did previously, by finding the ten best, their rankings in terms of accuracy measures, etc, This is due to the time and space of this paper, in addition to the samples now being more limited with fewer respondents and a smaller time span.
[44] Because of space issues, we only do this for the consensus.
[45] This is the same result as the one we got for the entire sample, presented in 6.2.1. However, now the Volcker disinflation period is excluded, hence this cannot be explained by the decreasing inflation in that special period. However, looking at figure 6.15, the inflation seems to be decreasing in this period also, indicating that the explanation proposed by Thomas (1999) might still be valid.

Table 6.27: The accuracy measures of the mean and median consensus forecasts in the period from1990q2, and for the consensus of the different industries.

Overview consensus accuracy measures for the industries and the total sample this period								
	All sample from 1990q2		Industry 1		Industry 2		Industry 3	
Measure	Mean	Median	Mean	Median	Mean	Median	Mean	Median
ME	-0.198	-0.173	-0.099	-0.092	-0.261	-0.224	-0.300	-0.322
MAE	0.729	0.728	0.751	0.740	0.744	0.728	0.732	0.730
RMSE	1.804	1.578	0.903	0.840	2.378	2.045	2.734	2.931
MNSE	2.159	1.888	1.104	1.027	2.796	2.404	3.344	3.585

with RMSE values of 0.90 and 2.38 for the two, respectively.[46]

We conclude that the consensus of the respondents in industry one is a bit better than the respondents in the other industries, at least in terms of the RMSE. Assuming that individuals employed in industry two have stronger strategic incentives than those employed in industry one, this finding in accordance with previous results presented by Laster et al. (1999).

6.4.2 Testing rationality of forecasts after the Philadelphia Fed took over the survey

We start testing the rationality of all forecasts from the second quarter of 1990, the period where we have information about the industry variable of the individuals. This allows us to compare the forecasts of the different industries with the forecasts of the total sample in the same period.[47] Important being aware of is that both the Volcker disinflation period, as well as the early period where the inflation was very high (in the beginning of the survey, as presented in figure 6.1) ended before 1990, and is therefore excluded from this sample. We expect that the exclusion of these unstable periods will affect the results in this sample.

As mentioned in the beginning of chapter six, a lot of previous research examines the performance of forecasts in different sub-periods, for example during booms and recessions. We have chosen not to examine such issues alone. However, our choice to examine the period after the second quarter of 1990 makes us able to at the same time compare this period with the whole sample, examining whether the accuracy and rationality of the forecasts have changed. In this later period they have had better control of the identification number of individuals, less gaps and missing values in the respondents' forecasts, and they have data

[46] Because the time period of forecasting is the same for the respondents in each industry, the MNSE measure, taking account of the standard deviation of the actual inflation, naturally shows the same patterns as RMSE.
[47] Because we only have three industry variables in the SPF, with the third one being very small, the analysis will mostly focus on comparing industry one and two. If one industry is better than the whole sample, the other one will naturally be worse, and vice versa. With this regard testing the whole sample may not seem very important. However, because of other interesting aspects in terms of examining this shorter sample (as explained in 6.4.2), we choose to analyse this sample alone.

that are more consistent than before.[48] These factors make this time period especially interesting when we compare the performance of forecasts in this period with the performance in whole sample. For the abovementioned reasons we can expect the forecasters to be more rational in this time period. Such a result will at the same time be in line with previous literature, finding that the forecasting performance have been better in a more recent time period (Croushore, 2006; Gerberding, 2006). The results of the consensus mean and median forecasts are presented in section 6.4.2.1, while the results of the individual respondents are discussed in 6.4.2.2.

6.4.2.1 The probability of the consensus being unbiased is larger in this shorter sample

The consensus mean and median forecasts of the period starting in the second quarter of 1990 are pictured in figure 6.15. The results of the tests are presented in table 6.28. The p-values of each test and the coefficients corresponding to the variables included in the regression are presented. Both consensus values pass the test of bias, with the probability values for the null hypothesis of unbiasedness being 28.9 % and 35.6 % for the mean and the median consensus, respectively. The conclusion of unbiasedness is the same as for the whole sample, but the p-values are higher, being 10.5 % and 11.3 % for the mean and median consensus of the whole sample, presented in table 6.14. Hence, the probability of the forecasts being unbiased have increased. This indicates more accurate respondents in this later period, a result in line with our expectations. The constant terms, thus, the errors are negative. This indicates overestimation, and is also a result in line with the results of the whole sample.

Turning to the efficiency tests, the respondents' consensus forecasts passes the first test where we add the lagged actual inflation. For the other efficiency tests the consensus fail the efficiency criterion. Hence, the consensus forecasts are not efficient and not strong-form rational in this period. This result is quite similar to the results of the consensus of the entire time period, indicating that the forecasts were not better in this later time period. The consensus passes one less test than the overall consensus, indicating that the forecasts have actually worsened. This result is a bit surprising, because the inflation level has been more stable in this period. However, the fact that the probability of unbiasedness is higher in this later period makes it hard for us to state that the forecasts have gotten worse.

[48]These issues are presented in 5.4.3, 5.4.2 and in a discussion of the consistency of the forecasts is presented in appendix 2.5.

Table 6.28: The results of the rationality tests of the mean and median consensus forecasts in the period after 1990q1.

Rationality tests for the consensus of the period where we have industry variables			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.198	-0.173
	α=0	0.289	0.356
Efficiency test 1: Lagged actual values	α (constant)	-0.493	-0.468
	β (lagged infl.)	0.152	0.151
	α=β=0	0.435	0.506
Efficiency test 2: Forecasted inflation	α (constant)	1.513	1.560
	β (forecasted infl.)	-0.730	-0.747
	α=β=0	0***	0***
Efficiency test 3: Lagged forecast error	α (constant)	-0.017	-0.011
	β (forecast error)	0.631	0.621
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	1.943	1.912
	β1 (lagged infl.)	0.560	0.585
	β2 (forecasted infl.)	-1.129	-1.207
	β3 (fed funds)	-0.095	-0.078
	β4 (unemployment)	-0.071	-0.057
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

When interpreting the coefficients we again find similar results as the whole sample. The lagged inflation coefficient is positive while the coefficients of the federal funds rate and the unemployment rate is negative, with all three indicating that the forecasters underreact to new information (the intuition behind this is explained in section 6.3.1).

6.4.2.2 The forecasts of the individual respondents are relatively similar

Table 6.29 presents the results of the individual respondents' forecasts. The majority of the individual respondents who pass each test for the different significance levels and the part they make of the sample are presented. Again, we are not able to run the test for all the individuals because some have missing lagged values needed to perform the Newey-West method.[49] The majority of the survey participants pass the test of bias; hence we can present them unbiased. The fraction of all individuals passing the test of bias on a 5 % significance level is 62.2%, very similar to the corresponding part of the whole sample, at 61.0 %

[49] This is because of the lack of sufficient forecasts, as explained in 5.4.4 and 6.3.2.2. This goes for the tests on the individual respondents in the different industries as well. The number of individuals that we are able to perform the tests for is shown in the tables where the tests results are presented.

Table 6.29: The results of the rationality tests of the individual respondents in the shorter sample starting in 1990q2. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests when the Philadelphia Fed conducted the survey				
	All sample	1 %	5 %	10 %
Test of bias: $\alpha=0$				
Total nmb	82	59	51	45
Part of all		0.720	0.622	0.549
Efficiency test 1: $\alpha=\beta=0$				
Total nmb	72	51	42	40
Part of all		0.708	0.583	0.556
Efficiency test 2: $\alpha=\beta=0$				
Total nmb	82	18	16	11
Part of all		0.220	0.195	0.134
Efficiency test 3: $\alpha=\beta=0$				
Total nmb	67	39	26	18
Part of all		0.582	0.388	0.269
Efficiency test 4: $\alpha=\beta1=\beta2=\beta3=\beta4=0$				
Total nmb	72	8	4	1
Part of all		0.111	0.056	0.014

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1F_t+\beta_2A_{t-4}+\beta_3U_t+\beta_4i_t+\varepsilon_t$

presented in table 6.15. Hence, we cannot claim any differences in terms of biasedness between the individual respondents in this limited sample and the entire sample.

Turning to the efficiency tests focusing on the 5 % significance level in table 6.29, we see that the majority of the individual respondents pass the first test of efficiency, but not the other three. This finding is the same as for the consensus of the whole sample. Because the majority of the individuals do not pass the efficiency criterion, they cannot be claimed rational. With the majority of the survey participants passing one efficiency test in this period, compared to none in the whole sample, the forecasts seems to have improved a bit when excluding the earliest data. This is in accordance with previous research by Gerberding (2006) and Croushore (2006).

6.4.3 Industry variable 1- financial service provider

This section examines the individuals employed in industry one. We start presenting the consensus mean and median forecasts in section 6.4.3.1. In section 6.4.3.2 we examine the performance of the individual respondents.

6.4.3.1 The consensus of industry one has a high probability of unbiasedness

Table 6.30 presents the results of the consensus of industry one together with the coefficients related to each added variable.[50] The consensus of industry one is unbiased with high p-values over 60 % for both the mean and the median. Hence, they are unbiased, as was the total sample this period. The p-value of the consensus in industry one is, however, much higher than for the whole sample starting in the second quarter of 1990. It is thus more probable that the consensus of industry one is unbiased, indicating better forecasting performance in this industry. This is in line with the accuracy measures presented for the

Table 6.30: Results for the rationality tests of the consensus mean and median forecasts of industry one.

Rationality tests for consensus, Industry 1			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.099	-0.092
	α=0	0.616	0.639
Efficiency test 1: Lagged actual values	α (constant)	-0.380	-0.424
	β (lagged infl.)	0.142	0.165
	α=β=0	0.748	0.694
Efficiency test 2: Forecasted inflation	α (constant)	1.712	1.734
	β (forecasted infl.)	-0.806	-0.816
	α=β=0	0***	0***
Efficiency test 3: Lagged forecast error	α (constant)	0.020	0.015
	β (forecast error)	0.644	0.642
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	1.882	1.897
	β1 (lagged infl.)	0.573	0.572
	β2 (forecasted infl.)	-1.175	-1.189
	β3 (fed funds)	-0.791	-0.075
	β4 (unemployment	-0.061	-0.060
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

[50] The coefficients indicate the same patterns as for the whole sample and for the shorter sample starting in the second quarter of 1990. The coefficients of the lagged inflation, the federal funds rate and the unemployment rate all indicate that forecasters underreact to new information.

industries and the whole sample in table 6.27. This is also in accordance with previous literature when we assume that the individuals in industry one have stronger accuracy incentives than strategic incentives compared with the other industries (Laster, et al., 1999).[51]

The consensus forecasts are weakly efficient based on efficiency test one. However, based on the other efficiency tests we can reject the null of efficiency. Because they do not pass all rationality tests, we conclude that the consensus of industry one is not strong-form rational, showing the same statistical results of all tests as the total sample this period. However, even though not strong-form rational, the respondents are quite accurate, because of the very high probability of unbiasedness. These results are also in line with previous literature (Croushore, 2006; Gerberding, 2006).

6.4.3.2 The majority of individuals in industry one are unbiased

Table 6.31 summarizes all tests performed on the individual respondents in industry one. Again we present the number of individuals passing the tests on different significance level. We have a total of 40 individuals employed in industry one that we are able to run the test of bias for.

When testing for bias 52.5 % of all individuals cannot be claimed biased at a 5 % significance level. We conclude that most of the individuals in this industry do not have biased forecasts. This is the same conclusion as for the consensus forecasts of this industry as well as the individual respondents of the whole sample. Even though the p-values for unbiasedness for the consensus was larger for this industry than in the entire time period, the part of all individuals being unbiased is smaller.

Performing efficiency test one, 52.8 % of the respondents in industry one are efficient at a 5% significance level. Hence, we conclude that most individuals are efficient and weakly rational based on this test. Again the result is similar as for the whole sample this period. Turning to efficiency test two, only 17.5 % of the total number of respondents can be claimed efficient. When adding the lagged forecast error in efficiency test three, we have to limit our sample to 33 respondents. None of the significance levels gives us over 50 % of the

[51] Because we only have three industries in the SPF, with industry three being undefined and including few respondents, our examination involves mostly a comparison between industry one and two. Laster et al. (1999) compares the industries and the consensus in this period, with the assumption that the consensus is the best, but since we only have (almost) two industries to compare, one of them will naturally be better than the consensus forecasts.

individuals being efficient. 42.4 % of the respondents are efficient on the "normal" 5 % significance level.

Because the respondents in this industry are employed in financial service provider firms, we expect them to have knowledge about the important actual economic values, such as the federal funds interest rate and the unemployment level that we add when running efficiency test four. However, at a 5 % significance level only two individuals cannot be claimed efficient. In percentages this leaves us with 5.7 % of the individuals being strong-form rational, hence the majority of the respondents are clearly not efficient in the financial service provider industry.

6.4.3.3 Concluding remarks of industry one

The overall conclusions regarding the consensus mean and median forecasts and the individual respondents in industry one is that the forecasters cannot be claimed strong-form rational. It is only the efficiency test where we add the lagged actual inflation, efficiency test one, where more than 50 % of the respondents and the consensus forecasts can be said to be

Table 6.31: The results of the rationality tests performed on the individual respondents employed in industry one. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests of individuals in industry 1: Financial service provider firms				
	All sample	1 %	5 %	10 %
Test of bias: $\alpha=0$				
Total nmb	40	23	21	20
Part of all		0.575	0.525	0.500
Efficiency test 1: $\alpha=\beta=0$				
Total nmb	36	22	19	18
Part of all		0.611	0.528	0.500
Efficiency test 2: $\alpha=\beta=0$				
Total nmb	40	9	7	6
Part of all		0.225	0.175	0.150
Efficiency test 3: $\alpha=\beta=0$				
Total nmb	33	16	14	12
Part of all		0.485	0.424	0.364
Efficiency test 4: $\alpha=\beta1=\beta2=\beta3=\beta4=0$				
Total nmb	35	2	2	1
Part of all		0.057	0.057	0.029

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1F_t+\beta_2A_{t-4}+\beta_3U_t+\beta_4i_t+\varepsilon_t$

efficient and weakly rational. In the other tests the consensus and the majority of respondents fail the rationality criterion. These results are the same as for the total sample this same period, indicating that the individuals in this industry do not particularly stand out when comparing with the other industries in the sample.

6.4.4 Industry 2- nonfinancial service provider

This section examines the respondents employed in the second industry, in nonfinancial service provider firms. Because this industry includes consultant firms, forecaster firms and university employed workers, we expect them to have stronger strategic incentives. Hence, we think that they value publicity and media attention more than respondents employed in industry one. Therefore they might value accuracy less than industry one, hence give less accurate forecasts. Again, the consensus forecasts are first examined, in section 6.4.4.1, before investigating the individual respondents, in 6.4.4.2.

6.4.4.1 The consensus of industry are unbiased, but with lower probability than industry one

The test results of the consensus mean and median forecasts are presented in table 6.32.[52] Also the consensus for industry two can be claimed unbiased with p-values of 16.4 % and 22.7 % for the mean and the median consensus, respectively. These values are smaller than those for both the total sample this period and industry one, indicating less accurate forecasts among industry two. This is in line with the accuracy measures presented in table 6.27, and with the hypothesis that these respondents have more strategic incentives and therefore make less accurate forecasts.

The consensus of the individuals in industry two also pass efficiency test one, but the null of rationality is rejected based on the other efficiency tests. Hence, we reject strong-form rationality for the consensus of industry two, again making the same conclusion as for industry one. However, the p-values are also lower than the corresponding ones for industry one, indicating a lower probability of the forecasts of individuals in industry two being efficient. Hence, it seems that the strategic incentives we believe to be stronger in this industry have resulted in them trying to make forecasts that stand out. This implies that they are taking larger risks, increasing the probability of making mistakes.

[52] Again the coefficients indicate that the forecasters to underreact to new information (the intuition is explained in section 6.3.1.

Table 6.32: Rationality tests for the consensus mean and median forecast of industry two.

Rationality tests for consensus, Industry 2			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.261	-0.224
	α=0	0.164	0.227
Efficiency test 1: Lagged actual values	α (constant)	-0.598	-0.609
	β (lagged infl.)	0.177	0.197
	α=β=0	0.237	0.282
Efficiency test 2: Forecasted inflation	α (constant)	1.484	1.502
	β (forecasted infl.)	-0.725	-0.728
	α=β=0	0***	0***
Efficiency test 3: Lagged forecast error	α (constant)	-0.035	-0.028
	β (forecast error)	0.620	0.607
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	1.947	1.959
	β1 (lagged infl.)	0.562	0.568
	β2 (forecasted infl.	-1.139	-1.171
	β3 (fed funds)	-0.094	-0.089
	β4 (unemployment	-0.068	-0.064
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

6.4.4.2 The majority of individual respondents in industry two are unbiased and efficient in terms of efficiency test one

Table 6.33 summarizes our test results of the individual respondents in industry two. The number of individuals employed in the nonfinancial service providing industry is 42. Testing for bias at a 5 % significance level we see that 66.7 % of all respondents employed in the industry cannot be claimed biased. Hence, most respondents in industry two are unbiased, fulfilling the first criterion for rationality. This result is somewhat different than for the consensus forecasts, with a larger majority of the individuals being unbiased in industry two than in industry one. Hence, the presumed strategic incentives of them wanting to make forecasts that stand out do not seem to have given the individuals more biased forecasts.

When testing for efficiency, we again start adding the lagged actual inflation. A bit more than half of the respondents are unbiased on a 5 % significance level. Hence, we conclude that most individuals are efficient, and weakly rational, in the sense that they do take the previous actual inflation into account when making forecasts. This result is very similar to the corresponding result of industry one.

Efficiency test two draws us to a different conclusion. Only 26.2 % of the individuals are efficient at a 5 % significance level. The fraction of efficient respondents is a bit higher than for industry one, otherwise the results are very similar. Because the forecast itself should not be able to explain some of the forecast error, the respondents cannot be said to be weakly rational based on this test. This result might be a bit surprising; especially if a lot the individual forecasters in this industry are employed in forecaster firms, who therefore should be very aware of their own forecasts. The reason for this result can be the mentioned strategic incentives, implying that other goals than making the most accurate forecast as possible exist.

The results when adding the previous forecast error are quite similar. 27 % of the respondents in this industry are efficient on a 5 % significance level; hence the majority is not efficient. Even though more than half 56.8 % of the respondents are efficient on a 10 % level, we conclude that the majority of respondents employed in the nonfinancial service providing industry are not weakly rational in terms of this test.

The results of efficiency test four give even weaker results. Again, we would expect the forecasters to be very much aware of the development of other economic variables. Looking

Table 6.33: Results for the rationality tests performed on the individuals employed in industry two. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests of individuals in industry 2: Financial service provider firms				
	All sample	1 %	5 %	10 %
Test of bias: $\alpha=0$				
Total nmb	42	33	28	24
Part of all		0.786	0.667	0.571
Efficiency test 1: $\alpha=\beta=0$				
Total nmb	38	26	20	18
Part of all		0.684	0.526	0.474
Efficiency test 2: $\alpha=\beta=0$				
Total nmb	42	12	11	7
Part of all		0.286	0.262	0.167
Efficiency test 3: $\alpha=\beta=0$				
Total nmb	37	21	10	5
Part of all		0.568	0.270	0.135
Efficiency test 4: $\alpha=\beta1=\beta2=\beta3=\beta4=0$				
Total nmb	38	6	2	0
Part of all		0.158	0.053	0.000

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1F_t+\beta_2A_{t-4}+\beta_3U_t+\beta_4i_t+\varepsilon_t$

at the joint null hypothesis there are, however, only two individuals who are efficient and strong-form rational at a 5 % significance level. In percentages the number is 5.3 % of all respondents, almost the exact same as for industry one.

6.4.4.3 Concluding remarks for industry two

We conclude that the individual respondents employed in the nonfinancial providing industry are unbiased, but cannot be claimed rational based on three of the efficiency tests. This goes for both the consensus and the majority of the individual respondents, leaving us with the same conclusion as for industry one and the total consensus. Hence, there are not many differences between the two industries in terms of rationality of the individual forecasters.

6.4.5 Industry 3- unknown

In the third industry there are very few individuals to examine. However, we do perform the tests on this sample as well, presenting the result of the consensus in table 6.34, and for the individuals in table 6.35. With the sample being very limited with only seven respondents located in this category, and for some tests even fewer due to the problems with the lags in the Newey-West regression (explained in 5.4.4) we should not put too much weight on this analysis.

The results for the consensus forecasts in table 6.34 show that the consensus of the individuals in industry three are unbiased and weakly efficient based on efficiency test one. In efficiency test two and four, efficiency is rejected at all significance levels. For efficiency test three we cannot reject efficiency if demanding 1 % significance level for rejecting the null hypothesis.

Turning to the tests on the individual respondents in table 6.35, the test of bias shows us that five out of seven, or 71.4 % of all respondents are unbiased. Hence, the majority of respondents in also this industry are unbiased. Performing the efficiency tests, 80 %, of the individuals are efficient when we add the lagged actual inflation, performing efficiency test one. Adding the forecast itself leaves us with only one, or 14.3 % of all individuals efficient, while the same number for efficiency test three is 80 %. Hence, they are efficient when adding the forecast error, but not when adding the forecast itself. This conclusion seems rather odd, but might be explained by the constrained sample that we examine. Adding the information set in efficiency test four leaves us with no individuals we can claim efficient, hence no respondents that are strongly rational at a 5 % significance level.

Table 6.34: Result of the rationality tests of the consensus mean and median forecast of industry 3.

Rationality tests for consensus, Industry 3			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.315	-0.338
	α=0	0.146	0.118
Efficiency test 1: Lagged actual values	α (constant)	-0.151	-0.243
	β (lagged infl.)	-0.044	-0.012
	α=β=0	0.454	0.361
Efficiency test 2: Forecasted inflation	α (constant)	1.675	1.681
	β (forecasted infl.)	-0.816	-0.820
	α=β=0	0***	0***
Efficiency test 3: Lagged forecast error	α (constant)	-0.103	-0.109
	β (forecast error)	0.420	0.443
	α=β=0	0.023**	0.016**
Efficiency test 4: Information set	α (constant)	2.121	2.084
	β1 (lagged infl.)	0.485	0.495
	β2 (forecasted infl.)	-0.945	-0.959
	β3 (fed funds)	-0.141	-0.136
	β4 (unemployment)	-0.119	-0.114
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

Table 6.35: Results of the rationality tests performed on the individuals employed in "industry" three. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests of individuals in industry 3: unknown categorized employment				
	All sample	1 %	5 %	10 %
Test of bias: α=0				
Total nmb	7	6	5	4
Part of all		0.857	0.714	0.571
Efficiency test 1: α=β=0				
Total nmb	5	5	4	4
Part of all		1.000	0.800	0.800
Efficiency test 2: α=β=0				
Total nmb	7	1	1	1
Part of all		0.143	0.143	0.143
Efficiency test 3: α=β=0				
Total nmb	5	5	4	2
Part of all		1.000	0.800	0.400
Efficiency test 4: α=β1=β2=β3=β4=0				
Total nmb	5	1	0	0
Part of all		0.200	0.000	0.000

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

The conclusions regarding the individuals with an "inconclusive" industry are that the majority and the consensus are unbiased and weakly efficient based on efficiency test one, the same findings as for the total sample this period and the other industries. The individual respondents are also rational when we add the forecast error, even though the consensus forecasts only passes this test on a 1 % significance level. When demanding strong-form rationality, the consensus forecasts as well as no of the individuals pass the criterion of efficiency.

6.4.6 Concluding remarks regarding the industries

Examining the tests performed on the different industries, there are few distinct differences to be found. The consensus' results are quite similar for all three industries and for the total sample this period. They all have unbiased forecasts that are efficient when performing looking at efficiency test one. They are, however, not efficient based on the other rationality tests. The fact that the results are so similar may not be very surprising, given that they are all professionals, who may have the same background in terms of education, etc. Hence, they have many of the same premises for making predictions. Another explanation can be that the strategic incentives of the forecasters in the SPF may be smaller than in other surveys, because some of them may not publish their forecasts.

One difference we find worth mentioning is the fact that the consensus mean and median forecasts of industry one have a larger probability of being unbiased than the consensus of industry two. This implies that the strategic incentives of getting publicity and attention of the individuals employed in industry two may exist. This result is in line with the results found by Laster et al. (1999).

We summarize our findings for the individual respondents in the different industries as well as the total sample in table A4.8 in appendix 4, listing the number of individuals in each industry for whom we cannot reject the null hypothesis of rationality. For industry one and two, as well as for the total sample we conclude with unbiasedness and efficiency for the majority of individuals in terms of efficiency test one. For the other tests most of the respondents in both these industries as well as the total period sample cannot be claimed efficient and rational. The number of efficient respondents when we add the lagged forecast error in efficiency test three is a bit higher for industry number one, but when we add the forecast itself in efficiency test two, it is the other way around. Hence, we cannot conclude that the respondents employed in one industry are better in than the other industry.

Even though the consensus mean and median forecasts of industry one have a larger probability of being unbiased, a larger fraction of the individuals in industry two can be claimed unbiased. Hence, it seems as though most of the individuals in industry two are more accurate, but that there are some who are worse than industry one, making their consensus forecasts' values lower. A possible explanation can be that the strategic incentives of getting attention, making forecasts that stand out, may be strong for some of the individuals in industry two. Therefore a minority may take larger risks and make larger mistakes, thus making the consensus values inaccurate.[53] Industry two includes individuals employed in manufacturing firms who probably have no strategic incentives and individuals employed in consulting and forecaster firms, who may have large strategic incentives. Thus, this explanation seems valid, and is also in accordance with the paper by Laster et al. (1999).

Looking at the third industry variable, we also conclude with unbiasedness and weak-form efficiency based on efficiency test one. We can also claim efficiency when we add the lagged forecast error. But because of the weakness of this analysis due to the very small number of individuals, it is hard for us to make any strong conclusions regarding industry number three. Therefore we cannot claim them different than the ones employed in the two other industries either.

[53] This should imply that the median consensus of industry two are better than the mean, an argument that is confirmed by looking at table 6.34, with the mean p-value being 16.4 % and the median being 22.7 %. However, the median consensus p-value is still lower than the consensus values for industry one.

6.5 The Volcker disinflation period

Paul Volcker was appointed chairman of the Board of Governors of the Federal Reserve Board in August 1979. At this time the inflation was as high as 11%, but within the next three years, Volcker reduced the inflation rate to 4% by using contractionary monetary policy. As stated by Mankiw et al. (2003), this change in policy makes it interesting to study inflation expectations in this period. This section examines accuracy and rationality of survey respondents during the Volcker disinflation, in addition to investigating the forecast performance of the survey when this period is excluded.

Previous papers have studied the forecasting performance of individuals during different sub-periods, finding that certain events have affected the forecasting performance (Mehra, 2002; Giordani & Söderlind, 2002). However, we do not have knowledge of many, at least not very recent papers examining the Volcker disinflation period alone. One exception is Mehra (2002), who found that excluding the Volcker disinflation period did not alter the results of the mean forecasts of the Livingston and Michigan surveys. However, the SPF forecasters seemed to perform better when the Volcker disinflation period was excluded. We test if the results of Mehra (2002) hold, and while Mehra (2002) looked at mean forecasts, we also focus on individual respondents. We do not have much knowledge of other studies investigating the performance of individual respondents during the Volcker disinflation. One exception is the mentioned paper by Mankiw et al. (2003), which finds that forecasters in the Michigan survey adjusted slowly to the disinflation. Hence, we hope to gain more knowledge of the forecast behaviour of individuals through this study.

The mean forecasts and the actual inflation are presented in figure 6.20. When Volcker gained his position the inflation was very high, and it peaked at about 10 % in 1980. During his first years in this position the inflation fell a lot. In the start of 1980 the inflation was high and the average forecasters underestimated the inflation. However, when the inflation started to fall the forecasts did not manage to “keep track” with the inflation. Hence, the average forecasts overestimated the inflation during this period. This pattern continued until around 1985.

Figure 6.21 presents the mean forecast of the SPF against the actual inflation in the “falling inflation period.” To picture the development we present the period 1979-1985 as the Volcker disinflation period, even though Volcker did not gain his chairman position before

Figure 6.20: The mean forecast and the actual inflation from 1974q4- 2010q4. The values in a given quarter are the forecast given of the next year inflation that quarter and the actual inflation for the next year.

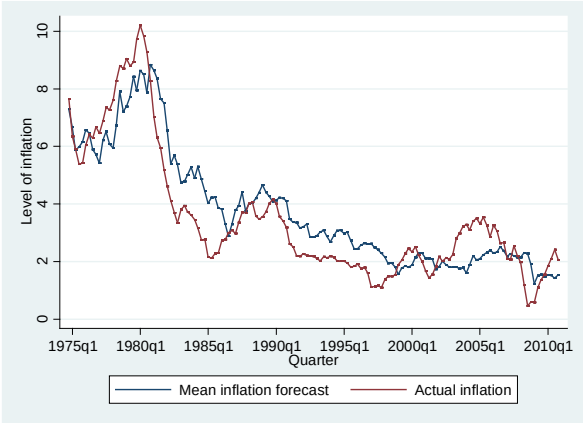
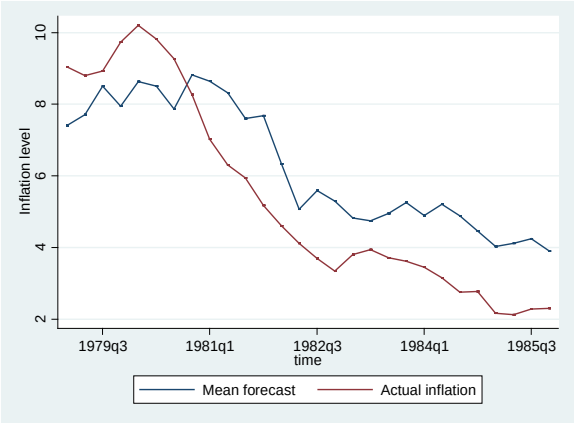


Figure 6.21: The mean forecast from the SPF and the actual inflation in the Volcker disinflation period, from 1979-1982. The values in a given quarter are the forecast given of the next year inflation that quarter and the actual inflation for the next year.



august 1979 and held this position until august 1987. When looking at the figure the forecast performance seems to have been relatively poor during this period. To examine if and how this period affected the forecasts, we start taking a look at the different accuracy measures in section 6.5.1. Then we turn to testing rationality of the forecasts in the Volcker disinflation period, section 6.5.2. In section 6.5.3 we examine the total the time span without this period.

6.5.1 The forecast accuracy is worse during the Volcker disinflation

To get an overview of whether the Volcker disinflation did affect the performance of the forecasts we calculate the accuracy measures for the consensus mean and median forecasts when we exclude the Volcker disinflation as well as for the Volcker disinflation period only. These measures are presented in table 6.36, together with the consensus of the whole sample.

Because of Mehra's (2002) findings we expect the accuracy of the consensus to be better when we exclude the Volcker disinflation period. The reasoning is that when a central bank changes their monetary policy, the previously established credible inflation "target" is altered. If the public have not discovered the new target level they may continue to expect a higher inflation than the realized one, thereby making systematic forecast errors (Thomas,

Table 6.36: The accuracy measures for the consensus mean and median forecasts in the whole sample, in the sample where the Volcker disinflation period is excluded and in the Volcker disinflation period only (1979-1985).

Accuracy measures for the consensus						
	The entire sample		Excluding the Volcker disinflation		Only the Volcker disinflation	
	Mean	Median	Mean	Median	Mean	Median
ME	-0.275	-0.280	-0.126	-0.093	-0.897	-1.066
MAE	0.865	0.884	0.699	0.708	1.557	1.624
RMSE	3.311	3.378	1.366	1.006	4.747	5.640
MNSE	2.216	2.260	1.006	0.741	2.875	3.416
Standard deviation	2.014	2.075	1.842	1.842	2.727	2.727

1999; Mehra, 2002).

From the accuracy measures presented in table 6.36, we see that the sample where the Volcker disinflation is excluded performs better than the sample that only includes the Volcker disinflation. It is also better than the total sample where the Volcker disinflation is not excluded. The ME is negative for all samples, indicating that consensus forecasts overestimate the inflation. This overestimation is, however, much higher during the Volcker disinflation than when we exclude this period. The mean error value of the mean consensus is -0.90 during the disinflation, compared to -0.13 when this period is excluded. Hence, some of the overestimation of the whole sample can have originated from this period with a decreasing inflation. This is in line with studies by DeLong (1997) and Thomas (1999), who suggests overestimation by survey participants when the actual inflation is declining. The explanation is probably that the forecasters did not see the decrease in inflation coming as fast and severe as it did when Volcker was appointed chairman of the Board of Governors of the Federal Reserve (Mehra, 2002; Giordani & Söderlind, 2002).

The RMSE measure (row three in table 6.36), which punishes large errors more than small ones, have much higher values for the consensus forecasts during the Volcker disinflation period than when this period is excluded. Hence, the sum of errors is larger, and there may be several very large errors in the Volcker disinflation period. Even when accounting for the larger dispersion in the actual inflation in this period, presented by the standard deviation of the actual inflation when we calculate the MNSE measure, the accuracy is worse during the Volcker disinflation period.[54] This indicates that even when we account for the changing level of the actual inflation in this period, the forecasters performed worse, a somewhat surprising result. Thomas (1999) finds that regime shifts can cause systematic errors in certain period, even when agents are fully rational. If we think of the disinflation as such a regime shift, this theory may explain this result.

6.5.2 The rationality of forecasts during the Volcker disinflation period

This section tests the rationality of the forecasts during the disinflation period. Using forecasts from the Volcker disinflation period only, we limit our dataset to fewer individuals than before. We have a total of 55 individuals in our sample of individuals during this period. However, due to the demands of the Newey-West method, we cannot test the rationality of all

[54] The Volcker disinflation period has larger MNSE values, presented in the fourth row in table 6.36.

these individuals using all efficiency test (explained in 5.4.4). The number of individuals that we have performed the tests for are shown together with the test results in table 6.38. We start presenting the results for the consensus in section 6.5.2.1 before continuing with the individual respondents in 6.5.2.2.

6.5.2.1 The consensus forecasts are biased and irrational during the Volcker disinflation

The results for the consensus are presented in table 6.37, showing the p-values for the different rationality tests, as well as the coefficients belonging to each variable. While the consensus mean and median forecasts for the whole sample could be claimed unbiased, the results for the consensus in the Volcker disinflation period seem to be more biased. This is not very surprising, considering the pattern in figure 6.21 and previous results by for example Mehra (2002). Both the mean and the median consensus forecasts can be claimed biased if we demand a 10 % significance level for rejecting the null. However, on a 5% significance level only the test of the median consensus forecasts rejects the null of unbiasedness. Both the mean and median consensus forecasts cannot be rejected unbiased on a 1 % significance level. Because the forecast performance in this period looks very poor in figure 6.21, it is somewhat surprising that the p-value of unbiasedness is not smaller. Even though we cannot claim the forecasts unbiased, the fact that we cannot reject unbiasedness at a 1 % significance level, and for the mean consensus not on a 5 % significance level either, is a bit surprising.

All efficiency tests have p-values of zero when using forecasts from the disinflation period only. The results of these tests are presented in table 6.37. Thus, as expected, the mean and median consensus forecasts during the Volcker disinflation period are not rational. The estimated coefficients for the lagged inflation are positive for both the mean and the median consensus in efficiency test one, while it is positive for the mean and negative for the median in efficiency test four. A positive estimated lagged inflation coefficient indicates that the respondents of the survey did not react enough to news about the past inflation, when thinking that a high inflation in one period should be followed by a high inflation in the next period (Mankiw, et al., 2003). Hence, most estimated coefficients for inflation are in line with the fact that the respondents did not manage to keep track of the disinflation in this period (Giordani & Söderlind, 2002).[55]

[55] The coefficients regarding the federal funds rate and the unemployment rate are again negative, indicating that the forecasters underreact to new information.

Table 6.37: Results of the rationality tests for the consensus mean and median forecasts during the Volcker disinflation period (1979-1985).

Rationality tests for consensus during the Volcker disinflation			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.897	-1.066
	α=0	0.083*	0.040**
Efficiency test 1: Lagged actual values	α (constant)	-2.333	-2.104
	β (lagged infl.)	0.183	0.112
	α=β=0	0***	0***
Efficiency test 2: Forecasted inflation	α (constant)	-3.774	-3.175
	β (forecasted infl.)	0.459	0.350
	α=β=0	0***	0***
Efficiency test 3: Lagged forecast error	α (constant)	-0.966	-1.111
	β (forecast error)	0.389	0.353
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	-0.613	-1.105
	β1 (lagged infl.)	0.281	-0.240
	β2 (forecasted infl.)	0.399	0.950
	β3 (fed funds)	-0.337	-0.309
	β4 (unemployment)	-0.971	-0.152
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

6.5.2.2 The majority of individual respondents are not rational during the Volcker disinflation, though a rather high fraction of them are

The results of the rationality tests performed for the individual respondents who responded to the survey during the Volcker disinflation period are presented in table 6.38.

Starting with the test of bias, we see that we can reject the null hypothesis of unbiasedness for the majority of individuals at all significance levels. At a 5 % level 41.8 % of the individuals are unbiased. Hence the individual forecasters have biased forecasts in this period. This result deviates from the results of the whole sample, indicating less rationality in the Volcker disinflation period. The fact that the forecasters perform worse during the Volcker period is expected. However, finding that we cannot claim biasedness for as many as 41.8 % of the individuals is somewhat surprising. Hence, the individuals did not perform very badly even though the inflation were on its way down more than they expected it to.

Continuing with the efficiency tests, the majority of the individuals fail the efficiency criterion for all tests. The closest proportion of individuals passing one of the tests at a 5 % significance level is 20 % of all individuals in efficiency test one and four.

Table 6.38: Results of the rationality tests performed for the individuals who forecasted during the Volcker disinflation period (1979-1985). The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests for the Volcker disinflation period: 1979-1985				
	All sample	1 %	5 %	10 %
Test of bias: α=0				
Total nmb	55	27	23	20
Part of all		0.491	0.418	0.364
Efficiency test 1: α=β=0				
Total nmb	41	6	2	2
Part of all		0.146	0.049	0.049
Efficiency test 2: α=β=0				
Total nmb	40	9	8	7
Part of all		0.225	0.200	0.175
Efficiency test 3: α=β=0				
Total nmb	26	6	3	2
Part of all		0.231	0.115	0.077
Efficiency test 4: α=β1=β2=β3=β4=0				
Total nmb	40	13	8	5
Part of all		0.325	0.200	0.125

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

6.5.3 The rationality of forecasts when the Volcker disinflation period is excluded

This section tests the rationality of the rest of the sample, the whole period excluding the Volcker disinflation period. Hence, we investigate whether the forecasters' performances are different when this "extraordinary" period is excluded from the sample. Section 6.5.3.2 examines the consensus forecasts, while section 6.5.3.2 examines the individual forecasts.

6.5.3.1 The consensus is unbiased, but not strong-form rational when excluding the disinflation period

Table 6.39 presents the results for the consensus when the Volcker disinflation period is excluded, showing the p-values for each test and the estimated coefficients to each variable. [56] For the test of bias the p-values for both the mean and the median consensus

[56] We do not focus on interpreting the coefficients here. However, they have the same signs as for the other performed tests, indicating underreacting of the new information by the respondents (with the coefficients of the lagged inflation being positive, and negative for the unemployment rate and the federal funds rate).

forecasts are much higher than during the Volcker disinflation period and then the total sample (presented in respectively section 6.5.2.1 and 6.3.1). With both p-values being higher than 10 %, we cannot reject the null of unbiasedness for both the mean and median consensus forecasts. Hence, the consensus of this period passes the first rationality criterion. This result is in line with Mehra's (2002) result; the forecast performance of the SPF improves when the Volcker disinflation period was excluded.

For the first efficiency test, the null of efficiency cannot be rejected. Adding the forecasts itself in efficiency test two gives us high probability values of the joint null being true. Hence, the consensus passes these two tests of efficiency. However, when adding the forecast error and the information set in efficiency test three and four, the p-values of the null hypothesis is zero for both the mean and the median consensus. Therefore the consensus does not pass these tests.

We conclude with the consensus passing the test of bias as well as two tests of efficiency when the Volcker disinflation period is excluded. As expected, this is a result much better

Table 6.39: Results for the rationality tests for the consensus mean and median forecasts when the Volcker disinflation period is excluded.

Rationality tests for the consensus when excluding the Volcker disinflation from the sample			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.126	-0.093
	α=0	0.406	0.550
Efficiency test 1: Lagged actual values	α (constant)	-0.590	-0.601
	β (lagged infl.)	0.166	0.183
	α=β=0	0.112	0.115
Efficiency test 2: Forecasted inflation	α (constant)	-0.260	-0.234
	β (forecasted infl.)	0.042	0.045
	α=β=0	0.657	0.764
Efficiency test 3: Lagged forecast error	α (constant)	0.024	0.044
	β (forecast error)	0.647	0.645
	α=β=0	0***	0***
Efficiency test 4: Information set	α (constant)	1.012	0.832
	β1 (lagged infl.)	0.500	0.549
	β2 (forecasted infl.)	-0.185	-0.248
	β3 (fed funds)	-0.210	-0.190
	β4 (unemployment)	-0.175	-0.147
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

than the results when looking at the Volcker disinflation only. Comparing with the consensus of the whole sample (presented in section 6.3.1), the conclusions are similar. One difference is that the probability values for unbiasedness and efficiency in efficiency test two are much higher for this sample than in the entire sample. Hence, we can say that the forecast performance have improved when the disinflation period is excluded, as Mehra (2002) suggests. However, the improvement does not seem very distinct.

6.5.3.2 Most individual respondents are unbiased, though not strong-form rational when the disinflation period is excluded

Table 6.40 presents the test results of the individual respondents' inflation forecasts in the sample where the Volcker disinflation period is excluded. Again we present number of individuals who pass the tests and the part they make of all individuals.

In the test of bias 69.8 % of the individuals are unbiased at a 5 % significance level.[57] Hence, as expected, the majority of the individual forecasts are unbiased when excluding the Volcker disinflation. Looking at efficiency test one, 47.1 % of all individuals are efficient at a 5 % significance level, a result very similar to the one we got when looking at the total sample (see section 6.3.2). The corresponding values for efficiency test two and efficiency test three are respectively 26.4 % and 41.6 %. When demanding strong-form rationality only 4.1 % of the individuals pass the rationality criterion. Hence, we conclude that individual respondents are not strong-form efficient and thus not strongly rational even when we exclude the Volcker disinflation period. Hence, the individual respondents do not seem to have improved a lot when the disinflation period excluded. This result does not correspond to Mehra's (2002) SPF result. Instead it is in accordance with his results for the Michigan survey and the Livingston survey.

[57] Also at a 10 % level, and naturally, also at a 1 % level, the majority also passes the test of bias.

Table 6.40: Results of the rationality tests performed on the individual respondents when the Volcker disinflation period is excluded from the sample. The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests when excluding the Volcker disinflation period:1979-1985				
	All sample	1 %	5 %	10 %
Test of bias: α=0				
Total nmb	129	101	90	77
Part of all		0.783	0.698	0.597
Efficiency test 1: α=β=0				
Total nmb	121	75	57	51
Part of all		0.620	0.471	0.421
Efficiency test 2: α=β=0				
Total nmb	129	42	34	23
Part of all		0.326	0.264	0.178
Efficiency test 3: α=β=0				
Total nmb	101	56	42	29
Part of all		0.554	0.416	0.287
Efficiency test 4: α=β1=β2=β3=β4=0				
Total nmb	121	7	5	2
Part of all		0.058	0.041	0.017

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

6.5.4 Concluding remarks regarding the Volcker disinflation period

The accuracy measures of the mean and median consensus forecasts presented show us that the forecast accuracy was worse during the Volcker disinflation period. When performing rationality tests, both the consensus and the majority of the individual forecasts have weaker rationality results than the entire sample. This leads us to the natural conclusion that the forecasts made in the Volcker disinflation period are worse than in the whole sample, as Mehra (2002) and Giordani and Söderlind (2002) suggests. However, even though rejected at a 5 % significance level for the median consensus, the p-values of the consensus tests of unbiasedness are not as low as we may expect when comparing the mean forecasts and the actual inflation in this period in figure 6.21. Additionally, even though the majority of the individual forecasters are biased, we could not reject unbiasedness for 41.8 % of the individuals. Hence, it does not seem like the forecasters were performing very badly even during this special disinflation period.

When we exclude the Volcker disinflation period, the results of the tests do not vary a lot from the results from the entire sample. One difference is that the probability values of

unbiasedness and efficiency in the consensus tests are higher in the bias test and in efficiency test two. For the individual forecasters one difference is that the part of individuals who are rational and efficient based on most efficiency tests are larger for the sample when the Volcker disinflation period is excluded than in the sample where we look at the Volcker disinflation period alone. The proportions are, however, not better than in the total sample. Because the conclusions of the tests are the same, the consensus and the individual respondents are not rational even when accounting for the Volcker disinflation period.

While the forecasts are worse during the Volcker disinflation period, leaving the period out of the sample does not "help" our rationality conclusion. The individual respondents and the consensus of the sample are not strong-form rational even with this period excluded. And even though the forecasters in the Volcker disinflation period have biased forecasts, the number of individuals with unbiased forecasts is actually higher than we expected. The p-value for unbiasedness for the mean and median consensus forecasts is also higher than expected. This is a result not completely in line with Mehra's (2002) SPF findings that the forecasters' performance increased when the Volcker disinflation period excluded.[58] The fact that the performance in the Volcker disinflation is not as bad as we expected, may be a contributing factor to why the sample where these are excluded do not perform particularly better than the entire sample.

The overall conclusion is that the Volcker disinflation period affected the forecasts to some degree, but cannot be claimed the reason to why the forecasters are not rational. A possible explanation can be that the forecasters in the survey are professionals. Being professionals, they may have known that Volcker planned decreasing the inflation, and therefore managed to make relatively good forecasts.

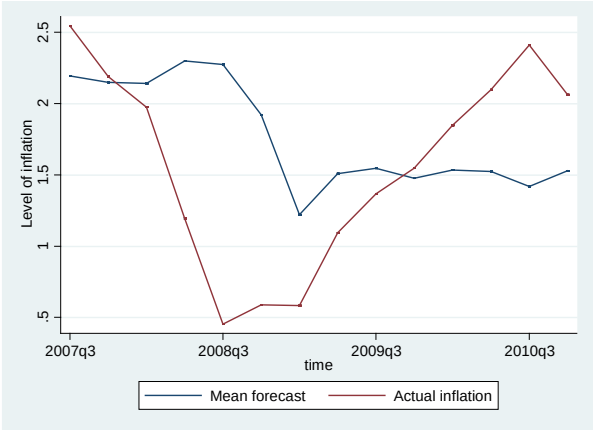
[58] The fact that our study is examined later than Mehra's study should be noted. The sample we use when excluding the Volcker disinflation is much longer than the sample of Mehra, implying that the effect of excluding the Volcker disinflation naturally is smaller.

6.6 The recent financial crisis

In this section we analyse the rationality of forecasters during the recent financial crisis. The data is relatively new, and we do not have knowledge of others examining this period. Hence, this analysis can bring something new to the existing literature. Previous literature states that disturbances in the economy, for example high unemployment, large government deficits and a moderate recession, can cause difficulties for forecasters (Su & Su, 1975).[59] We therefore believe that the recent financial crisis can have caused difficulties for the respondents of the SPF when predicting the next year inflation. With the disturbances in the economy starting in the summer of 2007, we choose to look at data from the third quarter of 2007 until the fourth quarter of 2010.[60] It is important to note that we do not have a lot of data to work with in this analysis. This makes the analysis weaker than it would have been if we had a large sample of individuals (and in addition a longer time sample) to examine.

Figure 6.22 presents the mean forecast of the individuals and the actual inflation in this period. In the figure there seems to have been large differences between the two.[61] In the third quarter of 2008 the error was over 1.5 %. Hence, the actual inflation in the third quarter of 2009 was much lower than the forecasters thought it would be in the third quarter of 2008. To examine if and how the forecasts have been affected by the financial crisis, we start looking at the accuracy measures of the consensus in section 6.6.1. The rationality tests of the forecasts are presented in section 6.6.2.

Figure 6.22: The mean forecast from the SPF and the actual inflation during the financial crisis, 2007q3 until 2010q4. The values in a given quarter are the forecast given of the next year inflation that quarter and the actual inflation for the next year.



[59] This is discussed in appendix one.
[60] Where the data in the fourth quarter of 2010 being forecasts for the expected inflation in the fourth quarter of 2011, and the actual data we compare it with the actual inflation in the fourth quarter of 2011.
[61] It is, however, important to be aware of the fact that the y-axe does not have a very wide range, from 0.5 % to 2.5 %.

6.6.1 The forecast accuracy is not worse during the financial crisis

The accuracy measures of the consensus mean and median forecasts during the financial crisis may give us a first hint to whether the financial crisis has made forecasting more difficult. The accuracy measures for the consensus forecasts in the financial crisis are presented in table 6.41. For comparison the forecast accuracy of the consensus in the total sample, in the sample where the Volcker disinflation is excluded and the sample starting in the second quarter of 1990 are also presented.

ME values of the consensus forecasts made during the crisis are negative, with values a bit smaller than for the total sample. This means that the respondents have been overestimating the inflation in this period. With large negative errors, at least in 2008 (visualized in figure 6.22), this seems to be correct, but the figure also shows us that the pattern is changing a lot in this period. When comparing with the total sample, the ME values are not worse during the financial crisis. It is important to be aware that the total sample includes early periods with large irregularities in the inflation, for example the Volcker disinflation. If we compare with the ME in the sample where the Volcker disinflation period is excluded, we see that the ME is much higher in the financial crisis. This also holds when comparing the consensus mean and median ME values for the financial crisis with the values in the sample starting in the second quarter of 1990.

MAE, row two in table 6.41, is not higher during the financial crisis either. If we compare with the sample where the Volcker disinflation period is excluded, the MAE values of the consensus are very similar. The RMSE values of the forecasts are much lower during the financial crisis than in the total sample, and also lower than the values for the sample where the Volcker disinflation period is excluded and for the shorter sample starting in the second quarter of 1990. Also the MNSE values are lower in the financial crisis sample than in the other three samples. This indicates that even when accounting for the lower dispersion in the actual inflation in this late period, the forecast accuracy is better during the recent crisis.

With the accuracy measures being relatively low, the forecasts during the financial crisis have not worsened. For several measures they are actually better than in the whole sample. Looking at figure 6.22 this is a bit surprising.

Table 6.41: The accuracy measures for the consensus mean and median forecasts in the whole sample, and in the financial crisis (2007-2010).

Accuracy measures for the consensus								
	The entire sample		The financial crisis		Excluding Volcker		Sample from 1990q2	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
ME	-0.275	-0.280	-0.200	-0.201	-0.126	-0.093	-0.198	-0.173
MAE	0.865	0.884	0.610	0.630	0.699	0.708	0.729	0.728
RMSE	3.311	3.378	0.747	0.751	1.366	1.006	1.804	1.578
MNSE	2.216	2.260	0.904	0.909	1.006	0.741	2.159	1.888

6.6.2 The rationality of forecasts during the financial crisis

This section presents the analysis of the rationality of the forecasts during the financial crisis. We follow the same pattern as we did in our previous sections, starting with the results for the consensus, in section 6.6.2.1. The analysis of the individual respondents is presented in section 6.6.2.2. The data from the financial crisis contain 52 individuals.

6.6.2.1 The consensus is not biased and are also efficient in most tests

Table 6.42 present the results for the consensus mean and median forecasts, containing the p-values for each test as well as each variable's coefficient.[62] With both the mean and median consensus having relatively high p-values, both over 50 %, we cannot claim the forecasts biased in this period. With p-values much higher than the corresponding ones for the total sample, presented in table 6.14, and in the sample starting in the second quarter of 1990, this gives us no indication of worsened forecasts during the financial crisis. Instead, the forecasters are more accurate and have a larger probability of being unbiased.

Turning to the efficiency tests, we see that the mean and median consensus pass test two and three, indicating relatively efficient forecasters. For efficiency test one we can reject efficiency at a 10 % level, though not at a "normal" significance level of 5 %. As in our other samples, the null of rationality is rejected for efficiency test four. Thus, the consensus forecasts in the financial crisis are not strong-form rational.

[62] In this analysis we will not focus on interpreting the coefficients. However, taking a look at them, in table 6.42, we see that they do not all have the same signs as for the other samples and tests. For example are the coefficients of the federal funds rate and the unemployment rate positive, leading us to the opposite conclusion than previously. It is important to be aware that this sample is relatively short, and that we do not have a lot of information about how new information have affected the inflation forecasts. Hence, we choose not to make strong conclusion regarding these patterns.

We conclude that the consensus mean and median forecasts during the current financial crisis performs pretty well. They pass the bias test with large p-values and we cannot reject efficiency in three out of the four tests. Hence, the forecasts seem to be more accurate and rational in this period. Because of the seriousness of the crisis, and what other literature has found about the rationality of forecasters in earlier crisis, this result seems rather surprising. The results may be explained with the level of and the change in the actual inflation being lower this period, indicating that forecasting is easier (Thomas, 1999). However, because the MNSE measure is lower in this period (presented in row four in table 6.41), the dispersion in the actual data does not seem to be the explanation. Thus, it may seem that forecasts actually have improved over the years, maybe due to more sophisticated analysis techniques and a better understanding of the economy. The conclusion that forecasts have improved is thus in accordance with previous research by for example Croushore (2006), even when the sample that we look at is the financial crisis.

Table 6.42: Results for the rationality tests for the consensus mean and median forecasts during the financial crisis (2007-2010).

Rationality tests for the consensus during the financial crisis			
Test	Hypothesis for rationality	Coefficients and p- values	
		Mean	Median
Test of Bias	α (constant)	-0.200	-0.201
	α=0	0.573	0.587
Efficiency test 1: Lagged actual values	α (constant)	0.917	1.043
	β (lagged infl.)	-0.818	-0.890
	α=β=0	0.081*	0.055*
Efficiency test 2: Forecasted inflation	α (constant)	1.571	1.534
	β (forecasted infl.)	-1.002	-0.981
	α=β=0	0.179	0.111
Efficiency test 3: Lagged forecast error	α (constant)	-0.364	-0.305
	β (forecast error)	-0.333	-0.272
	α=β=0	0.745	0.726
Efficiency test 4: Information set	α (constant)	-12.775	-9.982
	β1 (lagged infl.)	0.887	0.623
	β2 (forecasted infl.)	0.762	-0.875
	β3 (fed funds)	0.055	0.375
	β4 (unemployment)	1.146	1.013
	α=β1=β2=β3=β4=0	0***	0***

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1 F_t+\beta_2 A_{t-4}+\beta_3 U_t+\beta_4 i_t+\varepsilon_t$

* Rejects the null hypothesis on a 10% significance level

** Rejects the null hypothesis on a 5% significance level

*** Rejects the null on a 1% significance level

6.6.2.2 Individual respondents are unbiased, but not strong-form efficient

The results of the rationality tests performed on the individual respondents are presented in table 6.43. Again we present the part of all individuals who passes the tests and the number of individuals who fulfil the demands of the Newey-West method (which are explained in section 5.4.4).

When performing the test of bias we have 41 individuals to test. Almost all of those, 87.8 % cannot be claimed biased at a 5 % significance level. Thus, the majority of individual respondents are unbiased. With the number of unbiased individuals being high, we have no indication of them performing worse than before. There are instead a larger fraction of individuals who are unbiased, again leading towards the conclusion that the individual respondents have become better forecasters.

The efficiency tests give us fewer individuals who pass the tests. For most tests less than 50% of the individuals are efficient at a 5 % significance level. The exception is test three, where

Table 6.43: Results for the rationality tests performed on the individual respondents for the sample only containing the financial crisis (2007q3-2010q4). The consensus mean forecast's p-value and the part of the individuals who are better than this consensus is also presented. "Total nmb" is for "all sample" the number of individuals who we have performed the tests on, while the numbers to each significance level are the number of individuals who passes the test based on this level.

Overview rationality tests during the financial crisis: 2007q3-2010q4				
	All sample	1 %	5 %	10 %
Test of bias: $\alpha=0$				
Total nmb	41	38	36	32
Part of all		0.927	0.878	0.780
Efficiency test 1: $\alpha=\beta=0$				
Total nmb	35	27	15	11
Part of all		0.771	0.429	0.314
Efficiency test 2: $\alpha=\beta=0$				
Total nmb	41	21	14	9
Part of all		0.512	0.341	0.220
Efficiency test 3: $\alpha=\beta=0$				
Total nmb	29	23	19	17
Part of all		0.793	0.655	0.586
Efficiency test 4: $\alpha=\beta1=\beta2=\beta3=\beta4=0$				
Total nmb	34	4	2	1
Part of all		0.093	0.059	0.029

Test of bias: $A_t-F_t=\alpha+\varepsilon_t$

Eff. test 1: $A_t-F_t=\alpha+\beta\ A_{t-4}+\varepsilon_t$

Eff. test 2: $A_t-F_t=\alpha+\beta F_t+\varepsilon_t$

Eff. test 3: $A_t-F_t=\alpha+\beta(A_{t-4}-F_{t-4})+\varepsilon_t$

Eff. test 4: $A_t-F_t=\alpha+\beta_1F_t+\beta_2A_{t-4}+\beta_3U_t+\beta_4i_t+\varepsilon_t$

65.5 % pass the test. Hence, the forecasters in this period seem to be very aware of their previous forecast error. The period we are examining is a period with a lot of attention and focus on the economy. Thus, this finding may not seem too odd. For efficiency test one, 42.9% of all individuals are efficient, with the corresponding numbers for test two and four being 34.1 % and 5.9 %. When comparing with the whole sample, in table 6.18, and the shorter sample starting in the second quarter of 1990, in table 6.29, almost all tests have a larger, or similar part of individuals passing the tests during the financial crisis. Hence, there seem to be more individuals efficient during the financial crisis.

6.6.3 Concluding remarks about the forecasts during the financial crisis

Summing up our findings we see that neither the consensus nor the individual respondents have performed worse during the financial crisis. The consensus forecasts' accuracy measures are mostly lower than in the total sample and in the shorter sample starting in the second quarter of 1990. The consensus mean and median forecasts have high p-values for unbiasedness and relatively high p-values for efficiency. It is only in the test of strong-form efficiency we reject efficiency in this sample. Also the individual forecasters seem to have been performing well during the financial crisis. Almost all individual respondents are unbiased, and the fractions of efficient individuals are for many tests higher than in the corresponding tests for previous data samples.

The concluding remark is that the forecasters show no tendency of worse performances during the financial crisis. Several tests actually indicate that they have improved their forecasts, a conclusion opposite of what we would expect. The thought that the lower dispersion in the actual inflation in this period, as pictured when comparing figure 6.22 and figure 6.1, could explain some of this, does not seem to hold because of a low MNSE measure. Hence, it seems like the financial crisis have not made forecasting worse, and the previous finding that forecasters have improved over the years, seems to hold. It is important to note that the situation in the world economy is still quite anxious. We cannot be sure that the effect of the financial crisis have past, thus it will be preferable to examine the effect of the crisis for a longer sample than the one we have available. Also the fact that the dataset is quite small is a weakness of this analysis.

6.7 Conclusion

This aim of this paper has been to contribute to the broad existing literature of inflation expectations by examining the rationality of professional individual forecasters in a thoroughly manner. Using survey measures we compare consensus forecasts with individual forecasts, investigating if they are rational or not. We find that both the consensus and the majority of individual forecasts are quite accurate, though not strong-form rational, and that the forecasts seem to have improved over the time period. These results are in accordance with previous studies of the consensus (Croushore, 2006; Gerberding, 2006). However, we find that the behaviour of individuals varies. Both the accuracy measures and rationality tests performed on each individual reveal differences between the "best" and the "worst" forecasters. Because a lot of previous literature and most macroeconomic models presume that individuals have relatively similar expectations, it is important to highlight these findings (Mankiw, et al., 2003).

Previous literature has often stated that the consensus outperform individuals (Zarnowitz, 1984; McNees, 1987). Even though we find that the majority of individuals pass fewer efficiency tests than the consensus, we also find that a relatively large fraction of individuals outperform the consensus in several rationality tests. Hence, the consensus does not seem to be better than almost all individuals, as McNees (1987) states.

The industry variable in the SPF has to our knowledge not been examined previously. In our analysis the forecasts from professionals employed in different industries did not differ much. Using a previous study by Laster et al. (1999) we find that the forecasters employed in nonfinancial service provider firms can possess larger strategic incentives. These incentives can cause some of the forecasters in this "industry" to be less accurate than those employed in financial service provider firms. Our analysis of the Volcker disinflation period and the recent financial crisis find both the consensus and the individual forecasters to be more accurate and rational than we expected. Even though the majority of individuals were biased during the Volcker disinflation, there were many individuals for whom biasedness could not be claimed. In the financial crisis, both the consensus and the individuals performed better during the crisis than in the total sample and then the shorter sample starting in the second quarter of 1990.

Our analyses find individuals to be quite accurate, even when accounting for special episodes and different employment. Even though strong-form rationality is rejected for all tests, our results indicate that professional forecasts are relatively good. Our results also indicate differences between individuals. To find how and why individuals differ, maybe developing a new model or hypothesis regarding how individuals form their expectations could be an interesting topic for further research, and is a topic some forecasters already have begun examining (Mankiw, et al., 2003). As stated by Bernanke in his speech in Cambridge July 10, 2007: “a deeper understanding of the determinants and effects of the public's expectations of inflation could have significant practical payoffs” (Bernanke, 2007). Hence, getting better knowledge about how inflation expectations are formed is desirable, and the broad research of these expectations should continue.

Bibliography

Ang, A., Bekaert, G. & Wei, M., 2007. Do macro variables, asset markets, or surveys forecast inflation better?. *Journal of Monetary Economics*, Issue 54, pp. 1163-1212.

Ball, L. & Croushore, D., 2003. Expectations and the Effects of Monetary Policy. *Journal of Money, Credit and Banking*, August, 33(4), pp. 473-484.

Batchelor, R., 2000. The IMF and OECD versus Consensus Forecasts. *City University Business School*, August.

Batchelor, R. & Dua, P., 1995. Forecasters Diversity and the Benefits of Combining Forecasts. *Management Science*, pp. 68-75.

Bernanke, B. S., 2007. *Inflation Expectations and Inflation Forecasting.* Cambridge, Board of Governors of the Federal Reserve System.

Bernanke, B. S. & Mishkin, F. S., 1997. Inflation Targeting: A New Framework for Monetary Policy?. *Journal of Economics Perspectives*, Spring, 11(2).

Board of Governors of the Federal Reserve System, 2012a. *About the FOMC.* [Online]
Available at: http://www.federalreserve.gov/monetarypolicy/fomc.htm
[Accessed 27 May 2012].

Board of Governors of the Federal Reserve System, 2012b. *Board of Governors of the Federal Reserve System "Press Release".* [Online]
Available at: http://www.federalreserve.gov/newsevents/press/monetary/20120125c.htm
[Accessed 4 June 2012].

Bonham, C. S. & Dacy, D. C., 1991. In search of a "Strictly Rational" Forecast. *The Review of Economics and Statistics*, May, 73(2), pp. 245-253.

Bullard, J. & Mitra, K., 2002. Learning About Monetary Policy Rules. *Federal Reserve Bank of St. Louis, Working Paper 2000-001E.*

Bureau of Economic Analysis, 2011a. *Help for National Income and Product Account Tables.* [Online]
Available at: http://www.bea.gov/national/FA2004/NIPAHelp.htm
[Accessed 07 June 2012].

Bureau of Economic Analysis, 2011b. *Mission, Vission and Values.* [Online]
Available at: http://www.bea.gov/about/mission.htm
[Accessed 27 May 2012].

Bureau of Labor Statistics, 2012a. *About BLS.* [Online]
Available at: http://www.bls.gov/bls/infohome.htm
[Accessed 27 May 2012].

Bureau of Labor Statistics, 2012b. *Labor force characteristics.* [Online]
Available at: http://www.bls.gov/cps/lfcharacteristics.htm#unemp
[Accessed 27 May 2012].

Chew, J. & Price, C., 2008. *Introducing: An Industry Classification for the Survey of Professional Forecasters.* [Online]
Available at: http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/spf-industry-variable.pdf
[Accessed 16 March 2012].

Clements, M. P., 2004. Evaluating the survey of professional forecasters probability distributions of expected inflation based on derived event probability forecasts. *Empirical Economics*, 23 August.

Clements, M. P., 2006. Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. *Warwick Economic Research Papers.*

Clements, M. P., 2008a. Consensus and uncertainty: Using forecast probabilities of output declines. *International Journal of Forecasting*, pp. 78-86.

Clements, M. P., 2008b. Explanations of the inconsistencies in survey respondents forecasts. *Warwick Economic Research Papers*, 14 July.

Clements, M. P., 2008c. Rounding of probability forecasts: The SPF forecast probabilities of negative output growth. *Warwick Economic Research Papers.*

Croushore, D., 1993. Introducing: The Survey of Professional Forecasters. *Business Review*, November/ December.

Croushore, D., 2006. "An Evaluation of Inflation Forecasts From Surveys Using Real-time Data". *Working Papers Research Department.*

Croushore, D. & Stark, T., 1999. A Real-Time Data Set for Macroeconomists. *Working Paper No. 99-4.*

DeLong, J. B., 1997. *America's Peacetime Inflation: The 1970s.* [Online]
Available at: http://www.nber.org/chapters/c8886.pdf
[Accessed 20 May 2012].

Diebold, F. X., Tay, A. S. & Wallis, K. F., 1997. Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. *NBER Working Paper No. 215*, October.

Ehrbeck, T. & Waldmann, R., 1996. Why Are Professional Forecasters Biased? Agency versus Behavioral Explanations. *Quarterly Journal of Economics*, pp. 21-40.

Federal Reserve Bank of Philadelphia, 2008. *Survey of Professional Forecasters-Documentation.* [Online]
Available at: http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/spf-documentation.pdf
[Accessed 12 February 2012].

Gärtner, M., 2006. *Macroeconomics.* Essex: Pearson Education, Prentice Hall.

Gerberding, C., 2006. *Household versus expert forecasts of inflation: New Evidence from European survey data.* [Online]
Available at: http://www.nbp.org.pl/konferencje/bbm/gerberding.pdf
[Accessed 23 March 2012].

Giordani, P. & Söderlind, P., 2002. Inflation Forecast Uncertainty. *European Economic Review.*

Goetzmann, W. & Massa, M., 2003. Disposition Matters: Volume, Volatility and Price Impacts of Behavioral Bias. *Discussion Paper Series, Centre for Economic Policy Research,* Volume 4814.

Hansen, L. P. & Hodrick, R. J., 1980. Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *Journal of Political Economy,* October, 88(5), pp. 829-853.

Harri, A. & Brorsen, B. W., 2009. The Overlapping Data Problem. *Quantitative and Qualitative Analysis in Social Sciences,* 3(3), pp. 78-115.

Holden, K. & Peel, D. A., 1990. On Testing for Unbiasedness and Efficiency of Forecasts. *The Manchester School,* June, 58(2), pp. 120-127.

Keane, M. P. & Runkle, D. E., 1990. Testing the Rationality of Price Forecasts: New Evidence from Panel Data. *American Economic Review,* September, 80(4), pp. 714-735.

Kershoff, G. & Smit, B., 2002. Conducting Inflation Expectation Surveys in South Africa. *South African Journal of Economics,* March, 70(3), pp. 205-212.

Lamont, O. A., 2002. Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behaviour & Organization",* Volume 48, pp. 265-280.

Laster, D., Bennett, P. & Geoum, I. S., 1999. "Rational Bias in Macroeconomic Forecasts. *The Quarterly Journal of Economics,* February, pp. 293-318.

Llooyd, B. T. J., 1999. Survey Measures of Expected U.S Inflation. *The Journal of Economic Perspectives.*

Lovell, M. C., 1986. Tests of the Rational Expectations Hypothesis. *The American Economic Revies,* March, 76(1), pp. 110-124.

Mankiw, G. N., Reis, R. & Wolfers, J., 2003. Disagreement about Inflation Expectations. *NBER Working Paper Series,* June, Issue 9796.

McNees, S. K., 1987. Consensus Forecasts: Tyranny of the Majority. *New England Economic Review*, Nov/Dec.

McNees, S. K., 1992. How Large are Economic Forecast Errors?. *New England Economic Review*.

Mehra, Y. P., 2002. Survey Measures of Expected Inflation: Revisiting the Issues of Predictive Content and Rationality. *Economic Quarterly*, Summer, 88(3).

Mincer, J. A. & Zarnowitz, V., 1969. The Evaluation of Economic Forecasts. *Economic Forecasts and Expectations: Analysis of Forecasting Behaviour and Performance*, pp. 1-46.

Muth, J. F., 1961. Rational Expectations and the Theory of Price Movements. *Econometrica*, July, 29(3), pp. 315-335.

Newey, W. K. & West, K. D., 1987. A Simpler, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, May, 55(3), pp. 703-708.

Pearce, D. K., 1979. Comparing Survey and Rational MEasures of Expected Inflation: Forecast Performance and Interest Rate Effects. *Journal of Money, Credit and Banking*, November, 11(4), pp. 447-456.

Petersen, M. A., 2009. Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies*, 22(1), pp. 435-480.

Roberts, J. M., 1998. Inflation Expectations and the Transmission of Monetary Policy. *Board of Governors of the Federal Reserve System*, Oct, Issue 98.

Stark, T., 1997. Macroeconomic Forecasts and Microeconoic Forecasters in the Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Working Paper 97-10.*

Stark, T., 2011. *General Notes on the Philadelphia Fed's Real-Time Data Set for Macroeconomists (RTDSM) – Variables from the National Income and Product Accounts.* [Online]
Available at: http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/data-files/documentation/gen_doc_NIPA.pdf
[Accessed 19 March 2012].

Statistics Norway, 2012. *Prices.* [Online]
Available at: http://www.ssb.no/english/subjects/08/priser_tema_en/
[Accessed 11 June 2012].

Stekler, H. O., 2002. The Rationality and Efficiency of Individuals' Forecasts. In: *"A Comanion to Economic Forecasting"*. 02148-5020(MA): Blackwell Publishing, pp. 222-240.

Su, V. & Su, J., 1975. An Evaluation of ASA/NBER Business Outlook Survey Forecasts. In: *Explorations in Economic Research, Volume 2, number 4.* s.l.:NBER, pp. 588-618.

Thomas, J. L. B., 1999. Survey Measures of Expected U.S Inflation. *Journal of Economic perspectiver*, Fall, 12(4), pp. 125-144.

Thomson Reuters, 2012. *Reuters Ecowin Pro.* [Online]
Available at: http://thomsonreuters.com/products_services/financial/financial_products/a-z/ecowin_pro/#tab2
[Accessed 28 May 2012].

Zarnowitz, V., 1984. The Accuracy of Individual and Group Forecasts. *Journal of Forecasting,* Volume 3, pp. 11-26.

Zarnowitz, V., 1985. Rational Expectations and Macroeconomic Forecasts. *Journal of Business and Economic Statistics,* October, Volume 3, pp. 293-311.

Zarnowitz, V., 1992. *Business Cycles; Theory, History, Indicators, and Forecasting.* 60637(Chicago): The University of Chicago Press.

Zarnowitz, V. & Braun, P., 1993. Twenty-two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance. *Business Cycles, Indicators and Forecasting,* January, pp. 11-94.

Appendix

Appendix 1: Inflation forecasting in different time periods

In 6.1 we mention that it seems to be harder to predict the inflation in certain time periods. This could be because of special episodes, different policies, etc. Several articles mention this fact. In this appendix we will present some of those and what they have found.

Su and Su (1975) evaluated the SPF (at the time the ASA/NBER Business Outlook Survey) in an early study. Because it was written in an early stage of the survey, they did not have a lot of observations to investigate, and the article was also written in a challenging period. Some of the issues that they discuss could still be important to consider. They say that the forecasting period from 1968 to 1973 is generally considered to be a difficult period for forecasters. Su and Su (1975) claim that his was because of a high unemployment level, a moderate recession, a rapid inflation, a serious auto strike, a large government deficit as well as a foreign trade deficit. Even though our sample is restricted to after this period, the fact that these factors can make forecasting more difficult is something we should keep in mind. In addition to the fact that some of these factors may be present at some time in our data, the historical data that forecasters often use when making predictions will be of little value when coming from this period.

Croushore (2006) also points out some episodes that may be able to explain the poor performance of forecasters, like the fact that the inflation rose much higher than the forecasters believed after the oil-price shocks in the 1970s. He also states that the overlapping observations problem will be of importance when investigating these episodes.[63] Another problem with the SPF mentioned by Croushore (2006) is that in the early forecasts for the GNP deflator, they rounded the forecast to the nearest whole number, and this caused the forecasts to be quite erratic in these early years of the survey. This goes for 1968, 1969 and 1970. When looking at the forecasted inflation, for example in figure 6.1 in section 6.1, this seems hard to confirm because there were very few forecasts these years, with forecasted inflation missing for many of the quarters. However, with these years excluded from our data, this will not be a problem for us.

[63] This problem was discussed in 5.4.4. A shock will affect the actual values for several consecutive periods because the forecasts span a longer period than the sampling frequency, hence the forecast errors will be correlated.

Another paper that examines how the forecast performance has evolved is Mankiw et al. (2003). They look at how the disagreement between the forecasters has varied with the business cycle, and find that the disagreement between the economists in the SPF does not have a very strong obvious relationship with the state of the real economy. However, they do find that large changes in inflation, both positive and negative, are correlated with an increase in disagreement. Figure 6.1, presenting the mean forecasts of individuals and the actual inflation, suggests that this holds for our data as well The largest differences between the forecasted inflation and the actual inflation are in times when the inflation level was higher and more changing than it has the last years (in line with also other studies, like Su and Su (1975) and Croushore (2006)). Also when looking at the standard deviation of the inflation forecasts against the level of the actual inflation, in figure 6.5, this relationship seems to hold.

The article by Mankiw et al. (2003) also investigates the effect of the mentioned Volcker disinflation on the forecasts. They find that the expectations adjusted slowly to the regime change that the disinflation period presented. Even though they examine the Michigan survey, we think that this period also affected the SPF forecasters (we examine this in section 6.5). However, the effect may be smaller for the SPF forecasters, because they are economic professionals who should understand the impact of the disinflation to a larger extent than the consumers in the Michigan survey. Looking at figure 6.1-6.5 in section 6.1, it seems to be true that the Volcker disinflation affected the forecasts.[64] It appears that the forecasters did not see the fall in inflation coming as quick as it did, with the forecasted inflation "lagging" the actual inflation with a substantial difference between the actual and the expected inflation in this period (a relationship also mentioned by Giordani and Söderlind (2002) and Croushore (2006)). The degree of disagreement between forecasters, measured by the standard deviation presented in figure 6.5, also seem to be higher in these periods than later on.

Another paper that discusses the effect of specific events is a paper written by Lloyd (1999). Lloyd talks about unforeseeable "regime changes" and that failures of others, for example the central bank in keeping an inflation target, can cause fully rational agents to make systematic errors in certain periods. The mentioned effect of the Volcker disinflation period could be an example of such a regime change.

[64] This is a pattern that Mehra (2002) also found in the SPF. Mehra (2002) did however not find that excluding the Volcker disinflation period improved the results from the Michigan survey.

Appendix 2: The data

Appendix 2.1 Revised versus vintage data

When comparing survey data and actual data it is important to choose between revised or vintage actual data. The fully revised data is the newest value of the variable in question. If choosing vintage data, there are different sets to choose from, being the first one published or others published sometime after the first publications. For this to be possible it is necessary to have a real-time data set for the variables that one wants to look at.[65] For the NIPA variables the publications of the Bureau of Economic Analysis (BEA) are used as vintage data. The NIPA values undergo a systematic process of revision.[66]

Every fifth year there is a benchmark revision to the NIPA variables. In these revisions the base year level for the variables, thus the scale of the data, can change. This means that it is usually not appropriate to compare the level of an observation in one vintage with the level of the same observation in a different vintage if a benchmark revision span the two.[67]

Previous literature discusses whether to use revised or vintage data with different conclusions (Keane & Runkle, 1990; Croushore & Stark, 1999; Zarnowitz & Braun, 1993). Two issues are important in this discussion (Keane & Runkle, 1990). The first one is whether the respondents are trying to predict the initial or the revised data. If the first is true, then vintage data should be used, if not, using the revised data is appropriate. Keane and Runkle (1990) find that predictions on average were closer to the initial announcements than the revisions. They therefore argue that the forecasters are trying to predict the initial announcement, and choose to use vintage data when comparing. However, in the forecasting literature it is more common to analyse based on the latest variables, thus revised, data (Croushore, 2006). The reasoning behind this is that it is the final actual data that the forecasters are trying to predict, not some preliminary data. Zarnowitz and Braun (1993) claim that while the preliminary data

[65] A real- time dataset consists of the vintages; snapshots of the data at different times in the past, before the data were fully revised data. A vintage date is the date when the data were available for the public for the first time. This is discussed in (Stark, 2011) and in Stark and Croushore (1999)

[66] Near the end of the first month in each quarter the BEA releases the first estimate for the previous quarter. Revisions to this advance estimate, the preliminary and final estimates are released near the end of the following two months. After this, BEA releases annual revisions to estimates for the previous three years in its annual revision. Then, every few years, a benchmark revision is released, and these will usually affect all observations. In addition to incorporating new economic information, benchmark revisions often incorporate new statistical procedures and new definitions.

[67] This will be a problem if we want to compare our forecasts with the revised data, because the forecasts are not being revised with the base year changes (Federal Reserve Bank of Philadelphia, 2008). But, within a particular vintage, one can compare observations over time by for example computing growth rates.

are most closely related to what were available to the forecasters they may themselves be partly predictions. They could therefore themself deviate from the “truth,” represented by the last revised data. On the other hand, the fully revised data may contain a lot of benchmark revisions. To demand the forecasters to be responsible for all measurement errors that are corrected for by these is questionable.

The other issue Keane and Runkle (1990) states important is whether the data revisions are significant or predictable. If the data revisions are predictable, not systematic and not significant, the use of revised or vintage data should not be of big importance. Croushore and Stark (1999) find that the results for different vintages often are robust; meaning that this choice will not be of importance.

Appendix 2.2 The number of forecasts of the other pgdp levels for each individual

In 5.2, figure 5.1 and 5.2 we presented the number of forecasted pgdp2 and pgdp6 levels for each individual. We stated that those were pretty similar as for the other pgdp levels. These are presented in this appendix, in figures A2.1-A2.4.

Figure A2.1: The number of forecasted pgdp1 levels for each individual.

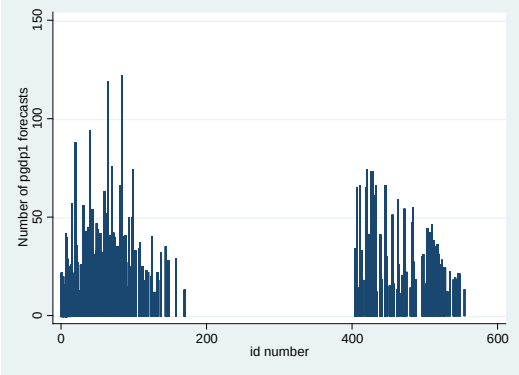


Figure A2.2: The number of forecasted pgdp3 levels for each individual.

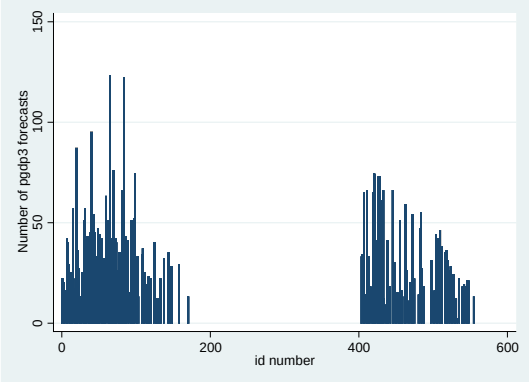


Figure A2.3: The number of forecasted pgdp4 levels for each individual.

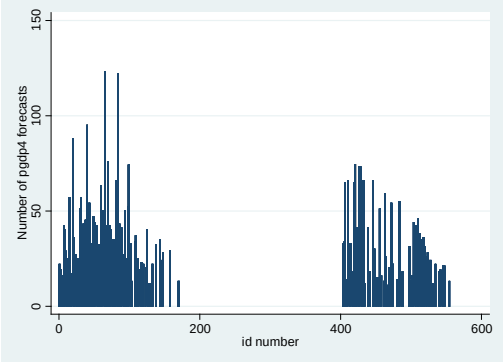
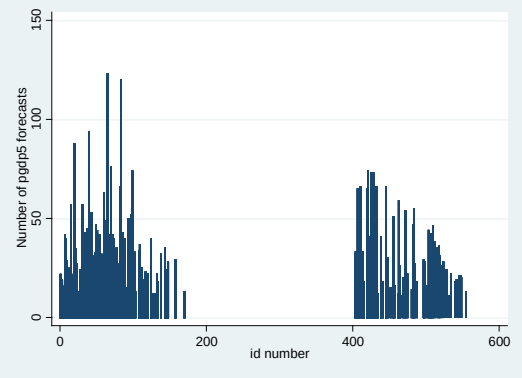


Figure A2.4: The number of forecasted pgdp5 levels for each individual.



Appendix 2.3 Altering the dataset to solve the problem of missing forecasts

In section 5.4.2 we discuss the problem of missing forecasts, and choose to start analyzing after the third quarter of 1974. However, another alternative could be to fill inn an estimated value of pgdp6. In this appendix we present this alternative and give an example of how this can be done.

Filling in an estimate of the missing values

Finding values for some of the missing pgdp6 levels that can give us an estimate of what the response of the individual is an alternative solution to the problem with missing forecasts. This is especially desirable for individuals where only a few forecasted levels are missing over a longer period of time. Doing this can give us more inflation forecasts and also more consecutive inflation forecasts.

Looking into previous studies, we have not found any other papers that have done something similar. To fill in for missing values is a task that needs consideration and carefulness, because we will be changing the original data. There are different approaches that could be thought of.[68] Here I will present one solution that we found could be good, to make a linear projection of other pgdp levels to find the missing value.

When examining the data we find that forecasters have often forecasted values of pgdp1-5, but not pgdp6. Hence, one alternative could be to use a simple linear projection to find a value for pgdp6 based on the earlier forecasts by the individual in that quarter. Two possible ways of doing this are:

$$1)\ alternative\ pgdp6 = pgdp5 * \frac{\left(\frac{pgdp2}{pgdp1}\right) + \left(\frac{pgdp3}{pgdp2}\right) + \left(\frac{pgdp4}{pgdp3}\right) + \left(\frac{pgdp5}{pgdp4}\right)}{4}$$

$$2)\ alternative\ pgdp6 = pgdp5 + \frac{(pgdp2 - pgdp1) + (pgdp3 - pgdp2) + (pgdp4 - pgdp3) + (pgdp5 - pgdp4)}{4}$$

Where the first one uses the relationship between the earlier observations and then multiplies it with pgdp5 to get pgdp6, while the other uses the difference and then adds it to pgdp5. The two methods give almost the same result.

[68] Examples are to make a liner projection if having the other pgdp levels necessary to do so, to fill in lead and lag values of pgdp2 and pgdp6, and to find out how the individual has performed compared to the mean before, and then fill in a value equivalent to those for the missing value.

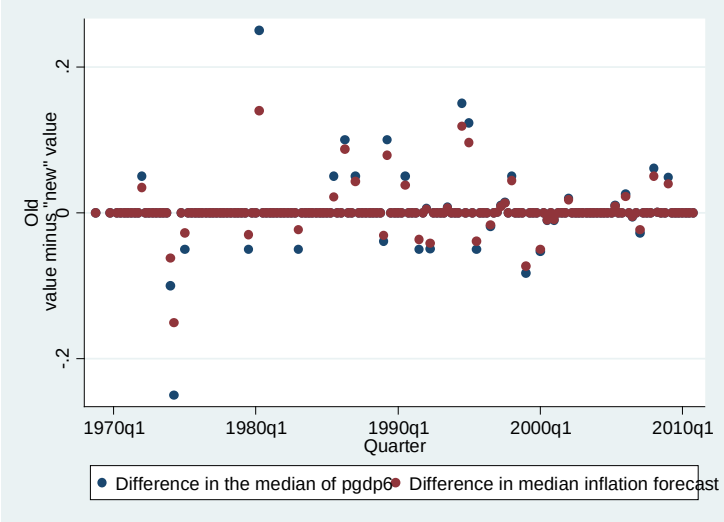
This linear projection method entails filling in 339 new values of pgdp6 when we have removed respondents with less than 12 responses. By the basic statistic showed in table A2.1 we see that this does not change the one-year-ahead forecast or the forecasted pgdp6 levels much, being 3.77% now and before (when comparing with the previous numbers when restricting the sample with those with 12 or more responses in table 5.1). Because we at the same time fill in values for pgdp6 in all the quarters that previously had no observation, thus increasing the number of forecasts per individual from 41.86 to 44.30, (presented in table 5.1) it seems as a reasonable method to use.

The difference between the new median inflation forecast and the "old," as well as how the pgdp6 levels have changed if we use this alternative, are presented in figure A2.5.

Table A2.1: Basic statistics of the forecasted values of the inflation after filling in a linear projection of pgdp6.

Statistics forecasted values and nmb of	Values after made a linear projection		
	Nmb forecasts per ind	pgdp6	Inflation forecast
Mean	44.295	145.334	3.772
Std	25.119	36.141	2.158
Min	11	105.7	-4.569
Max	123	247	31.137

Figure A2.5: The difference between the new median inflation forecasts and the old median inflation forecasts together with the difference made in the pgdp6 levels.



Appendix 2.4 Changing base year

When working with data from the SPF database we should be aware of the mentioned base year changes for several variables. Every fifth year, when there are benchmark revisions to the NIPA variables, the base year may change in addition to the data being revised (explained in section three). There have been several base year changes since the SPF survey began. Because the forecasted levels in the dataset have not been rescaled with the base year changes, the levels in the data set use the base year that was in effect when the questionnaire was sent to the forecasters. For the pgdp there have been seven base year changes that we should be aware of. Those were in 1976q1, 1986q1, 1992q1, 1996q1, 1999q4 and 2004q1, and are listed in table A2.2, (source: Federal Reserve Bank of Philadelphia, 2008, p. 10) and shown in figure A2.6.

Whether the forecasters manage to keep track with the base year changes in their forecasting is interesting.[69] In figure A2.6 we plot all individual forecasted levels of pgdp2 and pgdp6, showing that the individuals did keep track of the base year changes for pgdp2 and pgdp6. When working with percentage changes, the base year revisions do not have to be a problem, because the effect on the inflation rate is likely to be minor (Clements, 2004).[70] Because the individuals seems to keep track of the base year changes, and we are working with percentage changes when calculating the inflation, we do not think of the base year changes as a big problem.

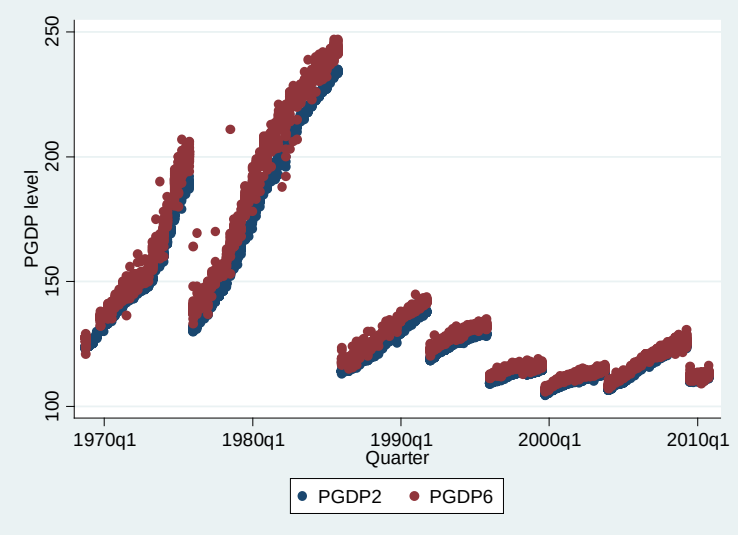
[69]This can also be done by checking the quarters where the base year changed, listing the individuals' responses these quarters and the previous quarter.

[70] However, if we want to compare the quarterly levels of pgdp with the actual levels, problems will occur. One problem is that the survey may ask for predictions further in the future than the next annual benchmark revision. A way of solving this is be to exclude all forecasts with horizons that extend beyond the date of systematic data revisions from the data (Keane & Runkle, 1990). One could also use vintage data when comparing (Clements, 2006).Vintage data will always have the same base year as the forecasts, being the data that was available around the time the forecast was made.

Table A2.2: Base year changes for the NIPA variables, which includes the GDP deflator. Source: SPF documentation p.8 (Federal Reserve Bank of Philadelphia, 2008).

Base Years for NIPA Variables in the Survey of Professional Forecasters Range of Surveys	Base Year
1968:Q4 to 1975:Q4	1958
1976:Q1 to 1985:Q4	1972
1986:Q1 to 1991:Q4	1982
1992:Q1 to 1995:Q4[2]	1987
1996:Q1 to 1999:Q3[3]	1992
1999:Q4 to 2003:Q4	1996
2004:Q1 to present	2000

Figure A2.6: Base year changes in forecasted pgdp levels. The forecasted levels of pgdp2 and pgdp6 are presented, making a downward "jump" each time the base year changes.



Appendix 2.5 Consistency of forecasts

When analysing data, it is important that the data that we use are reliable and consistent. Errors made by forecasters, such as extreme outliers can lead to forecasted values that do not seem reasonable. Extreme outliers can exist because of sloppy handwriting or other issues that makes the forecaster make mistakes (Giordani & Söderlind, 2002). To control for lacking consistency we can search the dataset for problems and eliminate them, and use robust methods when estimating, which will make the problems with outliers less severe (as for example the Newey-West method we are using discussed in 5.4.4).

We begin discussing the dispersion in the data, presented by the highest and lowest inflation forecast each quarter and by the standard deviation of the forecasts in figure 5.6 and 5.7, section 5. Looking at the figures, there seem to be some periods that have more extreme values than others. Thus, we should inspect the data to see if they seem consistent and correct. Doing this involves inspecting if all the pgdp levels forecasted by the given individual are extreme in the given period. This enables us to detect if one of the forecasted levels seems unreasonable compared to the others.

Extreme values can give biased results. Hence, it is important to locate them in order to assess their importance for our analysis. Figure A2.6 in appendix A2.4 present the forecasted pgdp2 and pgdp6 levels each quarter, and gives us a visual of potential outliers. As we expect, being a forecast of the current quarter pgdp, pgdp2 seem to be relatively consistent. However, pgdp6, seems to have some potential problematic outliers before the second base year change. The most serious one is located in the third quarter of 1978. In the same quarter one can find similar outliers in pgdp4 and pgdp5, which could imply that this is a forecaster who have made mistake or made a forecast which deviates from the consensus forecast. After some research, we find that the pgdp6 value reported here seems to be in line with the other pgdp levels that the given individual responded that quarter. Hence, it seems at though the respondent just made an optimistic forecast and can therefore not be seen as an outlier being in line with the forecaster's beliefs. The forecasted levels of pgdp as well as the calculated inflation forecast of this individual, number 47, is presented in table A2.3. The forecasted values give that this individual expects the inflation the next year to be 31.14%, over 20% higher than the mean and median values this quarter (with the mean being higher than the median, probably affected by this individual's high forecast).

Table A2.3: The forecasted values of the forecasted pgdp levels made by individual number 47 in 1978q3. The mean and median inflation forecast in the same quarter is also presented.

Outlier for id=47 in 1978q3							
Variable	pgdp1	pgdp2	pgdp3	pgdp4	pgdp5	pgdp6	Inflation forecast
Forecasted quarter	1978q2	1978q3	1978q4	1979q1	1979q2	1979q3	1979q3
id=47	150.7	160.9	172.5	185.0	197.7	211.0	31.137
Mean	150.673	153.681	156.908	160.131	163.181	165.908	7.910
Median	150.700	153.500	156.150	158.800	161.550	164.250	7.122

Because this was the largest outlier found it seems reasonable to believe that the other potential outliers are just a individual making a slightly more optimistic forecast than the consensus, hence we will not delete any of those from the dataset.

For the forecasted values to be claimed consistent we should to consider other things than potential outliers. One example is that if forecasts are consistent, the quarterly predicted pgdp levels should be relatively similar to the predicted annual levels. Testing this is possible in the quarters where we have annual forecasts available, from the third quarter of 1981 and onwards. In the first quarter of a year there should be consistency between pgdp6, which is the one-year-ahead forecasted pgdp level, and the annual average forecast the current year, pgdpa. Being forecasts of almost the same, these should be similar. Consistency should also exist between the level of pgdp6 and pgdpb, the annual-average forecast for the next year, when standing in the fourth quarter of a year. To investigate this we can find the percentage difference between the pgdp6 and pgdpa and plot this relationship against time when standing in the first quarter to see how this relationship has evolved. The same is done for the pgdp6 and pgdpb when standing in the fourth quarter of a year.

Figure A2.7 and A2.8 show these percentage differences between the forecasted inflation level one year ahead and the forecasted annual-average inflation level. The differences should be close to zero if the forecaster is consistent. However, that does not seem to hold, especially in the early 80's. In this period the differences between pgdp6 and pgdpa when standing in the first quarter varied from -3 % to +10 %. After Philadelphia Fed took over the survey, in early 1990, the problem became much less severe. This may imply that the forecasters were more aware of what they actually were forecasting. The pattern is quite similar for the pgdp6 and pgdpb when standing in the fourth quarter of a year, but the range of the percentage difference is smaller.

A possible solution to this consistency problem is to exclude values that are too extreme from the sample. This would make the dataset more robust and less exposed to outliers. We can, however, perform this consistency check from the third quarter of 1981 only. We have no way of checking the consistency for the earlier years, but looking at the results in figures A2.7 and A2.8 it is reasonable to believe that they are not too good. Only removing values from the 80's will not solve our problem, thus that is a bad solution. Another solution that seems more realistic is to only use data from 1990 and onwards, or use sub-samples that start after Philadelphia Fed took over the survey. Because of the huge amount of data that we will be missing by doing this, we will not do this when performing all tests, but we will however, also test the data only after the Philadelphia Fed took over the survey, thus using the more consistent period (this analysis is presented in 6.4.2).

Figure A2.7: Inconsistency between one-year-ahead inflation forecast standing in first quarter (pgdp6) and the forecasted average inflation current year (pgdpa).

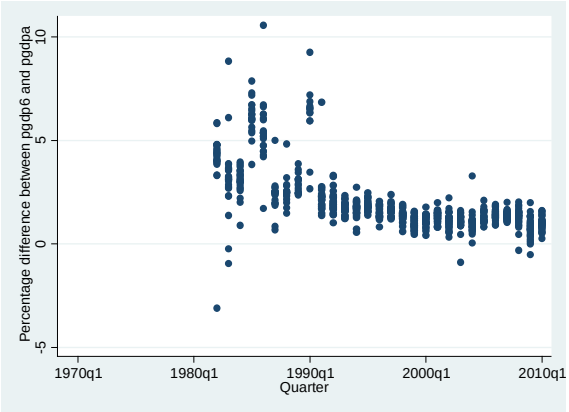
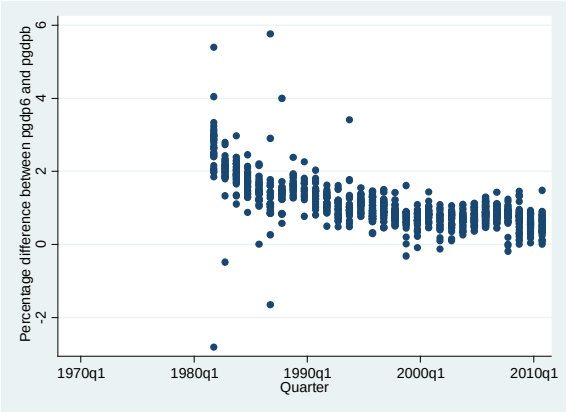


Figure A.2.8: Inconsistency between one-year-ahead inflation forecast (pgdp6) standing in the fourth quarter, and the forecasted average inflation for next year (pgdpb).



Appendix 3: Forecast accuracy

Appendix 3.1 The rankings of the best and the worst ten forecasters in terms of different accuracy measures

In 6.2 we rank the individual respondents in terms of having the lowest value of the different accuracy measures. The ones that were most accurate in terms of each measure are presented in table A3.1, while the ones who were least accurate, are presented in table A3.2. As mentioned in 6.2.2.3 we see that there is some overlapping in the best ones and in the worst ones. One example is individual number 472, who is ranked most accurate by ME, RMSE and MNSE.

Table A3.1: The most accurate respondents in terms of each accuracy measure

Best ten respondents								
Rank by best	ME		MAE		RMSE		MNSE	
	ID	Value	ID	Value	Id	Value	ID	Value
1	472	-0.001	531	0.508	472	0.010	472	0.0114
2	446	0.004	405	0.511	446	0.035	446	0.0415
3	448	0.020	422	0.514	448	0.107	145	0.1145
4	524	0.026	510	0.558	524	0.136	448	0.1400
5	431	0.028	502	0.579	145	0.143	524	0.1449
6	424	-0.029	544	0.581	424	0.144	424	0.1748
7	145	0.034	507	0.585	431	0.218	31	0.2217
8	65	0.046	546	0.593	465	0.301	431	0.2677
9	31	-0.059	465	0.595	502	0.313	65	0.3158
10	411	0.059	500	0.600	549	0.320	158	0.3224

Table A3.2: The least accurate respondents in terms of each accuracy measure.

Worst ten respondents:								
Rank by worst	ME		MAE		RMSE		MNSE	
	ID	Value	ID	Value	Id	Value	ID	Value
1	100	-2.251	148	2.4290	100	9.5484	100	12.975
2	23	-2.196	100	2.3654	60	9.1201	434	10.986
3	5	-1.865	23	2.2945	23	8.7834	440	10.114
4	47	-1.729	125	2.2709	47	8.2909	35	8.982
5	22	1.702	9	2.2147	35	8.2476	427	8.459
6	79	-1.630	47	2.2028	79	7.6445	23	8.361
7	13	-1.553	93	2.1874	66	7.2215	407	8.353
8	434	-1.538	31	2.1315	22	7.0188	66	7.920
9	69	-1.484	43	2.0680	5	6.7229	79	7.811
10	68	-1.484	5	1.8646	69	6.4697	60	7.689

Appendix 3.2 Summarizing the values of the accuracy measures and the number of forecasts per individual for the ten worst and the ten best respondents in terms of the accuracy measures

Table A3.3 summarizes the values of the accuracy measures. Values for mean and median consensus forecasts as well as the mean and median of the ten best and ten worst forecasters are presented for each accuracy measure.

Table A3.4 summarizes the mean and median number of forecasts for the consensus and the ten best and the ten worst for each accuracy measures. We see that the ten best tends to have made more forecasts than the worst for each measure, but that the number is not very much higher than the consensus number (as discussed in 6.2).

Table A3.3: Overview of the accuracy measures.

Overview accuracy measures	ME	MAE	RMSE	MNSE
Consensus mean	-0.275	0.865	3.311	2.216
Consensus median	-0.280	0.884	3.378	2.260
Mean of the ten best	0.013	0.562	0.173	0.175
Mean of the ten worst	-1.403	2.203	7.907	9.165
Median of the ten best	0.023	0.580	0.144	0.160
Median of the ten worst	-1.591	2.209	7.946	8.410

Table A3.4: Overview of number of forecasts for the ten most accurate and the ten least accurate in terms of each accuracy measure.

Overview nmb of individual forecasts	Consensus	ME	MAE	RMSE	MNSE
Mean number of forecasts:	41.80				
Ten best		47.90	25.30	33.90	43.40
Ten worst		17.10	23.40	24.70	27.30
Median number of forecasts:	42				
Ten best		43	23.5	27	32
Ten worst		17.5	21.5	20.5	20

Appendix 4: Rationality tests

Appendix 4.1 Ranking of the ten best individuals in each efficiency test in terms of the other efficiency tests

In the end of 6.2.2 we mention that we rank the individuals who had the highest p-values of being efficient for the different efficiency tests in terms of their ranking in the other efficiency tests. This left us with relatively little overlapping, pictured in the tables A4.1-A4.5 presented in this appendix.

Table A4.1: Bias test: the rankings in the other rationality tests of the individuals ranked highest in terms of the test of bias.

	Individual number	Rank efficiency test 1	Rank efficiency test 2	Rank efficiency test 3	Rank efficiency test 4
Bias test: ten best respondents in terms of ranking in the other tests	472	31	59	74	21
	446	33	79	86	80
	448	41	88	83	139
	145	36	2	19	10
	31	57	32	31	113
	524	9	19	21	15
	424	3	93	-	71
	93	22	42	16	91
	431	19	70	44	57
	158	96	62	20	67

Table A4.2-A4.5: Efficiency tests: the ranking of the ten best individuals in terms of each of the other rationality tests.

	Individual number	Rank test of bias	Rank efficiency test 2	Rank efficiency test 3	Rank efficiency test 4
Efficiency test 1: ten best respondents in terms of ranking in the other tests	429	19	85	103	64
	34	13	9	2	12
	424	7	93	-	72
	502	18	56	33	4
	543	40	72	5	95
	541	38	22	12	42
	65	20	14	25	122
	528	55	25	13	51
	524	6	19	21	15
	527	50	37	7	33

	Individual number	Rank test of bias	Rank efficiency test 1	Rank efficiency test 3	Rank efficiency test 4
Efficiency test 2: ten best respondents in terms of ranking in the other tests	98	12	11	24	55
	145	4	36	19	10
	124	33	62	37	82
	78	29	14	-	14
	546	22	25	14	73
	535	39	37	27	46
	549	17	28	6	52
	507	35	52	40	18
	34	13	2	2	60
	510	43	120	37	60

	Individual number	Rank test of bias	Rank efficiency test 1	Rank efficiency test 2	Rank efficiency test 4
Efficiency test 3: ten best respondents in terms of ranking in the other tests	144	34	64	31	59
	34	13	2	9	12
	465	14	54	113	111
	125	23	65	127	98
	543	40	5	72	95
	549	17	28	7	52
	527	50	10	37	33
	548	64	27	17	35
	500	24	88	77	62
	485	46	14	133	112

	Individual number	Rank test of bias	Rank efficiency test 1	Rank efficiency test 2	Rank efficiency test 3
Efficiency test 4: ten best respondents in terms of ranking in the other tests	488	108	80	81	57
	462	101	68	104	-
	432	98	103	111	-
	502	18	4	56	33
	60	138	130	100	114
	39	85	60	24	-
	42	41	56	47	-
	498	134	118	95	85
	520	53	40	28	60
	145	4	36	2	19

Appendix 4.2: The strong-form rational individuals are late forecasters, and have few responses

In section 6.3.2.2 we perform the efficiency tests on the individual respondents. We find that only four are strong-form efficient on a 5% significance level. These four are presented in this appendix.

There are, as mentioned, only four respondents who are strong-form rational at a 5% significance level. These are the four ones with the highest p-values in terms of the joint null hypothesis for efficiency test four holding, individual number 488, 462, 432 and 502. Table 6.26 in section 6.3.2.2 shows that all these responded quite late in the survey, all after the Philadelphia Fed took over the survey in the second quarter of 1990. The standard deviation of the actual inflation in their forecasting period is thus quite small, indicating that strong-form rationality is easier achieved if the actual inflation level is stable. This finding, indicating an improved forecast performance, is in line with the result in 6.2.2.2 as well as of other literature (Gerberding, 2006; Croushore, 2006). Other patterns are that three out of four seem to be underestimating the inflation, and that they all have made few forecasts. With the highest of these four's number of forecasts being 16, it again looks as if it is an advantage to not respond to the survey a lot of times. This is a rather strange conclusion, indicating no learning among the individual forecasters (again in line with the results found by Lamont (2002)).

Looking at table 6.22 in section 6.3.2.2 we see that these four do not seem to have particularly low accuracy levels and rankings. Only the fourth best, number 502 have relatively low accurate values for each accuracy measure. This is at the same time the only of the four who are overestimating the inflation on average.

Also when investigating their ranks in terms of the other rationality tests, as listed in table A4.1-A4.5 in appendix 4, we do not find any clear pattern of these four having low rankings. For efficiency test three, two of these do not have enough observations for us to be able to rank them in terms of this test. Hence, it does not seem that these strong-form rational forecasters have made especially good forecasts in terms of both accuracy and the other efficiency tests.

Appendix 4.3 An overview of the number of individuals rational and efficient in each industry

In 6.4.6 we sum up our results regarding the industry variables. Table A4.6 summarizes the results of the individual respondents in the different categories as well as for the whole sample this period, showing the total number of individuals and the part of them that pass the different tests.

Table A4.6: An overview of the rationality results of the individual respondents in each industry.

Overview total sample and all industries	All sample this period		Industry1	
	Total number of individuals	Part not rejecting the null on a 5% level	Total number of individuals	Part not rejecting the null on a 5% level
Test of bias	81	0.622	40	0.525
Efficiency test 1	72	0.583	36	0.528
Efficiency test 2	82	0.195	40	0.175
Efficiency test 3	67	0.388	33	0.424
Efficiency test 4	72	0.056	35	0.057
	Industry 2		Industry 3	
	Total number of individuals	Part not rejecting the null on a 5%	Total number of individuals	Part not rejecting the null on a 5%
Test of bias	42	0.667	7	0.714
Efficiency test 1	38	0.525	5	0.800
Efficiency test 2	42	0.262	7	0.143
Efficiency test 3	37	0.270	7	0.800
Efficiency test 4	38	0.053	5	0.000