# Order Matters: An Experimental Study on How Question Ordering Affects Survey-Based Inflation Forecasts*

Maxime Phillot and Rina Rosenblatt-Wisch
Swiss National Bank

Policymakers often rely on survey data when gauging expectations. To know the limits of survey data is thus crucial. We look at inflation expectations as measured through the Deloitte CFO Survey Switzerland and respondents' sensitivity to question ordering thereof. We investigate whether forecast inconsistencies—the discrepancies between point and density forecasts—as well as forecast accuracy change significantly depending on whether the point forecast or the density forecast is asked first. We find that forecast inconsistencies are sizable and order matters. Density forecasts seem to be less affected by question ordering than point forecasts and more accurate than point forecasts.

JEL Codes: E31, E37, E58.

## 1. Introduction

Expectations are key variables in macroeconomics. However, they are hardly measurable. One way to gauge expectations of households, professional forecasters, or firms is in the form of surveys.

*We are grateful to Britta Classen, Michael Grampp, and Dennis Brandes from Deloitte AG Switzerland for supporting our experiment by implementing it in the Deloitte CFO Survey Switzerland and providing us with the data. We would also like to thank Kenza Benhima, Andreas Fischer, Lucas Fuhrer, Jean-Paul Renne, Klaus Schmidt, the editor and two anonymous referees, as well as seminar participants at the SNB for their insightful and most valuable comments. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Swiss National Bank. Maxime Phillot completed part of the project while affiliated with the University of Lausanne. Author contact: Phillot: Swiss National Bank, Boersenstr. 15, CH-8022 Zurich, Switzerland; maxime.phillot@snb.ch. Rosenblatt-Wisch (corresponding author): Swiss National Bank, Boersenstr. 15, CH-8022 Zurich, Switzerland; rina.rosenblatt@snb.ch.

At least since Lucas (1972), economists have been widely assuming rational expectations of agents regarding future macroeconomic variables such as income and inflation. In other words, expectations are generally thought to be objectively and optimally formed given all available information. This also means that agents are assumed to be able at all times and under any circumstances to formulate clear and consistent answers. However, cognitive science has documented that seemingly innocuous factors such as purpose of the surveyor, topics covered, ordinary conversational norms, question length, wording and ordering, and many others can have a significant impact on survey responses; see, e.g., Sudman, Bradburn, and Schwarz (1996) or Tourangeau, Rips, and Rasinski (2000) for a review on the cognitive psychological theory behind surveys. These effects are known as *question effects*. Schuman and Presser (1981) provide insights into many empirical studies and experiments on question effects.

Recent review papers and books surveying the state of expectations' measurement, the literature on the role and nature of expectations in macroeconomics and finance, as well as a growing number of measurement efforts signal that the view is also changing in the economics profession (see, among others, Carroll 2017; Coibion et al. 2018; Coibion, Gorodnichenko, and Kumar 2018; Gennaioli and Shleifer 2018; Manski 2018).

One domain of expectations that has gained increased attention is the one of inflation expectations. For a recent review, see, for instance, Coibion et al. (2020). Inflation expectations are considered an important determinant in the transmission of monetary policy and are therefore closely monitored by central banks. However, to be a useful policy tool, it is important for policymakers to have reliable and robust data on inflation expectations. Knowing about and possibly avoiding question effects and other sources of measurement errors in inflation expectations is therefore crucial.

This paper addresses question effects in inflation expectations. To the best of our knowledge, it is the first one to study question ordering in inflation expectations. We analyze whether question ordering is crucial for forecast inconsistencies, i.e., the discrepancies between point forecasts and measures of central tendency derived from density forecasts, in inflation expectations. To do so, we make use of the Deloitte CFO Survey Switzerland, which contains two questions

about expected inflation in two years' time: the first question asks for a point forecast, and the second question asks for a density forecast. From 2014:Q4 until 2017:Q3 we set up an experiment and randomly assigned the order of these two questions to each of the survey respondents. We first assess whether there exists a persistent discrepancy—a forecast inconsistency—between point forecasts and measures of central tendency derived from density forecasts using non-parametric and parametric techniques. We then study whether these forecast inconsistencies change significantly depending on the specific order in which these two questions are asked, i.e., point forecast first and density forecasts second or vice versa. Finally, we analyze whether the potentially distortional effects of question ordering on consistency are relevant when thinking about forecast accuracy.

We find that (i) forecast inconsistencies are sizable in the data: approximately 18 to 25 percent of all forecasts are inconsistent. We also find that (ii) question ordering matters. Asking for the density forecast before the point forecast results in an approximately 5 percentage point increase in inconsistencies on average, whereby the question ordering affects mainly the answers to the point forecast, while the answers to the density forecast seem to be almost unaffected. In addition, (iii) forecasts are not equally distributed below and above their thresholds of consistency: central tendency measures derived from density forecasts generally reflect lower inflation expectations than point forecasts. This difference is statistically significant mostly for those who are asked the density forecast first. Finally, (iv) the answers to the density forecast question yield higher forecast accuracy than the answers to the point forecast question.

Our paper is in particular related to the following two strands of literature: First, there exists an empirical body in the economic literature that points towards the presence of question effects in inflation expectations. For instance, Bruine de Bruin et al. (2012) study the effect of question wording regarding inflation expectations of households[1] and Arioli et al. (2017) report that survey design such as wording, but also sample design and interview methodology, affect

[1]Initial results can be found in Van der Klaauw et al. (2008) and Bruine de Bruin et al. (2010).

responses to inflation expectations. Coibion et al. (2020) also report the sensitivity of inflation expectations to the design of questions, and Niu and Harvey (2021) analyze how context influences people's judgments in inflation rate surveys.

Second, one strand of literature studies forecast biases and inconsistencies by comparing point forecasts with density forecasts.[2] That point forecasts and measures of central tendency derived from density forecasts do not always match—so-called forecast inconsistencies—was acknowledged first by Engelberg, Manski, and Williams (2009). They assessed consistency using both a nonparametric and a parametric approach. They found that among those point forecasts that are inconsistent with their respective density forecast, a higher proportion underestimates inflation and overestimates GDP growth. Other contributions also point towards the fact that professional forecasters are not necessarily internally consistent and tend to provide point forecasts that are rosier than their density forecast; see, e.g., Garcia and Manzanares (2007), Boero, Smith, and Wallis (2008), or Clements (2009). The early literature comparing point and density forecasts explored uncertainty. Zarnowitz and Lambros (1987) found that the standard deviation of point forecasts tends to understate the mean dispersion of individual density forecasts, although they remain generally positively correlated. Giordani and Söderlind (2003) followed by comparing and discussing the relevance of both measures plus a third one, the variance of aggregate histograms, in capturing uncertainty.[3] They argue that they are all relevant depending on what one wishes to capture and find that disagreement is a reasonable proxy for uncertainty.

The remainder of this paper is structured as follows. Section 2 describes the data and our experiment. Section 3 investigates the effects of question ordering on forecast inconsistencies using nonparametric and parametric methods, and shows and discusses the results. Section 4 analyzes forecast accuracy. Section 5 provides a

[2]For a survey on density forecasts, their applications, evaluations, and limits, see Tay and Wallis (2000).

[3]Aggregate histograms are obtained by averaging over individuals the probability assigned to each bin.

discussion on the interpretation and the limitations of our results. Section 6 concludes.

## 2. Data

### *2.1 The Deloitte CFO Survey*

In this paper, we use data from the Deloitte CFO Survey conducted in Switzerland at a quarterly frequency since the third quarter of 2009. The survey covers the views of chief financial officers (CFOs) and group financial directors of companies in Switzerland from all relevant sectors on their outlook for business, as well as on financing, risks, and strategies. According to Deloitte, the sample is representative of the Swiss economy.

Each quarter, in March, June, September, and December, around 350 firms are contacted via e-mail to fill in the questionnaire. The number of respondents varies each quarter but is usually over 100 firms. The panel of participating CFOs changes over time. According to Deloitte, each quarter, 10 to 30 respondents are completely new to the sample, and the majority are respondents who either have been taking part for only a few surveys or who do not participate regularly. However, for reasons of anonymity, Deloitte does not provide us with the individual identifiers. We are thus unable to exploit any possible panel structure. Thus, we cannot track CFOs over time, and we treat our data set as a repeated cross-sectional study.

The survey is conducted online. It covers 20 questions that recur each quarter and approximately 10 questions unique to the financial conditions of the previous quarter. On the computer screen, the participants only see one question at a time. We thus know the order in which the questions are being presented to the interviewees. The participants do not have to provide answers for all the questions to complete the survey, and are allowed to go back and forth and edit their previous answers.

Since our focus lies on inflation expectations, we will mainly look at the following two questions:

1. In two years' time, what annual rate of inflation, as measured by the Swiss consumer price index, do you expect?

2. In two years' time, where do you expect the annual rate of inflation (Swiss consumer price index) to be?

$$(-\infty, -4], (-4, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], (2, 4], (4, +\infty)$$

The first question asks for a point estimate of two-year-ahead annual inflation rate (in percent), while the second offers a fixed number of intervals for the same rate, to which respondents are requested to assign probabilities. These intervals together form a symmetric eight-bin centered histogram. At the interval level, we interpret missing values as zeros. If the assigned probabilities do not add up to 100 percent, we normalize them so that they add up to 100 percent.[4] Moreover, for our analysis we exclude observations where either answer is missing. Appendix Section A.1 provides additional information about missing observations, the assigned probabilities, and their normalization.[5]

### *2.2 The Experiment*

As of 2014:Q1 we implemented the following experiment together with Deloitte: until then, questions 1 and 2 were always presented in the same order—point forecast first, density forecast second. From 2014:Q1 the order was assigned randomly to participants. The implementation took some time. Until 2014:Q3 we had to assign the ordering manually as follows: First, participants were asked about the point forecast, then about their density forecast. We then switched the ordering after approximately 50 percent of the CFOs whom we expected to participate in the respective quarter concluded the survey. We are fully aware that these two groups might have had quite different information sets each quarter. This in turn could have influenced their answers on inflation expectations. We therefore treat 2014:Q1 until 2014:Q3 as a trial period. From 2014:Q4 onwards, the computer program was adjusted such that the order of the two questions was completely randomized with no manual interference,

[4] All our results are robust to excluding the observations for which the probabilities do not add up to 100 percent.

[5] Appendix Table A.1 provides summary statistics about point forecasts, density forecasts, and firm characteristics.

giving each respondent a true 50 percent chance of seeing the question asking for the point forecast before the question asking for the density forecast or the other way around on their computer screen. The following analysis of inconsistency of the forecasts will be based on the sample with complete randomization, i.e., from 2014:Q4 until 2017:Q3.[6]

## 3. Forecast Inconsistencies

### *3.1 Methodology*

Generally, a forecaster is said to be internally consistent if he or she gives the same answer to two identical questions asked differently. In our case, each point forecast can reasonably be thought to match some statistic derived from the respective subjective probability distribution function underlying expectations over future inflation, which in turn should be summarized in each density forecast. In other words, if we knew each respondent's forecasting model and the statistic reported as the point forecast, we should be able to map density forecasts into point forecasts almost exactly. We could then confidently consider any difference as a *forecast inconsistency*. Unfortunately, with the data at hand we need to make assumptions to match density forecasts with their respective point forecasts.

There are two approaches to assessing consistency between point forecasts and density forecasts: the non-parametric and the parametric one. The non-parametric approach binds consistency by using the edges of each interval given in the survey but makes no further assumption regarding the underlying subjective distribution. The parametric approach, however, explicitly states the shape of the distribution and may rely on fitting techniques to obtain its parameters such as the mean and variance. The fundamental difference between these two methods lies in whether one wishes to assume how the probability mass is distributed *within* each bin. We therefore face a trade-off: While the non-parametric approach provides

[6]Deloitte modified its survey after 2017:Q3: Not only did the frequency change (from quarterly to biannually), but the questions were adjusted to be more in line with CFO surveys the company conducts abroad. We stopped our experiment at that time for this reason.

a more agnostic assessment, it does not give any information as to the *degree* of inconsistency one forecaster might show. In particular, under the non-parametric approach, we are only able to say whether a forecast is consistent or not, whereas the parametric approach tells us exactly by how much.

As for the parametric approach, we will follow Zarnowitz and Lambros (1987). This widely applied approach only assumes that the probability mass of density forecasts is located at the center of each bin. This allows us to compute the *midpoint* of each density forecast, i.e., its subjective mean.[7] A technical requirement however is to close the interval of the first and the last bin. In Question 2 (see Section 2) the first and the last bin is formulated as a one-sided open interval. To close the interval, we attribute the value $-6$ and 6, as it reproduces the length of 2 percentage points of inflation of the intervals, respectively, following and preceding them.[8] A drawback of this methodology is that it over-evaluates the variance under bell-shaped densities. In this respect, the so-called Sheppard's correction may help to obtain a more realistic estimate of the variance but is only computable if the bins are of the same size, which is not the case in the Deloitte CFO Survey. Notwithstanding, because we are not required to accurately evaluate the uncertainty surrounding density forecasts in our setup, we chose to follow the above-mentioned approach for its readability and simplicity.[9]

The following example should illustrate the difference between both approaches: If a forecaster assigns the probabilities 0.3, 0.4, 0.2, 0.1 to the bins $(-1, 0]$, $(0, 1]$, $(1, 2]$, $(2, 4]$, respectively, and 0 elsewhere, then the non-parametric approach binds midpoint consistency between $-1 \cdot 0.3 + 0 \cdot 0.4 + 1 \cdot 0.2 + 2 \cdot 0.1 = 0.1$ and $0 \cdot 0.3 + 1 \cdot 0.4 + 2 \cdot 0.2 + 4 \cdot 0.1 = 1.2$. The forecast is then considered consistent if the point forecast lies within $(0.1, 1.2]$. The

[7] Assuming the mass is uniformly distributed within each bin produces equivalent midpoint estimates.

[8] This choice is virtually irrelevant, since only 2.5 percent of the treatment sample assigned a probability greater than or equal to 10 percent to either of the extreme bins. All our results remain robust for other choices.

[9] Appendix Section A.4 describes an alternative approach which consists in fitting normal distributions to individual density forecasts by numerical optimization as in Giordani and Söderlind (2003). All our results are robust to such methodology, as shown in Section A.6 of the appendix.

lower (upper) bound accounts for the possibility that the forecaster always considered the lowest (highest) value of the bin while reporting the probabilities. By contrast, the parametric approach infers that the subjective midpoint be exactly $-0.5 \cdot 0.3 + 0.5 \cdot 0.4 + 1.5 \cdot 0.2 + 3 \cdot 0.1 = 0.65$, because it supposes that the forecaster always and exclusively considered the center of the bin. Any deviation of the point forecast from this value can then be associated with inconsistency.

As the point estimate question does not specify what statistic of the subjective probability distribution the respondent should report, although the use of the word "expect" points towards the use of expectation or mean as the relevant predictor, forecasters might report the median of their subjective distribution as their point forecast rather than the midpoint. To account for this case, we computed subjective medians as follows. In the non-parametric case, the subjective median is the first interval itself whose cumulative probability is 50 percent or more. In the parametric case, it is the middle of the same interval. By identifying the median in this way, we allow for potentially asymmetric density forecasts. Equivalently, one might be interested in assessing mode consistency. Because this requires further assumptions, we detail such analysis and show the robustness of our results thereto in Table A.4 in Section A.6 of the appendix.

### *3.2 Non-Parametric Approach*

Table 1 displays the results of the non-parametric approach. For each quarter of the experiment and by question ordering, it shows the percentage of respondents that gave a point forecast respectively within, below, or above their respective interval of consistency, evaluated according to the above-described non-parametric subjective midpoints and according to the non-parametric subjective medians. We denote such percentages by $\lambda_i^k$, where $i = P, D$ stands respectively for the group of respondents who were asked for a point forecast or a density forecast first, and $k = c, b, a$ stands respectively for consistent, below, and above. The last row depicts the pooled sample.

Focusing on midpoints, we observe a proportion of consistency that ranges from 74.5 to 96.1 percent for the $P$ group, and from

**Table 1. Midpoint and Median Forecast Consistency by Question Ordering**

| | Subjective Midpoint | | | | | | Subjective Median | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Consistent | | Below | | Above | | Consistent | | Below | | Above | |
| Quarter | $\lambda_P^c$ | $\lambda_D^c$ | $\lambda_P^b$ | $\lambda_D^b$ | $\lambda_P^a$ | $\lambda_D^a$ | $\lambda_P^c$ | $\lambda_D^c$ | $\lambda_P^b$ | $\lambda_D^b$ | $\lambda_P^a$ | $\lambda_D^a$ |
| 2014:Q4 | 80.3 | 82.1 | 9.8 | 1.8 | 9.8 | 16.1 | 72.1 | 71.4 | 14.8 | 1.8 | 13.1 | 26.8 |
| 2015:Q1 | 81.4 | 66.7 | 8.5 | 10.5 | 10.2 | 22.8 | 74.6 | 73.7 | 16.9 | 12.3 | 8.5 | 14.0 |
| 2015:Q2 | 74.5 | 62.0 | 14.5 | 8.0 | 10.9 | 30.0 | 67.3 | 60.0 | 21.8 | 4.0 | 10.9 | 36.0 |
| 2015:Q3 | 83.7 | 73.1 | 4.1 | 3.8 | 12.2 | 23.1 | 79.6 | 73.1 | 8.2 | 5.8 | 12.2 | 21.2 |
| 2015:Q4 | 77.2 | 71.9 | 17.5 | 1.8 | 5.3 | 26.3 | 78.9 | 73.7 | 15.8 | 1.8 | 5.3 | 24.6 |
| 2016:Q1 | 87.7 | 80.4 | 8.8 | 5.9 | 3.5 | 13.7 | 77.2 | 78.4 | 12.3 | 7.8 | 10.5 | 13.7 |
| 2016:Q2 | 76.5 | 81.5 | 17.6 | 3.7 | 5.9 | 14.8 | 52.9 | 64.8 | 29.4 | 7.4 | 17.6 | 27.8 |
| 2016:Q3 | 86.0 | 71.2 | 6.0 | 5.8 | 8.0 | 23.1 | 84.0 | 59.6 | 8.0 | 13.5 | 8.0 | 26.9 |
| 2016:Q4 | 80.4 | 81.3 | 7.8 | 4.2 | 11.8 | 14.6 | 80.4 | 79.2 | 5.9 | 4.2 | 13.7 | 16.7 |
| 2017:Q1 | 80.4 | 76.4 | 5.9 | 0.0 | 13.7 | 23.6 | 76.5 | 72.7 | 9.8 | 1.8 | 13.7 | 25.5 |
| 2017:Q2 | 75.6 | 79.6 | 8.9 | 2.0 | 15.6 | 18.4 | 68.9 | 75.5 | 17.8 | 6.1 | 13.3 | 18.4 |
| 2017:Q3 | 96.1 | 78.0 | 2.0 | 0.0 | 2.0 | 22.0 | 80.4 | 70.0 | 9.8 | 2.0 | 9.8 | 28.0 |
| Pooled | 81.6 | 75.3 | 9.4 | 4.0 | 8.9 | 20.8 | 74.4 | 71.0 | 14.3 | 5.7 | 11.3 | 23.3 |

**Note:** Each cell represents the percentage $\lambda_i^k$ of respondents from group $i = P, D$ falling in the category $k = c, b, a$, for the quarter in row. The subscript $P$ ($D$) denotes the respondents who were asked for a point (density) forecast first. The superscripts $c, b, a$ respectively denote whether the point forecast lies within, below, or above its level of consistency.

62 to 82.1 for the $D$ group. Quarterly consistency between the groups correlates by 19.4 percent, which indicates that time-varying macro factors exert a common pressure on consistency, although in a relatively low manner. Looking at the pooled sample tells us that those respondents who were asked for the point forecast before the density forecast were consistent in 81.6 percent of the cases, whereas those who were asked for the density forecast first were consistent in 75.3 percent of the cases. In other words, consistency (as defined by the non-parametric approach) of the $P$ group exceeded that of the $D$ group, on average, by a 6.3 percentage points margin. This difference of 6.3 percentage points is statistically significant at the 1 percent level as seen in Table A.9 in Section A.7 of the appendix.

More interestingly, a quick inspection of inconsistent forecasts reveals that the proportion of point forecasts that lie above and below their respective level of consistency is quite heterogeneous and depends on question ordering. For those who were asked for a point forecast first, the amount of under-evaluations of point forecasts relative to density forecasts ranges between 2 and 17.6 percent, while this amount ranges only between 0 and 10.5 percent for the other group. Conversely, the proportion of over-evaluated point forecasts varies from 2 to 15.6 percent for the $P$ group, while it goes from 14.6 to as much as 30 percent for the $D$ group.

In total, those who saw the question asking for a point forecast first understated inflation slightly more often (9.4 percent below versus 8.9 percent above), while the others almost systematically overstated inflation (4 percent below versus 20.8 percent above). In other words, being asked for the density forecast before the point forecast not only increases the amount of inconsistency but also makes it more likely for the point forecast to overstate the level of inflation as suggested by the density forecast.

The scrutiny of median non-parametric consistency gives the same general message. Interestingly, consistency occurs less often in the data when we evaluate consistency based on the relationship between point forecasts and subjective medians. This may indicate that forecasters actually link their point forecast to the mean of their density forecast rather than to the median thereof. Interestingly, Meyler and Rubene (2009) and Stark (2013) show a great reliance on judgment when producing a forecast and show that forecasters

are likely to be heterogeneous. The European Central Bank (ECB) and the Federal Reserve Bank of Philadelphia respectively issued a special questionnaire to gauge how their panelists compute and provide their predictions. The ECB reports that interviewees on average weight *judgment* as contributing up to 40 percent of their forecast. Approximately 80 percent of the respondents produce their density forecast solely based on judgment. When asked about which statistic they refer to for their point forecast, approximately 75 percent checked the mean, 20 percent the median, and 7 percent the mode. The Federal Reserve Bank of Philadelphia presents a similar picture: 80 percent of their interviewed panelists revealed that they rely on both mathematical models and judgment to form their forecasts. Notwithstanding, the $P$ group remains more consistent than the $D$ group by 3.4 percentage points. However, this difference is not significant, as we show in Table A.9 in the appendix. Moreover, the pattern in the discrepancies between excessively high and excessively low point forecasts as a function of question ordering is preserved.

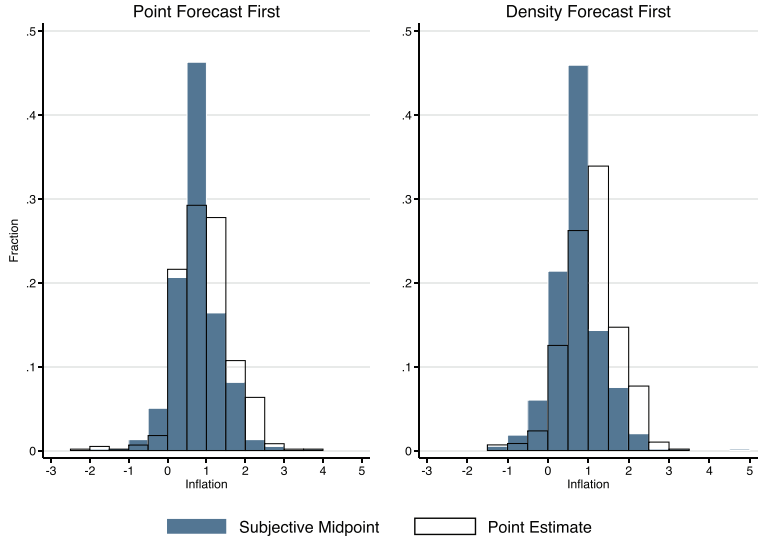
Overall, these results provide evidence that question ordering matters. In particular, asking for a point forecast before a density forecast seems to result in fewer occurrences of inconsistency. Furthermore, it appears that asking first for a point (density) forecast produces a slight (strong) tendency to report point forecasts reflecting a lower (higher) level of inflation than the respective subjective midpoints and medians. Therefore, our non-parametric assessment of consistency indicates that the effect of question ordering is twofold, for it both strongly affects the *amount* of inconsistencies and their *nature*.

### *3.3 Parametric Approach*

The parametric approach allows us to derive from the density forecasts some measures of central tendency that are in levels. However, as detailed above, it requires assumptions. The measure we are focusing on in our analysis is the midpoint, i.e., the subjective mean of density forecasts under the assumption that the probability mass is exactly located at the center of each bin.

Figure 1 plots, for each question ordering, the histogram of subjective midpoints against the histogram of point forecasts in the

**Figure 1. Point Forecasts and Subjective Midpoints by Question Ordering**

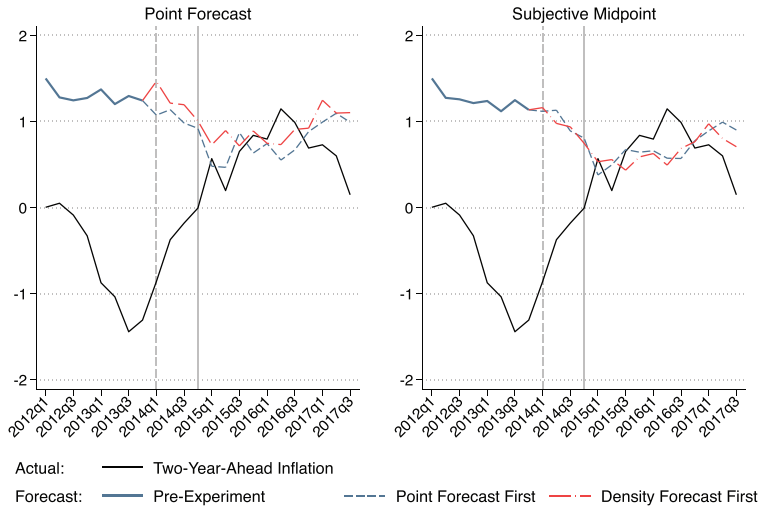



**Note:** The figure plots, for each group, the fraction of respondents (from 2014:Q4 to 2017:Q3) who reported a forecast corresponding to a certain level of inflation (in bins of size 0.5 percentage point), either directly (point forecast, in white) or indirectly (subjective midpoint derived from density forecast, in blue).

pooled sample (from 2014:Q4 to 2017:Q3). In particular, it shows the fraction of respondents who reported a forecast corresponding to a certain level of inflation (in bins of size 0.5 percentage point), either directly (subjective point forecasts, in translucent white) or indirectly (subjective midpoints, in blue).

On the one hand, it appears that the distribution of subjective midpoints is quite homogeneous between the groups (i.e., comparing the left and the right panel), with a fraction of almost 70 percent of all midpoints being comprised between 0 and 1 percent of inflation. On the other hand, however, it seems that the distribution of point forecasts shifts towards the center of the distribution of subjective midpoints when one jumps from the right to the left panel. Indeed, while approximately 50 percent of point forecasts lie between 0 and 1 percent of inflation for the $P$ group, only approximately 35 percent do for the $D$ group.

**Figure 2. Quarterly Point Forecasts and Subjective Midpoints**




**Note:** The dashed vertical line marks the implementation of the experiment, while the solid vertical line marks the starting point of our analysis.

Clearly, forecasters being asked for the point forecast before the density forecast generally give a point forecast that is more in line with the density forecast than forecasters facing the opposite ordering. In other words, we can already confirm the result from the non-parametric analysis, that forecasters tend to be less consistent when they first see the question about the density forecast.

Figure 2 breaks down point forecasts and subjective midpoints by group and quarterly averages, and plots them as a time series along with actual inflation. The dashed vertical line marks the implementation of the experiment (trial period), while the solid vertical line marks the starting point of our analysis (i.e., from 2014:Q4 to 2017:Q3). The black solid line is year-on-year inflation, lagged two years. The blue solid line represents quarterly averages of point forecasts, respectively subjective midpoints for the pre-experiment period. The blue dashed line depicts quarterly averages of point forecasts, respectively subjective midpoints for the $P$ group and the red dashed-dotted line the ones for the $D$ group. Recall that

before the implementation of the experiment (i.e., from 2012:Q1 to 2014:Q1), point forecasts were always asked before density forecasts. The blue dashed line, which shows the results of the group that sees the point forecast question first ($P$ group), can therefore be expected to follow the pattern of the solid blue line (like a control group)—any deviation of the red dashed-dotted line from the blue dashed line can thus be interpreted as the effect of flipping the question ordering (i.e., the treatment effect). Comparing the point forecasts between the two groups shows that the average point forecast of the $D$ group is somewhat persistently higher than that of the $P$ group (Figure 2, left panel). For the average midpoint, one can barely distinguish the two series (Figure 2, right panel).

Table 2 formalizes these observations for the pooled series. First, it shows the sample mean point forecast of the $D$ and the $P$ group and the respective standard deviation. The 0.16 percentage point difference in the mean point forecasts between the two groups is statistically significant, while the 0.94 ratio between their respective standard deviations is not.[10]

Second, Table 2 shows the mean probabilities assigned to each bin for both groups. The mean probabilities assigned to each bin can never be said to differ significantly between the groups. However, we reject equal variance of the assigned probability between groups for all but two of the eight bins. Interestingly, these two bins together comprise inflation from above zero to below 2 percent, and account on average for more than 70 percent of cumulated probability. Given that the Swiss National Bank defines price stability as an annual inflation rate below 2 percent and non-negative, this result suggests that credibility by forecasters about the capacity of the central bank to achieve its target is not affected by question ordering. In other words, inflation expectations seem to be too well anchored regarding "normal territories" for question ordering to affect the

[10]In Section A.9 of the appendix, we compare one-year-ahead exchange rate point forecasts, which are also elicited in the survey, between the two groups. This placebo test aims at putting the significant differences found in Table 3 in perspective: Significant differences in exchange rate forecasts between the two groups would cast doubt on the validity of our experiment. Table A.11 in the appendix shows no such pattern and strengthens the direct link between inflation-related differences and question ordering.

## Table 2. Comparison between Groups

| Variable | Mean | | | Std. Dev. | | |
|---|---|---|---|---|---|---|
| | $\mu_D$ | $\mu_P$ | $\mu_D - \mu_P$ | $\sigma_D$ | $\sigma_P$ | $\sigma_D/\sigma_P$ |
| *Inflation Expectations* | | | | | | |
| **Point Forecast** *Two-years-ahead inflation expectation* | 0.92 | 0.76 | 0.16*** (4.3) | 0.63 | 0.67 | 0.94 (0.9) |
| **Density Forecast** *Probability that two-years-ahead inflation lie within. . .* | | | | | | |
| $(-\infty, -4]$ | 0.11 | 0.07 | 0.04 (0.8) | 1.29 | 0.46 | 2.8*** (8.0) |
| $(-4, -2]$ | 0.45 | 0.64 | –0.19 (–1.5) | 1.59 | 2.75 | 0.58*** (0.3) |
| $(-2, -1]$ | 3.46 | 3.19 | 0.27 (0.6) | 8.67 | 7.60 | 1.14*** (1.3) |
| $(-1, 0]$ | 17.49 | 16.19 | 1.3 (1.3) | 18.28 | 16.33 | 1.12** (1.2) |
| $(0, 1]$ | 47.27 | 47.54 | –0.27 (–0.2) | 24.95 | 24.49 | 1.02 (1.0) |
| $(1, 2]$ | 24.37 | 25.82 | –1.45 (–1.2) | 20.82 | 21.12 | 0.99 (0.9) |
| $(2, 4]$ | 6.14 | 5.93 | 0.21 (0.4) | 10.02 | 9.18 | 1.09* (1.2) |
| $(4, +\infty)$ | 0.71 | 0.62 | 0.09 (0.4) | 4.46 | 2.85 | 1.56*** (2.5) |
| **Subjective Midpoint** | 0.66 | 0.68 | –0.02 (–0.6) | 0.61 | 0.60 | 1.03 (1.1) |

**Note:** This table shows the sample mean ($\mu_i$) and the sample standard deviation $\sigma_i$, from group $i = P, D$ for the sample between 2014:Q4 and 2017:Q3, i.e., during the experiment. $P(D)$ denotes the respondents who were asked for a point (density) forecast first. There are 637 (631) observations in the P (D) group. $t$ and $F$ statistics respectively for the mean- and variance-comparison tests are given in parentheses. Significance levels: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

variability of its associated probability (which, as we noted earlier, is centered around the same value for both groups). What question ordering does seem to affect is the (lack of) consensus as to the probability of rarer events, i.e., disinflation and high inflation. Nonetheless, identifying a pattern is difficult, for the significant differences in standard deviations between the two groups only reflect a cumulated probability of 30 percent.

Third, Table 2 reports the sample mean subjective midpoints of the density forecasts for both groups. The –0.02 percentage point difference in the sample mean midpoint between the two groups is not statistically significant.

Overall, the answers to the density forecast question seem to be less affected by question ordering than the answers to the point forecast question.

Nevertheless, to give a formal appraisal of the average treatment effect and its significance, we need to go one step further and compare the average inconsistencies *between* the two orderings. This is comparable to a difference-in-differences approach: because point forecasts are on average higher than subjective midpoints for both groups as shown in Figure 2, only the difference between the respective discrepancy captures the causal effect of question ordering. To this end, Table 3 summarizes by quarter the number of respondents $N_i$ and the average forecast inconsistency $\Delta_i$ for each group $i = P, D$ as well as the difference thereof, which captures the average treatment effect. The last column displays the $p$-value of the $t$-test that this difference $\Delta_D - \Delta_P$ is positive, under the null hypothesis that it is zero (assuming equal variances).

For every quarter of the experiment, the average treatment effect is positive. In 7 out of 12 quarters, it is significantly so at the 95 percent level. The pooled sample tells us that the discrepancy between point forecasts and subjective midpoints is on average positive and significantly higher by 0.18 percentage point of inflation for the $D$ group than for the $P$ group.[11] Clearly, imposing an alternative

[11] To put these numbers into perspective: In a historical and international comparison, Switzerland has low inflation (and interest) rates. The Swiss National Bank's primary goal is to ensure price stability and, as mentioned above, it defines price stability as a consumer price index (CPI) rate less than 2 percent per year and non-negative. Between January 1995 and September 2017, Swiss CPI

**Table 3. Forecast Inconsistencies and Treatment Effect**

| | Obs. | | Inconsistency | | Treatment Effect | |
|---|---|---|---|---|---|---|
| **Quarter** | $N_D$ | $N_P$ | $\Delta_D$ | $\Delta_P$ | $\Delta_D - \Delta_P$ | $p$-value |
| 2014:Q4 | 56 | 61 | 0.26 | 0.10 | 0.16 | 0.04 |
| 2015:Q1 | 57 | 59 | 0.19 | 0.10 | 0.09 | 0.20 |
| 2015:Q2 | 50 | 55 | 0.35 | –0.03 | 0.38 | 0.00 |
| 2015:Q3 | 52 | 49 | 0.30 | 0.20 | 0.10 | 0.15 |
| 2015:Q4 | 57 | 57 | 0.28 | 0.00 | 0.28 | 0.00 |
| 2016:Q1 | 51 | 57 | 0.13 | 0.08 | 0.05 | 0.29 |
| 2016:Q2 | 54 | 51 | 0.22 | –0.02 | 0.24 | 0.00 |
| 2016:Q3 | 52 | 50 | 0.22 | 0.12 | 0.10 | 0.21 |
| 2016:Q4 | 48 | 51 | 0.16 | 0.10 | 0.06 | 0.21 |
| 2017:Q1 | 55 | 51 | 0.25 | 0.10 | 0.15 | 0.04 |
| 2017:Q2 | 49 | 45 | 0.30 | 0.10 | 0.20 | 0.03 |
| 2017:Q3 | 50 | 51 | 0.39 | 0.07 | 0.32 | 0.00 |
| Pooled | 631 | 637 | 0.26 | 0.08 | 0.18 | 0.00 |

**Note:** The table displays, for each quarter of the experiment, the number of respondents $N_i$ and the average inconsistency $\Delta_i$ for each group $i = P, D$ as well as the difference thereof. The last column displays the $p$-value of the $t$-test that this difference $\Delta_D - \Delta_P$ is positive, under the null hypothesis that it is zero (assuming equal variances). The last row considers the pooled sample.

ordering by asking a density forecast before a point forecast causes forecast inconsistencies to widen significantly.

Thus, the results from the parametric approach confirm those of the non-parametric one by pointing towards the presence of question effects in surveys about inflation expectations. In particular, we find that asking for the density forecast before the point forecast results almost systematically in a statistically significant discrepancy between point forecasts and midpoints, with point forecasts overstating the level of inflation suggested by the density forecast. By contrast, asking for the point forecast first appears to produce

year-on-year inflation was on average 0.5 percent with a standard deviation of 0.9 percent. Between January 2012 and September 2017 (sample covered in this paper), Swiss CPI year-on-year inflation was on average –0.37 percent with a standard deviation of 0.55 percent. In view of the low inflation environment in Switzerland, the average treatment effect shown in Table 3 seems to be significant also from an economic perspective.

differences between midpoints and point forecasts that are of no statistical significance.[12]

All in all, we find marked evidence that question ordering distorts the internal consistency of two-year-ahead inflation forecasts: Question ordering not only affects the *amount* of inconsistencies, it also influences the *direction* in which the mismatch occurs. If question ordering affects consistency, is there anything to say about forecast accuracy? The next section sheds some light on this question.

## 4. Forecast Accuracy

We so far concentrated on the potentially distortionary effects of question ordering on consistency and saw that the answers to the density forecast question seemed to be less affected by question ordering than the answers to the point forecast question. Notwithstanding, and as far as policymakers are concerned, forecast *accuracy* matters when it comes to policymaking. Since central banks use inflation forecasts as intermediary targets, robustness and accuracy of their forecasts are desirable features.[13] Thus, robustness and accuracy of survey-based inflation expectations either serving as an input variable in forecasting inflation or serving as a forecast themselves should be a plus. A (potential) constant bias is either captured by the regression's intercept when estimated in levels or disappears in a regression when estimated in first differences. On the contrary, if answers are not robust over time due to, e.g., question effects, forecasting with such answers encompasses more uncertainty and might be misleading. However, even though we observed that the answers to the density forecast question seemed to be less affected by question ordering, it is not a priori certain if these answers forecast inflation more accurately than the answers to the point forecast do.

[12] In Section A.3 of the appendix we investigate the relationship between consistency and firm characteristics, such as the size of the firm or the economic sector. We find that characteristics such as uncertainty, firm size, and economic sector seem to play a role too: bigger firms from the service sector tend to be more consistent, and higher uncertainty is associated with more inconsistencies.

[13] For an early argument on the use of forecasts in policymaking, see Svensson (1997). For the theoretical limitations of such use, see Bernanke and Woodford (1997).

To test for forecast accuracy, we focus on the point forecasts and subjective midpoints of the density forecasts; thus we focus on our parametric assessment. As laid out in Section 2, the questions we cover ask about annual inflation in two years' time and the survey is conducted in March, June, September, and December. We therefore take as a reference value for realized inflation $\pi$ the 24-month-ahead year-on-year change of the Swiss consumer price index (CPI) in the respective month:

$$\pi^m_{t+24} = \frac{CPI^m_{t+24} - CPI^m_{t+12}}{CPI^m_{t+12}}, \tag{1}$$

where $m$ represents March, June, September, December and $t$ is time.[14]

Looking at Figure 2, we observe that both point forecasts and subjective midpoints overestimated inflation until the beginning of 2015 and were more aligned thereafter. The average point forecast of the $D$ group is somewhat persistently higher than that of the $P$ group. For the average midpoint, one can barely distinguish the two series.

As is standard in the forecast literature, we follow Diebold and Mariano (2002, hereafter DM) in order to determine which forecast is more accurate. Key to this approach is its account for serial correlation in the long-run variance (as opposed to regular $t$-tests).[15]

This leaves us with four different forecasts (two types of questions, i.e., point forecast (PF) or density forecast (DF), and two groups, i.e., point forecast first (P) or density forecast first (D), and six unique pairwise comparisons. Table 4 shows the test results. Each entry displays the column forecast mean squared error (MSE) minus the row forecast MSE. $(i, j)$ denotes forecast $i = PF, DF$ made by group $j = P, D$. A positive value reflects a relatively higher prediction error of the column forecast, and hence, higher accuracy of the row forecast.

[14]The question is formulated in a rather vague manner regarding realized inflation. We therefore also performed our calculations with different measures of realized inflation such as, e.g., year-on-year change of quarterly averages of the CPI in two years' time. Our results on accuracy remained robust to these changes.

[15]Appendix Section A.5 provides methodological details about DM tests.

**Table 4. Diebold-Mariano Tests for Predictive Accuracy**

| **Median Forecast** | | | | |
|---|---|---|---|---|
| | (*PF,P*) | (*PF,D*) | (*DF,P*) | (*DF,D*) |
| (*PF,P*) | | | | |
| (*PF,D*) | –0.04431 | | | |
| (*DF,P*) | 0.8875** | 0.1331$^{\dagger}$ | | |
| (*DF,D*) | 0.1282** | 0.1725* | 0.03945* | |
| **Mean Forecast** | | | | |
| | (*PF,P*) | (*PF,D*) | (*DF,P*) | (*DF,D*) |
| (*PF,P*) | | | | |
| (*PF,D*) | –0.05476$^{\dagger}$ | | | |
| (*DF,P*) | 0.0378* | 0.09205$^{\dagger}$ | | |
| (*DF,D*) | 0.08258 | 0.1195$^{\dagger}$ | 0.02749 | |

**Note:** $^{\dagger}p < 0.10$, $*p < 0.05$, $**p < 0.01$. Entries show the column forecast MSE minus the row forecast MSE. $(i, j)$ denotes forecast $i = PF, DF$ made by group $j = P, D$. Positive values imply higher accuracy of the row forecast.

The top panel of Table 4 shows median forecasts, while the bottom panel displays results for mean forecasts. If we compare (DF,.) with (PF,.) we observe that the subjective midpoint forecast is always more accurate than the point forecast, no matter whether the point forecast was asked first or second. Furthermore, comparing (PF,D) with (PF,P) and (DF,D) with (DF,P) indicates that to each question being asked first, the corresponding forecast yields higher accuracy.

From the analysis above, we know that point forecasts and density forecasts are closer to each other or more consistent when the point forecast is asked first. Moreover, asking for the point forecast first makes the point forecast slightly more accurate ((PF,D) versus (PF,P) in Table 4). However, again from Table 4, density forecasts are also more accurate when being asked first. This points towards the following trade-off: consistency of both forecasts together comes at the cost of subjective midpoint accuracy. Notwithstanding, Table 4 median and mean entries also show us that the gain of asking for density forecasts first rather than second lies between 0.027 and 0.039 in MSE terms ((DF,D) and (DF,P) entries)), while

that of asking for point forecasts first rather than second lies between 0.044 and 0.055 in MSE terms ((PF,D) and (PF,P) entries). Thus, asking for the point forecast first not only improves consistency, but also yields the higher benefits in MSE terms. Furthermore, and as already noted, Table 4 shows that density forecasts are still more accurate when being asked second than point forecasts when being asked first or second (see MSE of (DF,.) and (PF,.)).

Existing surveys come in different ways. The Federal Reserve Bank of Philadelphia Survey of Professional Forecasters (US-SPF) or the ECB Survey of Professional Forecasters (ECB-SPF) include both types of questions but present them to the respondents in different ways. While the US-SPF presents the point and subsequently the density forecast on different pages of the survey, the ECB-SPF asks both types of questions on the same page. Results of these answers are, e.g., used in forecasting and modeling: Ang, Bekaert, and Wei (2007), for instance, show that surveys are successful in forecasting inflation. They use, among other measures, inflation expectations of the US-SPF. Grishchenko, Mouabbi, and Renne (2019) include point and density forecasts of the US-SPF and density forecasts of the ECB-SPF when constructing inflation expectations, inflation uncertainty, and inflation-anchoring measures for the United States and the euro area. Our insights might be of practical relevance when designing new surveys or using existing ones. Some awareness of possible question effects might be indicated. Our findings suggest that the answers to the density forecast question seem to be less affected by question ordering than the answers to the point forecast question. In addition, in terms of forecast accuracy, the density forecasts seem to outperform the point forecasts. When both questions are being asked, our results indicate that one should ask for the point forecast first.

## 5. Discussion

Are our results in line with the literature on question effects we laid out in Section 1? Note, all surveys being analyzed so far had the same ordering: point forecast first, density forecast second. In line with this literature we find that forecast inconsistencies persistently occur. The literature on forecast inconsistencies also finds that point forecasts tend to underestimate inflation with respect to

their density forecast. Our findings regarding underestimation for the ordering point forecast first, density forecasts second are mixed. Our results of the non-parametric approach tend slightly towards underestimation, while our results of the parametric approach tend slightly towards overestimation.

Our contribution to this literature lies in the additional insight question ordering brings to the debate. When we switch the order of the questions, i.e., when the question about the density forecast precedes the question about the point forecast, we detect a clear overestimation both for the non-parametric and for the parametric approach. We observe that mainly the answers to the point forecast were affected by the switch in the order, while the answers to the density forecasts remained rather unaffected.

As mentioned in Section 2, our respondents are allowed to go back and forth when answering the questions. In addition, respondents can be in the sample repeatedly. Although the panel of participating CFOs changes over time, as also mentioned in Section 2, and there are 10 to 30 newcomers each quarter, we do not know how many and which CFOs repeatedly participated in the survey. We cannot track the number of "treatments" received by a given CFO, nor the length of the treatments. Some firms might have been presented repeatedly the density question first, while others might have been asked for the point forecast first. Others might have been switching constantly between the two questions. The data at hand do not allow us to exploit any possible panel structure.

One may be worried that the treatment effect weakens over time due to some learning process; see, e.g., Kim and Binder (2020).[16] This could indeed be the case. If some respondents edit their previous answers to improve the consistency of their answers, and if some respondents already know that both types of questions will be asked, this should work against our findings, since it should translate into fewer inconsistencies. Our average treatment effects estimates could therefore be interpreted as lower bounds on forecast inconsistency.

[16]Potential misspecifications arise in particular once we pool our panel data without properly accounting for individual-specific effects, which we however cannot observe. The inclusion of time fixed effects and clustered standard errors at the pseudo-individual level in the regressions in Sections A.3 and A.6.3 of the appendix control—however, only partly—for this potential issue.

Despite the fact that there are newcomers each quarter in the panel of participating CFOs and the order of the two questions was completely randomized with no manual interference, giving each respondent a true 50 percent chance of seeing the question asking for the point forecast before the question asking for the density forecast or the other way around on their computer screen, as time goes by, the expected number of "treatments" received should increase in the sample. A time trend could be a natural, albeit also imperfect proxy for a weakening of the treatment effects. Yet, we could not observe such a trend in our data.[17]

What is known as rounding or heaping at round numbers (see, e.g., Manski and Molinari 2010) could possibly influence our reported amount of inconsistencies. If it does, the observed amount could either increase or decrease. However, Gideon, Helppie-McFall, and Hsu (2017) show that patterns of rounding are not driven by question order in the context of financial questions. It is the difficulty of the question that affects rounding behavior. The response rate to both questions in detail described in Appendix Section A.1 does not give an indication that this is an issue—non-responses to the density forecast occurred, for example, equally often as those to the point forecast.

Could some form of anchoring be at play? One may argue that for a respondent who first sees the question about the density forecast, the point estimate will likely be anchored to the range that was shown in the density question. If anchoring is at play, we would expect the following for those who see the question about the density forecast first: (i) a lower variance of the point forecasts and (ii) a distribution of the point forecasts centered around zero due to the symmetry of the bins. It appears that the latter hypothesis can be discarded by looking at Table 2. The mean point forecasts of the $D$ group are further away from zero than those of the $P$ group. The former hypothesis seems to apply to a certain extent. The point estimates from respondents who see the question about the density forecast first have an overall standard deviation of 0.04 lower (in terms of inflation) than the other group. However, this difference is not statistically significant.[18] As far as the data can tell, anchoring

[17] See, e.g., Figure 2 and Table 3.

[18] Actually, the null hypothesis of equal variance cannot be rejected in any single quarter.

did not cause the point estimates to be more narrowly distributed. Of course, to study the effects of anchoring thoroughly, we would have to run further treatments. However, this goes beyond the scope of this paper and the data at hand do not allow us to draw further conclusions.

All in all, we find evidence that question ordering distorts the internal consistency of two-year-ahead inflation forecasts: Question ordering not only affects the *amount* of inconsistencies, it also influences the *direction* in which the mismatch occurs. But again, we only tested along the dimension of question ordering: With respect to question ordering, we found the answers to the density forecast question to be less sensitive than the ones to the point forecasts question. In terms of forecast accuracy, the answers to the density forecast question outperformed the answers to the point forecast question.

## 6. Conclusions

We showed that question ordering matters in economic surveys and is relevant for questions on inflation expectations. While the answers to the point forecast question were sensitive to the order in the survey, the answers to the density forecast question were basically unaffected. We found that inconsistencies between the point forecasts and measures of central tendency derived from density forecasts are sizable in the data and are increased if respondents see the question about the density forecast before the one about the point forecast. In terms of forecast accuracy, the answers to the density forecasts seem to outperform the answers to the point forecasts.

These results suggest that the design of surveys also matters in regard to economics. When gauging expectations on macroeconomic variables from surveys, policymakers and market participants alike should be aware that biases due to question effects might be at play. This should not imply that surveys are not a useful policy instrument; on the contrary, they deliver additional information compared to market data, or sometimes cover areas where no market data exist.

## Appendix

### *A.1 Data*

We compile our data by assembling Deloitte's quarterly surveys into a larger data set. Because we focus on forecast inconsistencies, we drop all observations for which either the point forecast or the density forecast on inflation expectations is missing (the survey does not force the box to be filled in). This occurred 246 times out of 1,514 for the experiment sample, and 129 out of 1,251 for the pre-experiment one. The number of respondents who answered neither question was 286 (of which 202 were from the experiment sample); 48 respondents (of which 24 were from the experiment sample) answered only the density forecast question; and 41 (of which 20 were from the experiment sample) answered only the point forecast question. A missing density forecast occurs when none of the intervals is used. When at least one interval contains a positive probability, we interpret unused intervals as zero-probability intervals.

Furthermore, the probabilities assigned to the intervals occasionally do not add up to a 100 percent (the survey does not require answers to do so). For 93.5 percent of all observations, however, the probabilities add up to 100 percent. For 97.7 percent of them, their sum is comprised between 90 and 110 or is equal to one. All the remaining observations range between 0.3 and 500. Nevertheless, to conserve the full information of our sample, we normalize all the probabilities so that they add up to 100 percent.

In addition to inflation expectations, the questionnaire provides information on the responding firm. In particular, three questions allow us to know more about the size, the openness, and the sector of the firm:

3. What was your company's turnover in the last financial year?

4. How much of your company's revenues are earned outside Switzerland?

5. In which sector does your company primarily operate?

Question 3 offers several intervals that we group into two categories: less than CHF 500 million (*low* turnover), and CHF 500

million or more (*high* turnover). Question 4 answers are regrouped as follows: less than one-third (*low* share), and one-third or more (*high* share). Question 5 suggests a list of several "sectors" from which respondents are allowed to select more than one answer. We group all combinations into three sectors: construction, manufacturing, and services.[19]

Table A.1 shows the summary statistics of the data we analyze. It shows the number of observations that were first assigned the point—respectively, the density—forecast and their respective sample mean. It also gives an overview of the average assigned probability for each bin of the density forecast. In addition, it reports details regarding the turnover, openness, and sector of the firms. The statistical analysis of the differences between the group that was first asked a density forecast and the group that was first asked a point forecast and of their forecast inconsistencies is the subject of the following section.

### *A.2 Visual Parametric Approach*

In a next step, we analyze the differences between point forecasts and subjective midpoints within each group, i.e., the so-called forecast inconsistencies. The panels on the left of Figure A.1 plot as a time series the quarterly averages of point forecasts (blue dashed lines) and subjective midpoints (green dashed lines), as well as the differences between the two (i.e., the forecast inconsistencies, red dotted lines) respectively for the pre-experiment sample (top panels) and the experiment sample broken down by question ordering (middle and bottom panels).[20] The red dashed lines surrounding the series of mean forecast inconsistencies are the lower and upper bounds of the 95 percent confidence interval (CI) for the difference between

[19] The groups are constructed to match the statistical classification of economic activities in the European Community (NACE) at best.

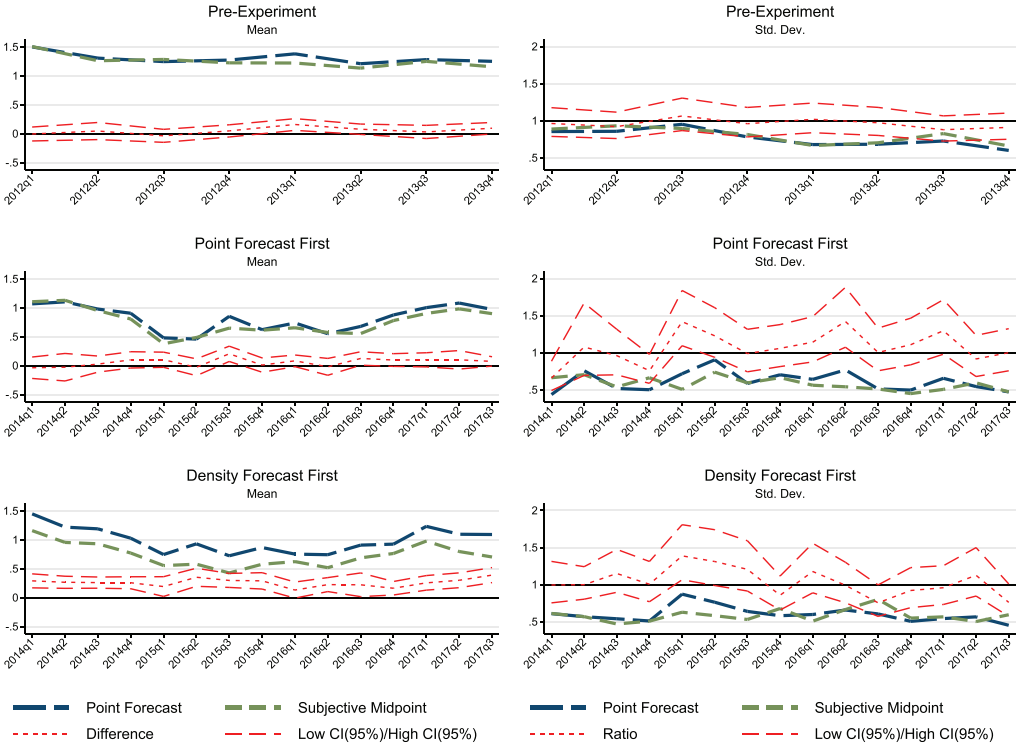
[20] Note that Figure A.1 also shows the trial period of the experiment (2014:Q1–2014:Q3) for each group, although, as we argued before, it does not provide a reliable assessment of the treatment effect. All the comments exposed here therefore do *not* consider this period, despite the robustness in doing so.

## Table A.1. Deloitte CFO Survey Summary Statistics

| Variable | Observations | | | Mean | | | Std. Dev. | | |
|---|---|---|---|---|---|---|---|---|---|
| | $N$ | $N_D$ | $N_P$ | $\mu$ | $\mu_D$ | $\mu_P$ | $\sigma$ | $\sigma_D$ | $\sigma_P$ |
| *Inflation Expectations* | | | | | | | | | |
| **Point Forecast** | **1,268** | **631** | **637** | 0.84 | 0.92 | 0.76 | 0.65 | 0.63 | 0.67 |
| *Two-years-ahead inflation expectation* | | | | | | | | | |
| **Density Forecast** | **1,268** | **631** | **637** | | | | | | |
| *Probability that two-years-ahead inflation lie within. . .* | | | | | | | | | |
| $(-\infty, -4]$ | | | | 0.09 | 0.11 | 0.07 | 0.96 | 1.29 | 0.46 |
| $(-4, -2]$ | | | | 0.54 | 0.45 | 0.64 | 2.25 | 1.59 | 2.75 |
| $(-2, -1]$ | | | | 3.33 | 3.46 | 3.19 | 8.15 | 8.67 | 7.60 |
| $(-1, 0]$ | | | | 16.83 | 17.49 | 16.19 | 17.33 | 18.28 | 16.33 |
| $(0, 1]$ | | | | 47.41 | 47.27 | 47.54 | 24.71 | 24.95 | 24.49 |
| $(1, 2]$ | | | | 25.09 | 24.37 | 25.82 | 20.98 | 20.82 | 21.12 |
| $(2, 4]$ | | | | 6.04 | 6.14 | 5.93 | 9.61 | 10.02 | 9.18 |
| $(4, +\infty)$ | | | | 0.66 | 0.71 | 0.62 | 3.73 | 4.46 | 2.85 |
| *Attributes* | | | | | | | | | |
| **Turnover** | **1,255** | **625** | **630** | | | | | | |
| *For the last financial year (millions CHF)* | | | | | | | | | |
| [0, 50) | 240 | 129 | 111 | | | | | | |
| (50, 100] | 191 | 86 | 105 | | | | | | |
| (100, 500] | 378 | 180 | 198 | | | | | | |
| (500, 1000] | 148 | 74 | 74 | | | | | | |
| (1000, $+\infty$) | 298 | 156 | 142 | | | | | | |
| **Openness** | **1,222** | **616** | **606** | | | | | | |
| *Share of revenues earned abroad* | | | | | | | | | |
| [0, 1/3) | 517 | 256 | 261 | | | | | | |
| [1/3, 2/3) | 118 | 56 | 62 | | | | | | |
| [2/3, 1] | 587 | 304 | 283 | | | | | | |
| **Sector** | **1,256** | **624** | **632** | | | | | | |
| Construction | 103 | 56 | 47 | | | | | | |
| Manufacturing | 590 | 294 | 296 | | | | | | |
| Services | 563 | 274 | 289 | | | | | | |

**Note:** Each cell represents the number of observations ($N_i$), the sample mean ($\mu_i$), or the sample standard deviation $\sigma_i$ from group $i = P, D$ for the sample between 2014:Q4 and 2017:Q3, i.e., during the experiment. $P$ ($D$) denotes the respondents who were asked for a point (density) forecast first.

**Figure A.1. The Effect of Question Ordering on Forecast Inconsistencies**




**Note:** Left panels plot quarterly averages of point forecasts and subjective midpoints, as well as their differences together with the 95 percent CI bands thereof assuming equal variances. Right panels plot quarterly standard deviations of the same variables, as well as their ratio together with the 95 percent CI bands thereof. Each row considers a different subsample: pre-experiment (2012:Q1–2013:Q4) and experiment (2014:Q1–2017:Q3) by question ordering.

the mean of point forecasts and the mean of subjective midpoints, computed separately for each quarter through two-sample mean-comparison $t$-tests assuming equal variances. In a very similar fashion, the right panels show the quarterly standard deviations of point forecasts and subjective midpoints as well as their ratios. The dashed lines surrounding these ratios are the bounds of the 95 percent CI

thereof, computed separately for each quarter through two-sample variance-comparison $F$-tests.

Focusing first on the left panels of Figure A.1 allows us to assess the effect of question ordering on forecast consistency. A quick comparison between the top and middle panels tells us that the $P$ group indeed follows the pattern of the pre-experiment sample. In fact, similar to prior to the experiment, those who submitted a point forecast first during the experiment provided on average point forecasts sometimes higher, sometimes lower than their respective midpoints, but for a difference that can almost never be considered significantly different from zero.

By contrast, forecasters from the $D$ group systematically submitted point forecasts that were higher on average than their subjective midpoints. Strikingly, this overstatement of inflation made by point forecasts relative to density forecasts is statistically significant at the quarterly level for almost every period. We thus observe a strong treatment effect: switching the order by asking for the density forecast before the point forecast exerts an upward pressure on point forecasts relative to midpoints, thereby producing an increase in forecast inconsistencies.

Finally, looking at the right panels of Figure A.1 provides an indication of the plausibility of our results. The standard deviation of point forecasts (or subjective midpoints) is a measure of *disagreement* and is often used in the literature as a proxy for general uncertainty.[21] We interpret the quasi-permanent conservation of the null hypothesis (i.e., that standard deviations are equal) for both question orderings as evidence that question ordering affects the amount of inconsistencies, but not the general level of disagreement among forecasters. In other words, asking for a density forecast first intensifies the discrepancies between point forecasts and midpoints, but does so without distorting their respective dispersion. We can thus exclude that the experiment itself came as a surprise, which would in turn drive our results.

[21] Because the quarterly sample size of each question ordering is half the size of the pre-experiment sample, the volatility of the series becomes mechanically lower.

**Table A.2. Recoding Attributes in Binary Variables**

| Variable | Observations | | |
|---|---|---|---|
| | $N$ | $N_D$ | $N_P$ |
| *Attributes* | | | |
| **Turnover** | **1,255** | **625** | **630** |
| *For the last financial year (millions CHF)* | | | |
| $0 = [0, 500)$ | 809 | 395 | 414 |
| $1 = [500, +\infty)$ | 446 | 230 | 216 |
| **Openness** | **1,222** | **616** | **606** |
| *Share of revenues earned abroad* | | | |
| $1 = [0, 1/3)$ | 517 | 256 | 261 |
| $0 = [1/3, 1]$ | 705 | 360 | 345 |
| **Sector** | **1,256** | **624** | **632** |
| 0 = Construction & Manufacturing | 693 | 350 | 343 |
| 1 = Services | 563 | 274 | 289 |

**Note:** The table displays the number of observations for each attribute after recoding them into binary variables. For the original data, see Table A.1.

### *A.3 Regression Analysis*

What drives forecast inconsistencies? As we have already noted, question ordering does. However, other factors such as firm characteristics or uncertainty might very well be influencing the discrepancy between density forecasts and point forecasts. To address this question, we make use of the firm's attributes present in our data, define a measure of uncertainty, and estimate two models: a logistic regression and a linear regression.

Recall that we have information about the turnover, the share of revenues earned abroad, and the operating sector of the respondent's firm. To be parsimonious, we recode these three attributes into binary variables. For each of them, Table A.2 displays the threshold we chose as well as the number of observations falling in each category by question ordering. Note that neither group is over- or underrepresented in terms of their question

ordering, so that we can exclude that attrition is correlated with the attributes.[22]

We coded the dummy variables so that we expect the value 1 to be associated with more consistency. First, we consider a turnover greater than or equal to CHF 500 million to be a high turnover. A higher turnover should reflect a higher size and access to better data, or a greater need for quality forecasts. Second, we define a share of revenues earned abroad between zero and one-third as low openness. Arguably, a domestically oriented firm is more likely to depend on national rather than international prospects, and thus to monitor local prices accurately. Finally, firms from the services sector could be associated with higher levels of technology or financial market knowledge, and thus with more rigorous forecasts.

As a measure of uncertainty at the individual level, we argue as Clements (2010) that the number of bins that are assigned a positive probability by the respondent is a good proxy for the variance of the density function underlying forecasters' expectations over future inflation. The advantage of this measure is that it is non-parametric and readily available.[23] Similarly, we recode this variable as a dummy whose value is 1 if the number of bins used by the respondent is lower than or equal to 3 (i.e., if the forecast is of high certainty) and 0 otherwise.

We turn now to our baseline model, the logistic regression (logit). In this respect, suppose we have $N$ independent realizations $\{y_j\}_{j=1}^{N}$ of a random variable $Y_j$. Let $Y_j \sim \text{Bernoulli}(\lambda_j^c)$ and $y_j$ be equal to 1 if respondent $j$ is consistent and 0 otherwise. We can then model the probability $\lambda_j^c$ using a linear predictor function according to

$$\text{logit}(\lambda_j^c) = d_j\alpha + x_j'\beta + z_j\gamma, \tag{A.1}$$

where $d_j$ is a dummy for the treatment group, $x_j$ a vector of attributes, $z_j$ a measure of uncertainty, and $\alpha$, $\beta$, $\gamma$ a set of parameters. We then estimate the regression coefficients in Equation (A.1) through maximum likelihood estimation.

[22]Table A.10 in Section A.8 explores the correlations between firm attributes. There is little correlation present in the data.

[23]Appendix Section A.6 explores the robustness of our results to using parametric evaluations of subjective dispersion.

This specification makes use of our non-parametric assessment of consistency. The idea here is to predict the likelihood that a forecaster will produce a consistent forecast based on his or her question ordering, the characteristics of his or her employing firm, and the uncertainty surrounding his or her forecast. However, since the coefficients that Equation (A.1) yields are log odds and thereby difficult to interpret, we compute and report the marginal probability changes evaluated at means associated with a discrete change away from the reference category. Because we coded our binary regressors such that switching away from the reference category (i.e., from zero to one) should increase the probability of being consistent, our estimates will tell us by how much it does at the margin for an average respondent.[24]

As an alternative model, we use our parametric assessment of consistency and estimate the following linear regression (LR):

$$-|\Delta_j| = d_j\tilde{\alpha} + x_j'\tilde{\beta} + z_j\tilde{\gamma} + \tilde{\varepsilon}_j, \tag{A.2}$$

where $|\Delta_j|$ is the absolute difference between the point forecast and the midpoint given by respondent $j$, $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$ a new set of parameters, and $\tilde{\varepsilon}_j$ are i.i.d. errors.

Considering the distance in absolute terms and negating it makes our two specifications comparable, because it recovers our notion of consistency in levels. In particular, all else equal, each coefficient can be interpreted as the average marginal increase in closeness between subjective midpoints and point forecasts produced by switching away from the reference category of the dummy regressors. While Equation (A.1) provides estimates to be interpreted in terms of probabilities, Equation (A.2) yields estimates in terms of percentage points of inflation. Therefore, the linear regression model will serve us both as an assessment of the robustness of the results under the logit and as an indication of their economic significance.

Table A.3 displays the results from our two specifications based either on midpoint consistency (columns 1 and 2) or on median

[24]Recall that the logistic transform is a non-linear combination of the regressors, so that we need to fix their value to assess marginal probability changes. We take their sample respective means.

**Table A.3. Logit and LR of Midpoint and Median Consistency on Attributes**

| | Subjective Midpoint | | Subjective Median | |
|---|---|---|---|---|
| | **Logit (1)** | **LR (2)** | **Logit (3)** | **LR (4)** |
| Point Forecast First | 0.0635** | 0.0716** | 0.0285 | 0.102** |
| | (2.82) | (3.96) | (1.06) | (4.36) |
| High Certainty | 0.0723** | 0.0513 | 0.103*** | 0.0726** |
| | (3.21) | (1.84) | (5.54) | (3.24) |
| High Turnover | 0.0668*** | 0.0560*** | 0.0333 | 0.0439** |
| | (3.69) | (4.61) | (1.27) | (3.34) |
| Low Openness | 0.00959 | 0.00818 | 0.00964 | 0.0438* |
| | (0.40) | (0.37) | (0.55) | (2.62) |
| Services Sector | 0.0352* | 0.0416 | –0.0194 | –0.00636 |
| | (2.06) | (1.54) | (–1.00) | (–0.36) |
| Constant | | –0.502*** | | –0.597*** |
| | | (–19.95) | | (–22.63) |

**Note:** $t$-statistics in parentheses. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. $N = 1,217$. Logistic regression (logit) models use the proportion of consistent forecasts as the dependent variable, whereas linear regression (LR) models use the negative absolute difference between the point forecast and the central tendency measure derived from the density forecast. Logit coefficients represent marginal probability changes evaluated at means. All models include time fixed effects, and standard errors are clustered at the quarterly level.

consistency (columns 3 and 4). Note that to account for potential global time-varying factors, we include in all our models time fixed effects.[25,26] As mentioned above, the logit coefficients (odd columns) show the marginal increase in the probability of being consistent

[25]A caveat of our approach, however, is that we cannot control for individual fixed effects. This is not a problem insofar as forecasters' ability to be consistent through time is *not* correlated with our regressors. In other words, we need to make the assumption that this ability is unobservable by employers and that there is no self-selection of better forecasters into certain types of firms. Since this may be argued to be a somewhat strong assumption, one should interpret our estimates as upper bounds.

[26]All our results are robust to the non-inclusion of time fixed effects.

induced by a discrete change of the variable in row, when all the other variables take their mean value. LR coefficients (even columns) express the average percentage-point increase in closeness between the center of the density forecast and the point forecast associated with the same change. Standard errors are clustered at the quarterly level.[27]

Overall, the results confirm our previous findings that question ordering matters. Focusing on subjective midpoints first, column 1 indicates that an average forecaster (in terms of its other characteristics) is as much as 6.35 percent more likely to submit a consistent forecast if he or she is asked for a point forecast first. Column 2 tells us that such ordering makes the point forecast on average closer to the subjective midpoint by 7.16 basis points.[28]

In addition, Table A.3 suggests that consistency depends on some firm attributes and certainty as well. First, being more certain induces a marginal increase of 7.23 percent in the probability of being consistent. However, it does not seem to exert a significant effect on the closeness between subjective midpoints and point forecasts. Second, if the respondent works in a firm with a high turnover, the probability marginally increases by 6.68 percent, and reduces the distance between the midpoint and the point forecast by 5.6 basis points. Third, we cannot say that openness has an impact on consistency in either specification. Fourth, although consistency in levels does not seem to significantly vary with the sector in which the firm operates (column 2), it appears that switching to the service industry marginally increases the probability for the average forecaster to be consistent by 3.52 percentage points. Finally, the constant reflects part of our previous results, saying that a rather uncertain forecaster working in a small open construction or manufacturing firm hands in a point forecast on average 0.5 percentage point away from the midpoint if he or she sees the question about the density forecast first.

[27] In Section A.6, we construct pseudo-identifiers based on the combination of firms' turnover, openness, and sector, and we cluster standard errors at this level to account for heteroskedasticity. Our results are robust.

[28] A basis point is a hundredth of a percent of inflation.

We now inspect median consistency to assess the robustness of these results. The logistic regression in this context (column 3) globally suggest a similar picture but with somewhat less statistical significance. In fact, only certainty turns out to significantly raise the probability of consistency. Moreover, the service-sector dummy now exerts a negative marginal effect—although not significant—on such probability. Column 4 on the other hand reveals that the linear regression model performs better than its counterpart from column 2. We can indeed infer that all our dummy variables except for the sectoral one provokes a positive and significant effect on forecast consistency as measured by the closeness between subjective point forecasts and the median of the density forecasts.

Interestingly, the positive effect on consistency of question ordering and certainty is of higher magnitude than in column 2. This result together with the non-significance of the corresponding coefficient in column 3 indicates that question ordering makes little difference in the marginal probability that the misalignment between the median and the point forecast exceeds a relevant threshold but stills makes this misalignment on average greater by 10.2 basis points when the density forecast is asked first. In addition, the constant term reveals a greater discrepancy than in column 2. This reinforces our previous argument that subjective midpoints capture the information relevant to point forecasts better than medians.

### *A.4 Distribution Fitting*

As mentioned in Section 3, assuming that all the mass of the density forecast lies at the center of each bin tends to overstate the level of uncertainty if the underlying distribution is thought to be bell-shaped. Although we do not make explicit use of the second moment of the density forecasts, it is worth considering an alternative parametric approach to assess the robustness of our results.

Thus, following Giordani and Söderlind (2003), we can assume that each forecaster's density forecast is normally distributed, and we can solve for each individual parameter through numerical optimization. Formally, we would like to estimate for each respondent

$j \in \{1, \ldots, N\}$ the subjective mean $\mu_j$ and the subjective variance $\sigma_j^2$ according to

$$\min_{\hat{\mu}_j, \hat{\sigma}_j^2} \sum_{k=1}^{K} (P[L_k < Z_j \leq U_k] - p_{j,k})^2,$$

where $K$ is the number of bins, $Z_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$; $L_k$ and $U_k$ respectively denote the lower and upper bound of bin $k$; and $p_{j,k}$ denotes the probability associated by respondent $j$ to bin $k$. In other words, we pick the set of parameters that minimizes the sum of squared differences between the probability mass lying under the curve of a normal density following these parameters, and the probability mass assigned by the respondent.

When the number of bins used by the respondent does not exceed two, the fitting may not be satisfying. To address this issue in a simple manner we assume (i) a uniform distribution within the bin if only one bin is used, and (ii) that the mass lies at the center of the bin if exactly two bins are used. The first assumption avoids a subjective variance of 0. Note that this procedure slightly differs from the one used by Giordani and Söderlind (2003) since we have to address bins of different sizes. However, these two special cases occur in less than 20 percent of our experiment sample, and thus should not be of critical importance.

Overall, this specification is somewhat more restrictive than the one we use in the paper, as it assumes that the underlying distribution is symmetric and unimodal. Nevertheless, it is appealing in that it equates the mean, the mode, and the median. Appendix Section A.6 shows the results associated with this approach.

### *A.5 Diebold-Mariano Tests*

Suppose we have two competing forecasts $\{\hat{x}_{it}\}_{t=1}^{T}$ and $\{\hat{x}_{jt}\}_{t=1}^{T}$ of the same time series $\{x_t\}_{t=1}^{T}$, with respective resulting forecast errors $\{\hat{e}_{it}\}_{t=1}^{T}$ and $\{\hat{e}_{jt}\}_{t=1}^{T}$. Let $g(x_t, \hat{x}_{kt}) = g(\hat{e}_{kt})$ be an arbitrary loss function of the realization and the prediction $k = i, j$, or equivalently, the forecast error. We test the null hypothesis of equal predictive accuracy, i.e., whether the expected loss differential is zero, $E[d_t] \equiv E[g(e_{it}) - g(e_{jt})] = 0$.

Under stationarity and short memory of the loss differential series $\{d_t\}_{t=1}^{T}$, Diebold and Mariano (2002) propose the following test statistic:

$$S = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \overset{a}{\sim} N(0,1),$$

where $\bar{d}$ is the sample mean loss differential, and $\hat{f}_d(0)$ is a consistent estimate of the spectral density of the loss differential at frequency 0.[29,30]

Typical criteria for the loss function $g(\cdot)$ include mean squared error (MSE), mean average error (MAE), and mean average percentage error (MAPE). Following, we make use of the MSE, $g(\hat{e}_{kt}) = \hat{e}_{kt}^2$.[31]

Note that the DM test compares *one* prediction per competing forecast for each observed period, so we have to summarize our data along the panel dimension. To that end, the two most natural statistics are the mean and the median. This leaves us with four different forecasts (two types of questions, i.e., point forecast (PF) or density forecast (DF), and two groups, i.e., point forecast first (P) or density forecast first (D), and six unique pairwise comparisons.

### *A.6 Robustness*

#### *A.6.1 Non-Parametric Mode*

One could argue that the mode of the density forecast is reported as the point forecast rather than the midpoint or the median. To

[29]Such an estimate requires to choose the maximum order of the lag to consider when computing the long-run variance of the loss differential series from its autocovariance function. Diebold and Mariano (2002) suggest $k-1$, where $k$ is the forecast horizon ($k = 8$ quarters in our case). Alternatively, one can use the Schwert criterion, which is $12 * (T/100)^{1/4}$ and which yields something between 7 and 8. We select 7, consistent with both criteria. Furthermore, to ensure non-negativity of the estimate of the spectral density, we use a Bartlett kernel.

[30]As noted by Harvey, Leybourne, and Newbold (1997), with few time-series observations and long-horizon forecasts, test size distortions likely exist. Table A.8 shows the DM tests with bootstrapped standard errors.

[31]Our results are robust to these different criteria.

**Table A.4. Mode Consistency by Question Ordering**

| | Subjective Mode | | | | | |
|---|---|---|---|---|---|---|
| | Consistent | | Below | | Above | |
| Quarter | $\lambda_P^c$ | $\lambda_D^c$ | $\lambda_P^b$ | $\lambda_D^b$ | $\lambda_P^a$ | $\lambda_D^a$ |
| 2014:Q4 | 73.8 | 73.2 | 16.4 | 3.6 | 9.8 | 23.2 |
| 2015:Q1 | 74.6 | 75.4 | 18.6 | 12.3 | 6.8 | 12.3 |
| 2015:Q2 | 70.9 | 64.0 | 20.0 | 4.0 | 9.1 | 32.0 |
| 2015:Q3 | 85.7 | 73.1 | 10.2 | 7.7 | 4.1 | 19.2 |
| 2015:Q4 | 77.2 | 71.9 | 17.5 | 7.0 | 5.3 | 21.1 |
| 2016:Q1 | 80.7 | 74.5 | 14.0 | 13.7 | 5.3 | 11.8 |
| 2016:Q2 | 64.7 | 64.8 | 29.4 | 7.4 | 5.9 | 27.8 |
| 2016:Q3 | 82.0 | 59.6 | 10.0 | 13.5 | 8.0 | 26.9 |
| 2016:Q4 | 86.3 | 79.2 | 9.8 | 6.3 | 3.9 | 14.6 |
| 2017:Q1 | 78.4 | 65.5 | 13.7 | 9.1 | 7.8 | 25.5 |
| 2017:Q2 | 73.3 | 79.6 | 17.8 | 8.2 | 8.9 | 12.2 |
| 2017:Q3 | 80.4 | 74.0 | 11.8 | 2.0 | 7.8 | 24.0 |
| Pooled | 77.2 | 71.2 | 15.9 | 7.9 | 6.9 | 20.9 |

**Note:** See Table 1 for details.

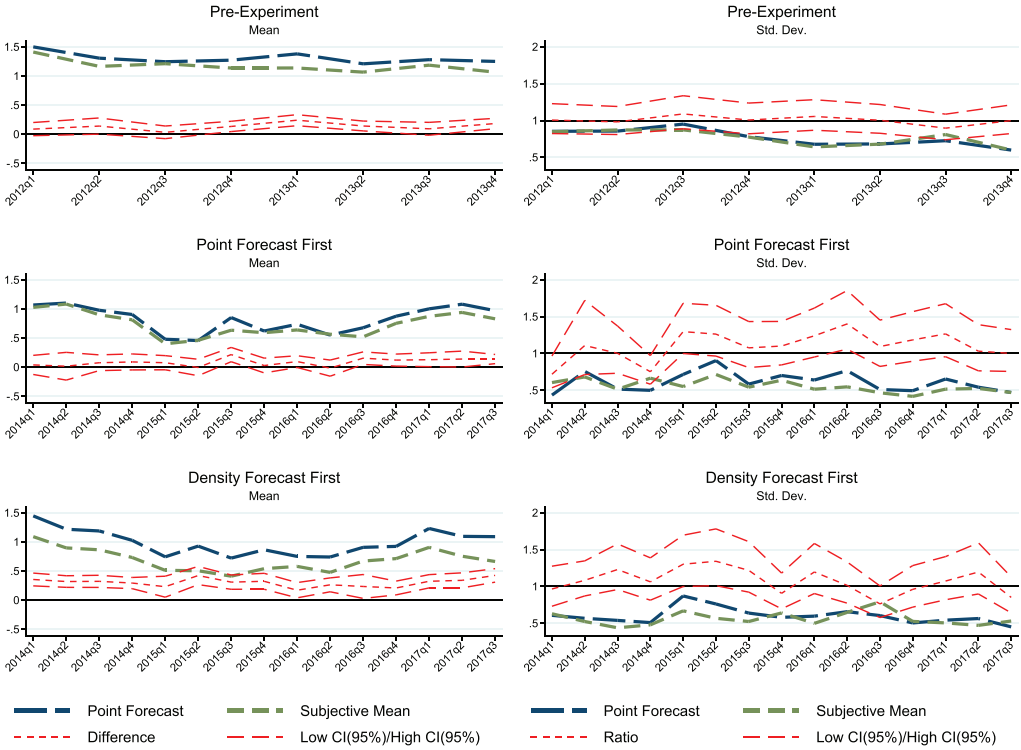
address the plausibility of such a hypothesis, we apply the same non-parametric exercise to this statistic.

The non-parametric subjective mode is taken as the bin itself to which the highest probability is assigned. When the highest probability is assigned to more than one bin, we take the bin that is closest to the midpoint. We do so because it prevents the need to address cases involving bins of different sizes, or cases of multi-modal density forecasts.

For each quarter and by question ordering, Table A.4 displays the proportion of respondents whose point forecast lies respectively within, below, or above its consistency level.

All the conclusions drawn from Table 1 are conserved. Asking for a point forecast first yields point forecasts that are more frequently mode consistent than asking for a density forecast first by a 6 percentage points average margin. Moreover, we observe a stronger

**Figure A.2. Treatment Effect Under an Alternative Parametric Approach**




**Note:** See Figure A.1 for details.

discrepancy between the two question orderings in regard to inconsistent forecasts. In particular, an inconsistent point forecast is much more likely to lie below its consistent level if the point forecast is asked first, but much more likely to lie above it if the density forecast is asked first.

### *A.6.2 Normally Fitted Parameters*

Figure A.2 shows the effect of question ordering on forecast inconsistencies when we use the subjective means stemming from the normal

**Table A.5. Forecast Inconsistencies and Treatment Effect**

| | Obs. | | Inconsistency | | Treatment Effect | |
|---|---|---|---|---|---|---|
| **Quarter** | $N_D$ | $N_P$ | $\Delta_D$ | $\Delta_P$ | $\Delta_D - \Delta_P$ | $p$-value |
| 2014:Q4 | 56 | 61 | 0.31 | 0.09 | 0.22 | 0.01 |
| 2015:Q1 | 57 | 59 | 0.24 | 0.09 | 0.15 | 0.09 |
| 2015:Q2 | 50 | 55 | 0.41 | 0.01 | 0.39 | 0.00 |
| 2015:Q3 | 52 | 49 | 0.31 | 0.22 | 0.09 | 0.15 |
| 2015:Q4 | 57 | 57 | 0.32 | 0.04 | 0.29 | 0.00 |
| 2016:Q1 | 51 | 57 | 0.17 | 0.10 | 0.08 | 0.18 |
| 2016:Q2 | 54 | 51 | 0.25 | –0.01 | 0.26 | 0.00 |
| 2016:Q3 | 52 | 50 | 0.24 | 0.16 | 0.08 | 0.25 |
| 2016:Q4 | 48 | 51 | 0.21 | 0.13 | 0.08 | 0.16 |
| 2017:Q1 | 55 | 51 | 0.33 | 0.12 | 0.20 | 0.01 |
| 2017:Q2 | 49 | 45 | 0.34 | 0.14 | 0.19 | 0.02 |
| 2017:Q3 | 50 | 51 | 0.42 | 0.14 | 0.28 | 0.00 |
| Pooled | 631 | 637 | 0.29 | 0.10 | 0.19 | 0.00 |

**Note:** See Table 3 for details.

density fitting approach described in Section A.4 of this appendix instead of the subjective midpoints.

Clearly, the picture yields the same general interpretation as to the effect of question ordering on forecast inconsistencies. For the experiment sample, although we observe a slightly higher degree of inconsistencies for the $P$ group compared to using midpoint (see Figure A.1), these quarterly average inconsistencies remain of rather low statistical significance. By contrast, the null hypothesis of equal means between point forecasts and subjective means can still be rejected for every single quarter regarding the $D$ group. The treatment effect under this specification remains qualitatively unchanged, as indicated by Table A.5. Indeed, with a significant average difference in inconsistencies of 0.19 percentage point of inflation, we reject the null hypothesis that this difference is zero at the quarterly level in 7 out of 12 cases.

Table A.6 displays the results from the linear regression estimated in Equation (A.2) when we use the subjective means from

**Table A.6. LR of Subjective Mean Consistency on Attributes**

| | Subjective Mean | |
|---|---|---|
| | **(1′)** | **(2′)** |
| Point Forecast First | 0.0870*** | 0.0862*** |
| | (4.68) | (4.93) |
| High Certainty | 0.0425 | 0.105** |
| | (1.53) | (3.26) |
| High Turnover | 0.0537** | 0.0559** |
| | (3.94) | (4.41) |
| Low Openness | 0.0230 | 0.0148 |
| | (1.19) | (0.71) |
| Services Sector | 0.0142 | 0.0107 |
| | (0.47) | (0.36) |
| Constant | –0.486*** | –0.515*** |
| | (–16.75) | (–17.75) |

**Note:** See Table A.3 for details.

the fitted normal distributions instead of the midpoints in the computation of the dependent variable.

The cells are therefore the coefficients from the linear regression of the absolute difference in negative terms between the point forecasts and the subjective means on question ordering, certainty, and firm characteristics. Column 1′ considers the exact same variable of certainty as in the paper (cf. Table A.3), while column 2′ considers an alternative dummy variable based on the subjective standard deviation derived from the normal fitting approach. Namely, its value is 1 if the subjective standard deviation is less than or equal to 0.6, and 0 otherwise. The value of 0.6 was chosen because it is the median subjective standard deviation in the full sample.

Comparing column 1′ here with its counterpart in Table A.3 (i.e., column 2) leads to the exact same conclusions. Furthermore, looking at column 2′ reinforces the view that more certainty (at the individual level) is associated with a higher degree of consistency. Using a parametric measure of certainty, namely, the subjective standard

deviation recoded into a dummy variable, makes the corresponding coefficient highly significant.

Overall, we argue that all our results are robust to adopting the alternative parametric approach described in Section A.4 of the appendix, which consists in fitting normal distributions to individual density forecasts. Using the subjective means instead of the midpoints yields the same general conclusions.

### *A.6.3 Accounting for (Some) Heteroskedasticity*

Arguably, the forecasts (and forecast errors) produced by a given CFO are likely autocorrelated. Thus, one may be worried that the standard errors calculated in the regression analysis (Section A.3 of this appendix) are unreliable due to heteroskedasticity. Ideally, one would like to cluster standard errors at the individual level. But because we do not observe the individual identifiers from our panel data, we cannot.

To circumvent the issue, we generate what we call pseudo-identifiers based on the combination between each firm sector, turnover, and share of revenues earned abroad. In practice, each combination is assigned a unique identifier that likely tracks firms (or groups of firms) along the time dimension. Though they do not uniquely identify firms in the data, they allow to improve on the interpretation of our data as repeated cross-samples. Moreover, firms have unlikely switched categories over the time of our experiment. By clustering standard errors at the pseudo-individual level, we therefore allow for heteroskedasticity at the firm (or group of very similar firms) level.

Table A.7 shows the results of the regression analysis when standards errors are clustered at the pseudo-individual level. Compared to Table A.3, results on question ordering remain robust. In fact, this approach only weakens the significance of some of the controls: Being a firm from the service sector is no longer significant in column 1, more certainty is no longer associated with higher levels of forecast consistency from column 2, and openness does not seem to matter anymore in column 4. Note that controlling for pseudo-individual fixed effects as well yields the same general conclusions.

**Table A.7. Logit and LR with Clustered Standard Errors**

| | Subjective Midpoint | | Subjective Median | |
|---|---|---|---|---|
| | **Logit (1)** | **LR (2)** | **Logit (3)** | **LR (4)** |
| Point Forecast First | 0.0635** | 0.0716** | 0.0285 | 0.102** |
| | (2.87) | (3.21) | (1.00) | (3.12) |
| High Certainty | 0.0723** | 0.0513* | 0.103*** | 0.0726** |
| | (2.95) | (2.03) | (4.16) | (3.07) |
| High Turnover | 0.0668** | 0.0560** | 0.0333 | 0.0439 |
| | (2.72) | (2.70) | (1.08) | (1.84) |
| Low Openness | 0.00959 | 0.00818 | 0.00964 | 0.0438 |
| | (0.38) | (0.35) | (0.28) | (1.74) |
| Services Sector | 0.0352 | 0.0416 | –0.0194 | –0.00636 |
| | (1.39) | (1.81) | (–0.58) | (–0.26) |
| Constant | | –0.502*** | | –0.597*** |
| | | (–13.57) | | (–13.75) |

**Note:** $t$ statistics in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. $N$ = 1,217. Logistic regression (logit) models user the proportion of consistent forecasts as the dependent variable, whereas linear regression (LR) models use the negative absolute difference between the point forecast and the central tendency measure derived from the density forecast. Logit coefficients represent marginal probability changes evaluated at means. All models include time fixed effects, and standard errors are clustered at the pseudo-individual level. See Appendix Section A.6 for details.

### *A.6.4 Diebold-Mariano Tests with Bootstrapped Standard Errors*

Table A.8 shows the DM tests with bootstrapped standard errors with 1,000 replications. Compared to Table 4, the significance is only slightly affected downwards.

## *A.7 Treatment Effect in Non-Parametric Approach*

In Table 1 we show, by group, the percentage of respondents whose point forecast is respectively consistent with, below and above the central tendency of the corresponding density forecast. For the sake of clarity, the table does not show the significance of the differences between groups. Doing so would amount to assessing the

**Table A.8. Diebold-Mariano Tests for Predictive Accuracy**

| **Median Forecast** | | | | |
|---|---|---|---|---|
| | $(PF,P)$ | $(PF,D)$ | $(DF,P)$ | $(DF,D)$ |
| $(PF,P)$ | | | | |
| $(PF,D)$ | –0.04431 | | | |
| $(DF,P)$ | 0.08875* | 0.1331 | | |
| $(DF,D)$ | 0.1282* | 0.1725* | 0.03945* | |
| **Mean Forecast** | | | | |
| | $(PF,P)$ | $(PF,D)$ | $(DF,P)$ | $(DF,D)$ |
| $(PF,P)$ | | | | |
| $(PF,D)$ | –0.05476 | | | |
| $(DF,P)$ | $0.03728^{\dagger}$ | $0.09205^{\dagger}$ | | |
| $(DF,D)$ | 0.08258 | $0.1195^{\dagger}$ | 0.02749 | |

**Note:** $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$. Entries show the column forecast MSE minus the row forecast MSE. $(i,j)$ denotes forecast $i = PF,DF$ made by group $j = P,D$. Positive values imply higher accuracy of the row forecast. Significance stars correspond to bootstrapped standard errors with 1,000 replications.

treatment effect of our experiment, which we document in a more sophisticated way through the parametric approach (namely, in Table 3).

Since it still may be of interest, Table A.9 displays the differences between the two groups in the proportions along with the $p$-value for the corresponding (not shown) $t$-statistics. The null hypothesis is that the proportion of respondents is the same regardless of the question ordering, and the test assumes equal variance.

As seen in Table A.9, the difference in midpoint consistency ($\Delta\lambda^c$) is statistically significant when one considers the pooled sample (last row). This observation is also true for the difference in mean inconsistencies both for the share of forecasts lying below and for the share of forecasts lying above their consistent level. In other words, respondents who are asked first for a density forecast tend to be less consistent (as opposed to the other group) and when they are inconsistent, they tend to overestimate inflation more and to underestimate it less.

## Table A.9. Group Differences in Forecast Consistency

| | Subjective Midpoint | | | | | | Subjective Median | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Consistent | | Below | | Above | | Consistent | | Below | | Above | |
| Quarter | $\Delta\lambda^c$ | $p$ | $\Delta\lambda^b$ | $p$ | $\Delta\lambda^a$ | $p$ | $\Delta\lambda^c$ | $p$ | $\Delta\lambda^b$ | $p$ | $\Delta\lambda^a$ | $p$ |
| 2014:Q4 | –1.8 | 0.80 | 8.1 | 0.07 | –6.2 | 0.32 | 0.7 | 0.93 | 13.0 | 0.01 | –13.7 | 0.06 |
| 2015:Q1 | 14.7 | 0.07 | –2.1 | 0.71 | –12.6 | 0.07 | 0.9 | 0.91 | 4.7 | 0.48 | –5.6 | 0.35 |
| 2015:Q2 | 12.5 | 0.17 | 6.5 | 0.30 | –19.1 | 0.01 | 7.3 | 0.44 | 17.8 | 0.01 | –25.1 | 0.00 |
| 2015:Q3 | 10.6 | 0.20 | 0.2 | 0.95 | –10.8 | 0.16 | 6.5 | 0.45 | 2.4 | 0.64 | –8.9 | 0.24 |
| 2015:Q4 | 5.3 | 0.52 | 15.8 | 0.00 | –21.1 | 0.00 | 5.3 | 0.51 | 14.0 | 0.01 | –19.3 | 0.00 |
| 2016:Q1 | 7.3 | 0.30 | 2.9 | 0.57 | –10.2 | 0.06 | –1.2 | 0.88 | 4.4 | 0.45 | –3.2 | 0.61 |
| 2016:Q2 | –5.0 | 0.53 | 13.9 | 0.02 | –8.9 | 0.14 | –11.9 | 0.22 | 22.0 | 0.00 | –10.1 | 0.22 |
| 2016:Q3 | 14.8 | 0.07 | 0.2 | 0.96 | –15.1 | 0.04 | 24.4 | 0.01 | –5.5 | 0.38 | –18.9 | 0.01 |
| 2016:Q4 | –0.9 | 0.91 | 3.7 | 0.45 | –2.8 | 0.68 | 1.2 | 0.88 | 1.7 | 0.70 | –2.9 | 0.69 |
| 2017:Q1 | 4.0 | 0.62 | 5.9 | 0.07 | –9.9 | 0.20 | 3.7 | 0.66 | 8.0 | 0.08 | –11.7 | 0.13 |
| 2017:Q2 | –4.0 | 0.64 | 6.8 | 0.14 | –2.8 | 0.72 | –6.6 | 0.48 | 11.7 | 0.08 | –5.0 | 0.51 |
| 2017:Q3 | 18.1 | 0.01 | 2.0 | 0.32 | –20.0 | 0.00 | 10.4 | 0.23 | 7.8 | 0.10 | –18.2 | 0.02 |
| Pooled | 6.4 | 0.01 | 5.5 | 0.00 | –11.8 | 0.00 | 3.4 | 0.17 | 8.6 | 0.00 | –12.0 | 0.00 |

**Note:** $\Delta\lambda^k$ is the difference $(\lambda_P^k - \lambda_D^k)$ between the two groups in the proportions shown in Table 1 for each category $k = c, b, a$, that is, whether the point forecast lies within, below, or above its level of consistency. $p$ denotes the $p$-value for the $t$-test that this difference is zero, assuming equal variance.

**Table A.10. Correlation between Attributes**

| | High Turnover | High Openness | Service Sector | High Certainty |
|---|---|---|---|---|
| High Turnover | 1 | | | |
| High Openness | 0.04 | 1 | | |
| Service Sector | –0.04 | –0.30*** | 1 | |
| High Certainty | 0.02 | 0.00 | –0.00 | 1 |

**Note:** ***$p < 0.001$. Each entry is the pairwise Pearson correlation between the row and the column variable.

In regard to median consistency, it appears that the difference in the pooled sample ($\Delta\lambda^c$) is not significant, although the differences in the distribution of inconsistencies ($\Delta\lambda^b$ and $\Delta\lambda^a$) are. This generally confirms our previous finding that when a point forecast is inconsistent, it is more likely to be above (below) its consistency level if the density forecast was asked first (second).

As for quarterly differences, they are hardly significant. This can be generally explained by the small sample size. Nevertheless, all the figures, when they are significant, offer a consistent view and lead to the same conclusions as above.

### *A.8 Correlation Between Attributes*

Table A.10 displays correlations between firm attributes displayed in Table A.2 together with the significance thereof. As shown in the table, there is little correlation present in the data. In fact, only firm sector and openness correlate significantly—unsurprisingly, as service firms tend to be more domestically oriented.

### *A.9 Effect of Question Ordering on One-Year-Ahead Exchange Rate Forecasts*

One may argue that the tests ran in this paper fail to provide a sensitive benchmark about the observed level of forecast inconsistencies. In order to put the inflation-related inconsistencies into perspective, we run a placebo test. In the survey, respondents are asked to provide a one-year-ahead point forecast for the exchange rate of the

## Table A.11. Exchange Rate Forecasts and Question Ordering

| Quarter | Obs. | | Mean Forecast | | Treatment Effect | |
|---|---|---|---|---|---|---|
| | $N_D$ | $N_P$ | $\tau_D$ | $\tau_P$ | $\tau_D - \tau_P$ | $p$-value |
| *EUR/CHF* | | | | | | |
| 2014:Q4 | 56 | 60 | 1.206 | 1.207 | –0.001 | 0.781 |
| 2015:Q1 | 54 | 59 | 1.071 | 1.065 | 0.006 | 0.493 |
| 2015:Q2 | 50 | 54 | 1.045 | 1.044 | 0.001 | 0.899 |
| 2015:Q3 | 52 | 49 | 1.073 | 1.088 | –0.014 | 0.038 |
| 2015:Q4 | 56 | 56 | 1.085 | 1.084 | 0.002 | 0.753 |
| 2016:Q1 | 51 | 56 | 1.090 | 1.091 | –0.001 | 0.903 |
| 2016:Q2 | 52 | 50 | 1.107 | 1.096 | 0.012 | 0.117 |
| 2016:Q3 | 51 | 50 | 1.096 | 1.094 | 0.002 | 0.655 |
| 2016:Q4 | 48 | 51 | 1.070 | 1.082 | –0.011 | 0.118 |
| 2017:Q1 | 54 | 51 | 1.093 | 1.079 | 0.014 | 0.221 |
| 2017:Q2 | 48 | 42 | 1.094 | 1.108 | –0.014 | 0.252 |
| 2017:Q3 | 49 | 51 | 1.131 | 1.140 | –0.009 | 0.164 |
| Pooled | 621 | 629 | 1.098 | 1.099 | –0.001 | 0.753 |
| *USD/CHF* | | | | | | |
| 2014:Q4 | 53 | 59 | 0.984 | 0.985 | –0.001 | 0.965 |
| 2014:Q1 | 53 | 58 | 0.992 | 0.984 | 0.008 | 0.437 |
| 2015:Q2 | 49 | 54 | 0.970 | 0.969 | 0.002 | 0.841 |
| 2015:Q3 | 52 | 49 | 0.988 | 0.984 | 0.005 | 0.525 |
| 2015:Q4 | 54 | 54 | 1.019 | 1.008 | 0.011 | 0.176 |
| 2016:Q1 | 51 | 53 | 1.002 | 1.006 | –0.004 | 0.553 |
| 2016:Q2 | 52 | 49 | 1.004 | 0.993 | 0.010 | 0.251 |
| 2016:Q3 | 50 | 49 | 0.994 | 0.995 | –0.001 | 0.900 |
| 2016:Q4 | 46 | 51 | 1.009 | 1.009 | –0.000 | 0.990 |
| 2017:Q1 | 52 | 51 | 1.016 | 1.007 | 0.009 | 0.499 |
| 2017:Q2 | 47 | 40 | 0.987 | 0.989 | –0.003 | 0.853 |
| 2017:Q3 | 47 | 50 | 0.983 | 0.980 | 0.003 | 0.699 |
| Pooled | 606 | 617 | 0.996 | 0.992 | 0.004 | 0.209 |

**Note:** The table displays, for each quarter of the experiment, the number of respondents $N_i$ and the average exchange rate forecast $\tau$ for each group $i = P, D$ as well as the difference thereof. The last column displays the $p$-value of the $t$-test that this difference $\tau_D - \tau_P$ is different than zero (assuming equal variances). The last row considers the pooled sample.

Swiss franc vis-à-vis the euro and the U.S. dollar. For this placebo test, exchange rate forecasts are good candidates because they are specific and quantitative, and are being asked shortly before the inflation forecasts. Significant differences between the two groups $P$ and $D$ would cast doubt on our treatment, as they would question the direct link between observed differences in the inflation forecasts and question ordering itself.

In this respect, Table A.11 displays, for each quarter of the experiment, the number of respondents $N_i$ and the average exchange rate forecast $\tau$ for each group $i = P, D$ as well as the difference thereof, for both currency pairs. The last column displays the $p$-value of the $t$-test that this difference $\tau_D - \tau_P$ is different than zero (assuming equal variances). The last row of each panel considers the pooled sample.

There does not seem to be a pattern in the difference in forecasts between the two groups. Out of the 24 quarterly tests run in Table A.11, we reject the null hypothesis of equal mean forecast only once at the 10 percent level (in 2015:Q3 for the EUR/CHF), which falls well within the type-I error rate. This provides additional evidence that discrepancies in inflation-related forecasts between groups are explained by question ordering rather than fortuitous, unobservable differences.

## References

Ang, A., G. Bekaert, and M. Wei. 2007. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54 (4): 1163–1212.

Arioli, R., C. Bates, H. C. Dieden, I. Duca, R. Friz, C. Gayer, G. Kenny, A. Meyler, and I. Pavlova. 2017. "EU Consumers' Quantitative Inflation Perceptions and Expectations: An Evaluation." ECB Occasional Paper No. 186.

Bernanke, B., and M. Woodford. 1997. "Inflation Forecasts and Monetary Policy." *Journal of Money, Credit and Banking* 29 (4): 653–84.

Boero, G., J. Smith, and K. F. Wallis. 2008. "Uncertainty and Disagreement in Economic Prediction: The Bank of England Survey of External Forecasters." *Economic Journal* 118 (530): 1107–27.

Bruine de Bruin, W., W. Van der Klaauw, J. S. Downs, B. Fischhoff, G. Topa, and O. Armantier. 2010. "Expectations of Inflation: The Role of Demographic Variables, Expectation Formation, and Financial Literacy." *Journal of Consumer Affairs* 44 (2): 381–402.

Bruine de Bruin, W., W. Van der Klaauw, G. Topa, J. S. Downs, B. Fischhoff, and O. Armantier. 2012. "The Effect of Question Wording on Consumers' Reported Inflation Expectations." *Journal of Economic Psychology* 33 (4): 749–57.

Carroll, C. D. 2017. "Heterogeneity, Macroeconomics, and Reality." Sloan-BoE-OFR Conference on Heterogeneous Agent Macroeconomics, U.S. Department of the Treasury.

Clements, M. P. 2009. "Internal Consistency of Survey Respondents' Forecasts: Evidence Based on the Survey of Professional Forecasters." In *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, ed. J. Castle and N. Shephard, 206–26. Oxford University Press.

———. 2010. "Explanations of the Inconsistencies in Survey Respondents' Forecasts." *European Economic Review* 54 (4): 536–49.

Coibion, O., Y. Gorodnichenko, and S. Kumar. 2018. "How Do Firms Form Their Expectations? New Survey Evidence." *American Economic Review* 108 (9): 2671–2713.

Coibion, O., Y. Gorodnichenko, S. Kumar, and M. Pedemonte. 2020. "Inflation Expectations as a Policy Tool?" *Journal of International Economics* 124 (May): Article 103297.

Coibion, O., Y. Gorodnichenko, S. Kumar, and J. Ryngaert. 2018. "Do You Know That I Know That You Know . . .? Higher-Order Beliefs in Survey Data." NBER Working Paper No. 24987.

Diebold, F. X., and R. S. Mariano. 2002. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 20 (1): 134–44.

Engelberg, J., C. F. Manski, and J. Williams. 2009. "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters." *Journal of Business and Economic Statistics* 27 (1): 30–41.

Garcia, J. A., and A. Manzanares. 2007. "Reporting Biases and Survey Results: Evidence from European Professional Forecasters." ECB Working Paper No. 836.

Gennaioli, N., and A. Shleifer. 2018. *A Crisis of Beliefs. Investor Psychology and Financial Fragility.* Princeton University Press.

Gideon, M., B. Helppie-McFall, and J. W. Hsu. 2017. "Heaping at Round Numbers on Financial Questions: The Role of Satisficing." *Survey Research Methods* 11 (2): 189–214.

Giordani, P., and P. Söderlind. 2003. "Inflation Forecast Uncertainty." *European Economic Review* 47 (6): 1037–59.

Grishchenko, O., S. Mouabbi, and J.-P. Renne. 2019. "Measuring Inflation Anchoring and Uncertainty: A US and Euro Area Comparison." *Journal of Money, Credit and Banking* 51 (5): 1053–96.

Harvey, D., S. Leybourne, and P. Newbold. 1997. "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting* 13 (2): 281–91.

Kim, G., and C. Binder. 2020. "Learning-Through-Survey in Inflation Expectations." Available at https://ssrn.com/abstract=3790834.

Lucas, R. E. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4 (2): 103–24.

Manski, C. F. 2018. "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise." In *NBER Macroeconomics Annual 2017*, Vol. 32, ed. M. Eichenbaum and J. A. Parker, 411–71 (chapter 5). University of Chicago Press.

Manski, C. F., and F. Molinari. 2010. "Rounding Probabilistic Expectations in Surveys." *Journal of Business and Economic Statistics* 28 (2): 219–31.

Meyler, A., and I. Rubene. 2009. "Results of a Special Questionnaire for Participants in the ECB Survey of Professional Forecasters (SPF)." MPRA Paper No. 20751.

Niu, X., and N. Harvey. 2021. "Context Effects in Inflation Surveys: The Influence of Additional Information and Prior Questions." *International Journal of Forecasting* 38 (3): 988–1004.

Schuman, H., and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context in Attitude Surveys.* New York: Academic.

Stark, T. 2013. "SPF Panelists' Forecasting Methods: A Note on the Aggregate Results of a November 2009 Special Survey." Federal Reserve Bank of Philadelphia.

Sudman, S., N. M. Bradburn, and N. Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass.

Svensson, L. E. 1997. "Inflation Forecast Targeting: Implementing and Monitoring Inflation Targets." *European Economic Review* 41 (6): 1111–46.

Tay, A. S., and K. F. Wallis. 2000. "Density Forecasting: A Survey." *Journal of Forecasting* 19 (4): 235–54.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Van der Klaauw, W., W. Bruine de Bruin, G. Topa, S. Potter, and M. F. Bryan. 2008. "Rethinking the Measurement of Household Inflation Expectations: Preliminary Findings." Staff Report No. 359, Federal Reserve Bank of New York.

Zarnowitz, V., and L. A. Lambros. 1987. "Consensus and Uncertainty in Economic Prediction." *Journal of Political Economy* 95 (3): 591–621.