HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

Measuring Inflation Expectations: How the Response Scale Shapes Density Forecasts

Christoph Becker

Peter Duersch

Thomas Eife

AWI DISCUSSION PAPER SERIES NO. 723
January 2023

Measuring Inflation Expectations: How the Response Scale Shapes Density Forecasts

Christoph Becker* Peter Duersch† Thomas Eife‡

January 13, 2023

Abstract

In density forecasts, respondents are asked to assign probabilities to pre-specified ranges of inflation. We show in two large-scale experiments that responses vary when we modify the response scale. Asking an identical question with modified response scales induces different answers: Shifting, compressing or expanding the scale leads to shifted, compressed and expanded forecasts. Mean forecast, uncertainty, and disagreement can change by several percentage points. We discuss implications for survey design and how central banks can adjust the response scales during times of high inflation.

JEL codes: C83, D84, E31

Keywords: Inflation, density forecast, experiment

Acknowledgements

We thank Kenza Benhima, Mohammed Bulutay, Christian Conrad, Timo Dimitriadis, Zeno Enders, Frank Heinemann, Luba Petersen, and Michael Weber for their helpful comments. We are especially grateful to Norbert Schwarz for his advice and for answering our many questions. We also would like to thank participants of the following conferences and seminars: LAGV in Marseille, HKMetrics in Mannheim, EF in Bonn, WTEM and DIW in Berlin. Financial support from the University of Heidelberg is gratefully acknowledged. Finally, we would like to thank the Bundesbank for including our questions in its survey on consumer expectations.

*Heidelberg University, christoph.becker@awi.uni-heidelberg.de.

†University of Mannheim, duersch@xeeron.de.

‡Heidelberg University, thomas.eife@awi.uni-heidelberg.de.

1 Introduction

The idea that managing expectations is an essential part of central banking emerged already before the financial crisis of 2008/2009 (e.g., Woodford, 2005; King *et al.*, 2008), but it was only when central banks reduced interest rates to levels around zero during the crisis that this "nonconventional" tool became widely adopted. With the interest rate, one of their main policy tools, incapacitated, central banks tried to "guide" households' and firms' expectations about future interest rates and future inflation (forward guidance). An important part of this management of expectations is their measurement which is typically done via large representative surveys. Following the lead of the Survey of Consumer Expectations (SCE), which was established by the Federal Reserve Bank of New York in 2013, numerous other central banks recently initiated surveys that include density (probabilistic) forecasts.[1] In density forecasts, respondents are given a response scale showing pre-specified intervals and are then asked to assign probabilities to the intervals that best represent their beliefs about the expected outcome of a variable (e.g., inflation rate or GDP growth).

The experimental results we present in this paper show that inflation expectations elicited via density surveys are highly dependent on the specifics of the response scale. Shifting or compressing the response scale used in the question causes respondents to shift or compress their answers. For example, we can vary respondents' mean inflation forecast from −0.32% to 8.21% simply by changing the center of the response scale. While these examples are extreme, it is clear that density forecasts cannot provide information about how well respondents' inflation expectations are "anchored" around a certain value (e.g., around the central bank's target). Often, we are more interested in changes over time rather than the levels of the forecasts but these, too, require a careful interpretation. Similarly, we can double respondents' average uncertainty (the standard deviation of their response) from 3.08% to 6.08% when we expand the scale by doubling its width. A somewhat striking finding is

[1]Examples are the Consumer Expectations Survey conducted by the European Central Bank (which surveys representative samples of households in all countries of the Euro area) and similar surveys conducted by the central banks of Canada, France, Germany, the Netherlands, Ukraine, and the United Kingdom.

that, on average, the number of intervals respondents use for their answer is roughly stable around five, independently of whether we compress or expand the scale. These findings make it unclear how well we can measure respondents' "true" uncertainty using density surveys.

Survey researchers have long been aware that even minor variations in the wording of a question or in the design of a questionnaire may strongly affect the responses (see Schwarz (2010) for an overview and Payne (1951) and Sudman & Bradburn (1974) for early contributions). The recent literature on inflation expectations also addresses these points. Phillot & Rosenblatt-Wisch (2018) discuss the effect of question ordering on respondents' forecast consistency. The effect of a question's wording on responses is discussed in Bruine de Bruin *et al.* (2012), Manski (2018), Coibion *et al.* (2020) and Coibion *et al.* (2022). Our focus here is on variations of the response scale. In a widely influential study, Schwarz *et al.* (1985) show that shifting the response scale in an interval question (where respondents are asked to pick a single interval) may lead to shifts of the responses. This phenomenon is a robust finding in survey research and has been replicated in various other studies (Schwarz, 2010, gives an overview). We extend this research agenda to density questions. In particular, we use the NY Fed's SCE question on inflation expectations as our baseline and employ a battery of 12 treatments to systematically test if, and if so, how changes to the scale affect the results. Four treatments study the effect of shifting the response scale and four treatments study the effect of compressing or expanding the scale. Given the specific characteristic of the SCE that the center intervals are narrower than the other closed intervals, the final four treatments study the effect of combining or splitting up existing intervals. We collect data on two different subject pools: A representative sample of 1,300 respondents from the United States on which we ran all 13 treatments, and a representative sample of more than 4,000 respondents from Germany on which the Bundesbank ran three of our treatments.

The rest of the paper is organized as follows. Section 2 describes the experimental design and provides details on the hypotheses and the data collection for the US sample. Section 3 presents the results for the US sample and Section 4 presents the results for the Bundesbank panel. Section 5 interprets the results and addresses policy implications including a discussion about how the response scale could be modified when higher actual inflation rates compel central banks to adjust the scale.

2 Experimental design

Our US experiment consists of an online survey on 12-months ahead inflation. For inflation, we use questions from the New York Fed Survey of Consumer Expectations (SCE). Specifically, participants answer a density forecast (designated Q1 in our survey, Q9 in the SCE), a binary question whether participants expect inflation or deflation (Q2, Q8v2 in the SCE) and a point forecast of the inflation rate (Q3, Q8v2part2 in the SCE). The density forecast (Q1) covers the inflation rates from smaller than -12% (deflation) to larger than 12% (see e.g. Figure 2.1). The two outermost intervals are open, capturing in principle all values from -∞% to -12% and 12% to ∞%. In contrast to original ordering of the SCE, we move the density forecast ahead of the other two questions. This is to prevent the other two questions (especially the point forecast) from confounding answers in the density forecast, which we are primarily interested in. Apart from the ordering, we adopt the wording used by the Fed for the questions and answers. For the point forecast (Q3), the wording depends on the answer to Q2: If a respondent stated in Q2 that they expect an inflation in the next 12 months, the wording in Q3 asked them for a point forecast for the inflation rate. If they answered deflation in Q2, they were likewise asked for the expected deflation rate in Q3. This aspect was also taken from the original Fed survey.

After each of these three questions, participants are asked to indicate how certain they feel about their answer on a 6-item Likert scale (ranging from Very Uncertain to Very Certain). This allows us to later control for the participants' degree of confidence in their answers. Following these six main questions, participants are asked to answer a short questionnaire about age, gender, education, knowledge of the Fed's inflation target, political orientation, state of residence and three questions on financial literacy (Lusardi & Mitchell, 2014). The questionnaire also includes a control question to test the attentiveness of participants.

In the survey, participants face one of 13 different treatment conditions. In *Baseline*, the intervals of the scale given in the density forecast are identical to the SCE. The other 12 treatment conditions introduce different variations to the response scale. All other questions are the same across all treatments. Hence, our design allows to isolate the effect of changing the scale of the density forecast on the forecast itself, but also on subsequent assessments of 12-months ahead inflation via other question types. The 12 treatment conditions are grouped into three categories: *Shift* treatments, *Compression* treatments and *Centralization* treatments. The following three subsections present the different categories in greater detail.

2.1 Shift treatments

In the *Shift* treatments, the response scale is shifted towards either inflation or deflation, keeping all other parameters (e.g. number of intervals, their relative widths) constant. This means that the center of the scale moves away from zero compared to the *Baseline*. Thus, the *Shift* treatments allow us to test how respondents' forecasts are influenced by different centers of the scale. We implement both shifts in two different degrees, resulting in a total of four *Shift* treatments: *ShiftMinus12*, *ShiftMinus4*, *ShiftPlus4*, and *ShiftPlus12*. Figure 2.1 shows an overview of the four *Shift* treatments, with the *Baseline* as a reference. In *ShiftMinus12* and *ShiftMinus4* we subtract 12 and 4 respectively from the interval limits. Conversely, in *ShiftPlus4* and *ShiftPlus12*, we add respectively 4 and 12 to the interval limits. This means that the response scales in *ShiftMinus12* and *ShiftPlus12* only have a single open interval to indicate either belief in inflation (in *ShiftMinus12*) or deflation (in *ShiftPlus12*). A participant in *ShiftMinus12*, for example, who believes that prices will increase in the next 12 months (implying a positive inflation rate), would need to assign a probability of 100% to the one remaining open interval that captures this belief.

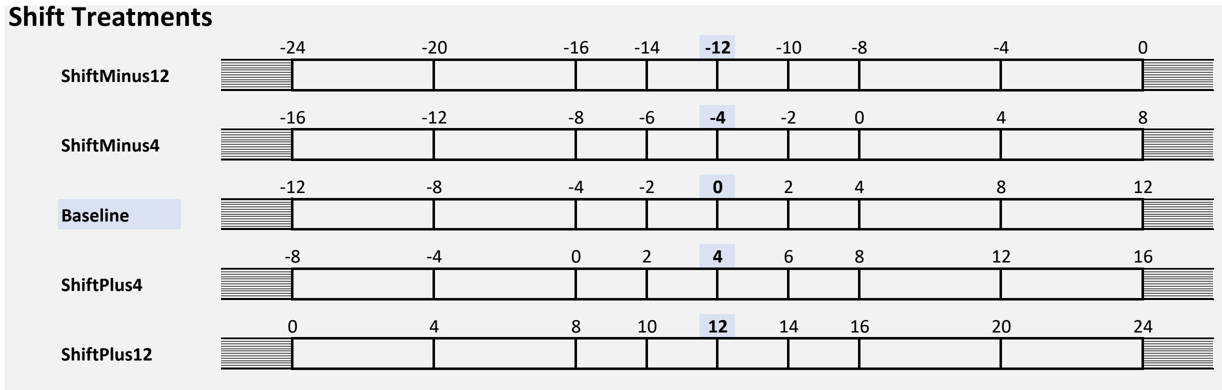

Figure 1: **Shift Treatments**: The figures shows the response scales of the *Shift* treatments with *Baseline* as reference. In the Shift treatments, the response scales shifts along the real line leaving all else (number of intervals, width of intervals) unchanged.

2.2 Compression treatments

In the four *Compression* treatments, the interval limits of the response scale are multiplied by a fixed factor, keeping the number and the relative size of the intervals unchanged. For factors below 1, this implies that the scale is being compressed around the midpoint. Factors above 1 result in an expansion. The *Compression* treatments help us to understand how changing the width of the response scale affects the responses. As before, we implement both compression and decompression with two different degrees giving us four *Compression* treatments: *Compression0.25*, *Compression0.5*, *Compression2*, and *Compression4*. An overview of these treatments is provided in Figure 2.2. In *Compression0.25* and *Compression0.5* the interval limits are multiplied by 0.25 and 0.5 and thus provide scales that zoom in more closely to inflation rates close to zero. In comparison, *Compression2* and *Compression4* widen the intervals of the response scale. As Figure 2.2 demonstrates, this results in values now being explicitly included in intervals that would have been part of the open intervals in *Baseline*. While *Compression2* and *Compression4* thus allow participants to better communicate beliefs about high inflation (deflation) rates, they also imply a coarser image of participants' inflation beliefs around the midpoint.

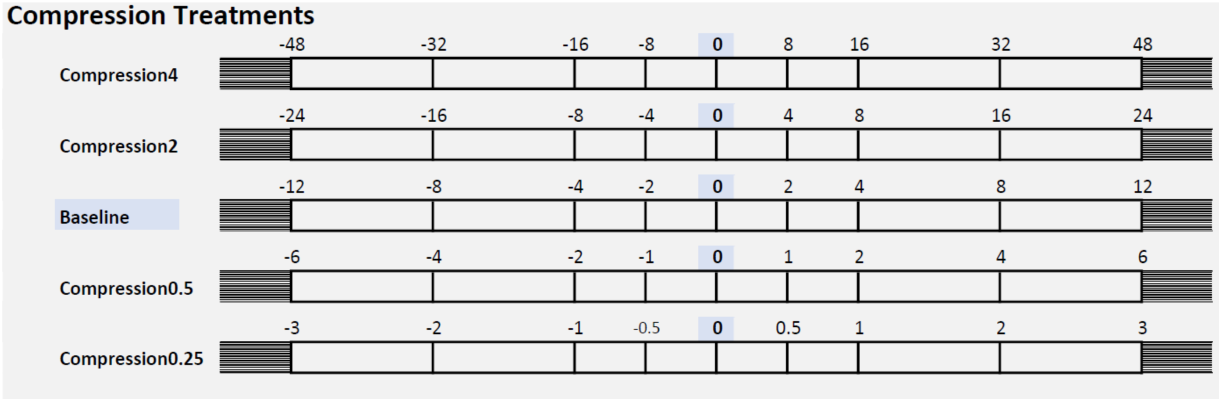

Figure 2: **Compression Treatments**: The figure shows the response scales of the *Compression* treatments with *Baseline* as reference. In the *Compression* treatments, the interval limits of the response scale are multiplied by a constant factor leaving all else (number of intervals and center of the scale) unchanged.

2.3 Centralization treatments

Finally, the four *Centralization* treatments vary the number of bins around the center of the scale. Differently to the other two treatment categories, where the scale is either shifted or compressed, the overall range of the scale is identical to *Baseline*. Instead, we either split existing intervals around the center or we combine them, thus changing the overall number of intervals. Similarly to the *Compression* treatments, this allows for a finer or coarser image of participants' inflation beliefs around the center, however, without changing the range of the scale itself. As with the other treatment categories, the centralizations are implemented with two different degrees, giving us the final set of four treatments: *Centralization6*, *Centralization8*, *Centralization12*, and *Centralization14*. Figure 2.3 depicts all four treatments relative to *Baseline*. In *Centralization6* and *Centralization8* the center intervals are combined, such that the overall number of intervals decreases to 6 or 8, respectively. Participants in these treatments thus can only give very coarse beliefs. In *Centralization12* and *Centralization14*, on the other hand, we split the intervals around the center allowing participants to more finely express beliefs in this range.

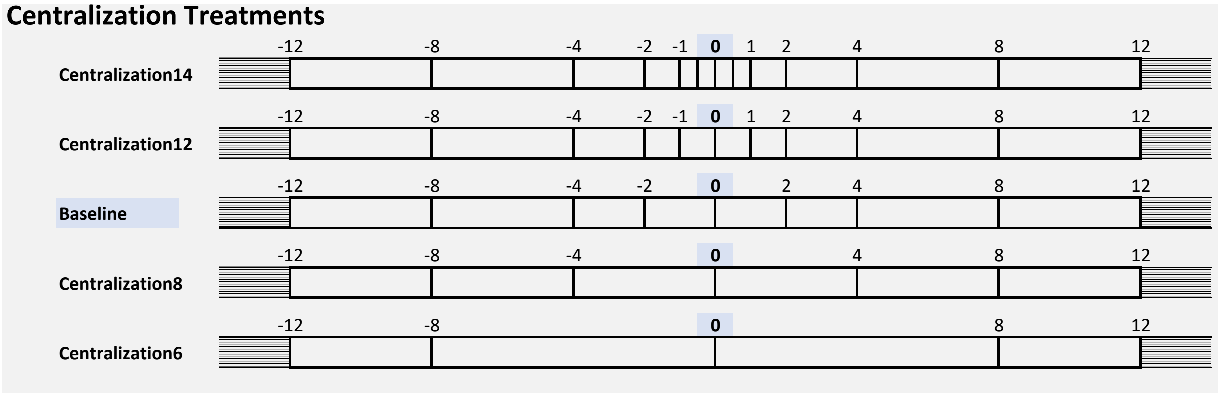

Figure 3: **Centralization Treatments**: The figure shows the response scales of the *Centralization* treatments with *Baseline* as reference. In the *Centralization* treatments, the center intervals are split apart or merged with adjacent intervals—decreasing or increasing the overall number of intervals. Everything else (center and width of the scale) is unchanged. Centralization8 has all closed intervals of the same size.

2.4 Hypotheses

The hypotheses were preregistered with the design of the experiment on the AEA RCT registry (`www.socialscienceregistry.org/trials/8716`). The study received ethics approval from the German Association for Experimental Economic Research (`gfew.de/ethik/apyKIJdX`).

2.4.1 Across treatment hypotheses

First, we test how our treatment scales affect key moments, such as the mean or dispersion of respondents' answers, across treatments. If participants use the scale provided by the experimenters as a guideline that informs their answers, we would expect their answers to follow changes to the scale relative to the *Baseline*. That is, for the *Shift* treatments, the middle point of the scale is shifted away from zero. Thus, we would expect participants to also shift their stated distribution in the same direction.

Hypothesis 1: *In the Shifting treatments, the reported distributions of inflation expectations shift in the direction of the scale shift.*

After we elicit the density forecast (Q1), we ask participants in a binary question whether they expect inflation or deflation in the next 12 months (Q2), and elicit a point forecast (Q3). If our treatments interventions affect the inflation beliefs beyond the density forecast in our survey, we expect the answers to Q2 and Q3 also to differ from those obtained in *Baseline*. The *Shift* treatments here provide us with two intuitive predictions to test. As the midpoint of the scale is shifted in the *Shift* treatments, this might influence participants to move their expectations further towards inflation or deflation. If participants answer the binary question consistent with their stated distribution, we expect different rates of participants expecting inflation or deflation in the *Shift* treatments compared to *Baseline*. Furthermore, we also expect such an effect to carry over to the point forecast.[2] Hence, we test:

Hypothesis 2: *In the Shifting treatments, the incidence of expecting deflation is lower [higher] for positive [negative] shifts of the scale. The incidence of expecting inflation is higher [lower] for positive [negative] shifts if the scale.*

Hypothesis 3: *In the Shifting treatments, the point forecast is higher [lower] for positive [negative] shifts of the scale.*

For the *Compression* and *Centralization* treatments, the midpoint is held constant at zero. Here, we would expect a distribution that is still centered on the same value as if the participant were part of *Baseline*. However, both treatments vary the range and/or number of bins on the scale. In the *Compression* treatments, as the scale is either stretched out or squeezed together, a participant trying to assign a positive probability to each bin would end up with a more dispersed distribution in the first case and a less dispersed distribution in the second case. Conversely, the *Centralization* treatments split or combine bins around the center of the scale. A distribution with the same dispersion as in the *Baseline* could be obtained if the probability mass is simply split proportional to the split of the bin (or combined in

[2]For the *Compression* and *Centralization* treatments, matters look different: Assuming that the point forecast corresponds to the central moment of the density forecasts in some way, point forecasts should not be affected by our treatment interventions, even if participants behave in line with Hypotheses 4 and 5. However, this would strongly depend on the actual, individual distribution of beliefs, whether the point forecast is actually linked to the density forecast, and if so, which central moment (mean, mode or median) is actually used to inform the point forecast. Hence, we will test for differences across treatments for the *Compression* and *Centralization* treatments in an exploratory fashion, without having a specific hypothesis concerning possible effects.

proportion to the combination of bins). As previous literature has demonstrated, however, decision makers commonly deviate from this behavior. Instead, the sum of probabilities assigned to a subset of events tend to exceed the probability assigned to the overarching event (Tversky & Koehler, 1994; Sonnemann *et al.*, 2013). Accordingly, we would expect the interventions of Centralization treatments to lead to a similar effect. This in turn would again affect the dispersion of the stated distribution. According to these considerations, we test the following directional hypotheses:

Hypothesis 4: *In the Compression treatments, the reported distributions are more [less] dispersed, in the less [more] compressed treatments.*

Hypothesis 5: *In the Centralization treatments, the reported distributions are more [less] dispersed, if the number of bins in the central part of the scale is lower [higher].*

2.4.2 Within treatment hypotheses

We conduct tests within the treatments to investigate how consistent participants answer the questions and how our results are moderated by personal characteristics. First, we test the consistency of answers to our three main questions within each treatment. For example, assume a participant in the *ShiftMinus12* treatment states a density forecast with its central moments in the deflation part of the scale. A consistent participant should then also report to expect deflation in the upcoming 12 months in the binary question and should also give a negative inflation rate in the point forecast. Thus, we test:

Hypothesis 6: *Subjects report consistent inflation forecasts.*

In addition, we assume that the effects of our treatment interventions might depend on personal characteristics. Participants with a higher financial literacy or higher education level might be better informed about monetary policy and thus be less susceptible to changes to the scale. Similarly, participants that know the inflation target of the central bank might be more anchored towards this target, expecting the central bank to rein in the inflation rate if it deviates. Additionally, such participants might also feel surer that their answers are right. In line with these deliberations, we test two further, directional, hypotheses:

Hypothesis 7: *Subjects with better education/financial literacy/knowledge of the inflation target are affected less by the treatment interventions.*

Hypothesis 8: *Subjects with better education/financial literacy/knowledge of the inflation target are more certain in their answers.*

2.5 Implementation

The survey was programmed in oTree (Chen *et al.*, 2016). The main experiment consisting of all 13 treatments was conducted online using Prolific (`www.prolific.co`), a UK based commercial subject pool. On Prolific, we recruited a representative sample of the US population. (stratified along sex, age and ethnicity). Data collection started on December 17th and finished on December 19th. In total, 1301 participants completed our survey, with 100 participants per treatment condition, except *Baseline*, which had 101 participants. For the data analysis, 20 participants were dropped for providing beliefs in the density forecast that did not add up to 100, see Table 1.[3] Participants were paid a fixed amount of £1 (worth \$1.33 at the time of the experiment) for completing the survey. On average, it took participants 5:44 minutes to finish the survey. Based on our payment, participants earned on average an hourly wage of \$16.40, well above the average hourly earnings on Prolific.

3 Results of the US Survey

Does the response scale used in the density question affect the survey responses? Figure 4 provides a first view of the evidence. The figure shows the distribution of respondents' mean inflation expectation for all treatments. The *Shift* treatments are shown in the left, the *Compression* treatments in the center, and the *Centralization* treatments in the right panel of the figure. The mean responses are calculated from a smoothed response following the

[3]Participants whose probabilities did not add up to 100 were prompted once to correct their answer, but submitting an answer whose probabilities did not sum up to 100 was possible.

Treatment		Response scale			Demographics		
Nr	Name	#	$\odot$	Span	obs	female	age
1	Baseline	10	0	24	101	47.5%	45.2
2	ShiftMinus12	10	-12	24	99	45.5%	44.5
3	ShiftMinus4	10	-4	24	99	41.4%	47.1
4	ShiftPlus4	10	4	24	98	46.9%	43.5
5	ShiftPlus12	10	12	24	99	47.5%	43.5
6	Compression4	10	0	96	99	60.6%	46.7
7	Compression2	10	0	48	99	50.5%	43.7
8	Compression0.5	10	0	12	96	46.9%	45.1
9	Compression0.25	10	0	6	100	61.0%	45.8
10	Centralization14	14	0	24	97	44.3%	44.9
11	Centralization12	12	0	24	96	47.9%	43.8
12	Centralization8	8	0	24	99	55.6%	46.1
13	Centralization6	6	0	24	99	46.5%	44.2
	Average				**98.6**	**49.4%**	**44.9**

Table 1: **Descriptive Statistics by Treatment**: Number of intervals (#), center of response scale ($\odot$), span of the closed intervals, number of respondents (obs), percentage share of female, and average reported age.

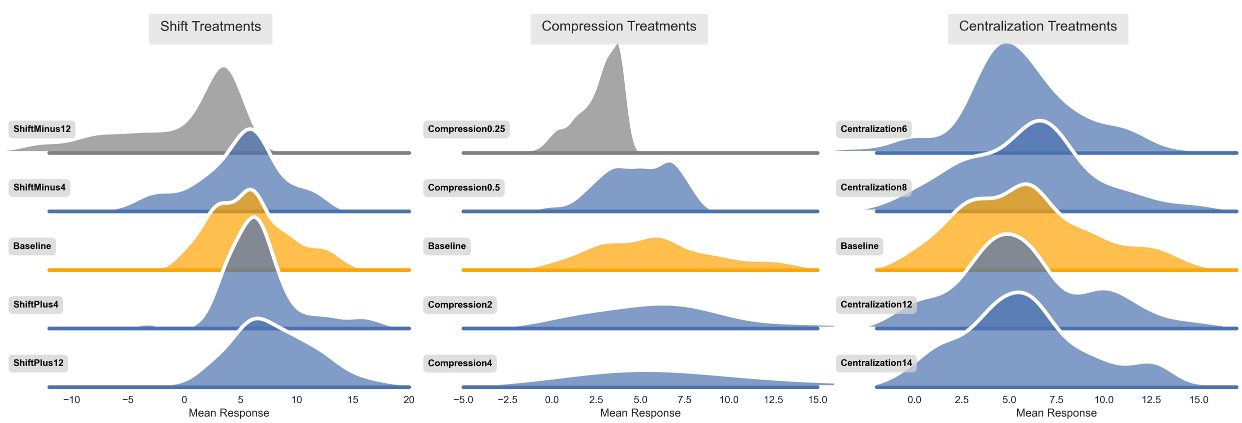

Figure 4: **Distribution of mean forecast**: Kernel density estimates by treatment. Each of the three panels shows the distributions of one of the three treatment types. *Shift* treatments in the left panel, *Compression* treatments in the center panel, and *Centralization* treatments in the right panel. Each panel uses a common y-axis with the *Baseline* shown in orange in the center for comparison. Treatments with large probability mass in the open intervals in gray.

Treatment		**Statistics**									
		Average Mean Forecast				**Average Forecast Uncertainty**				**Disagreement**	
Nr	Name	beta		mass-at-mid-point		beta		mass-at-mid-point		beta	mass-at-mid-point
1	Baseline	5.56	T-Test	5.87	T-Test	3.08	T-Test	3.81	T-Test	3.48	3.63
2	ShiftMinus12	-0.32	(1) 0.000 ***	-0.53	(1) 0.000 ***	3.78	(2) 0.037 **	3.36	(2) 0.228	5.87	5.67
3	ShiftMinus4	4.31	(1) 0.011 **	4.64	(1) 0.014 **	3.30	(2) 0.438	4.10	(2) 0.333	4.13	4.26
4	ShiftPlus4	6.59	(1) 0.019 **	6.83	(1) 0.030 **	3.38	(2) 0.347	4.15	(2) 0.308	3.51	3.58
5	ShiftPlus12	8.21	(1) 0.000 ***	8.39	(1) 0.000 ***	3.59	(2) 0.049 **	4.12	(2) 0.278	4.43	4.47
6	Compression4	10.98	(2) 0.000 ***	11.77	(2) 0.000 ***	8.85	(1) 0.000 ***	11.09	(1) 0.000 ***	11.55	12.39
7	Compression2	6.23	(2) 0.353	6.76	(2) 0.237	6.08	(1) 0.000 ***	7.61	(1) 0.000 ***	6.34	6.54
8	Compression0.5	4.55	(2) 0.013 **	4.81	(2) 0.012 **	1.84	(1) 0.000 ***	2.16	(1) 0.000 ***	2.00	2.03
9	Compression0.25	2.61	(2) 0.000 ***	2.66	(2) 0.000 ***	1.07	(1) 0.000 ***	1.10	(1) 0.000 ***	1.31	1.26
10	Centralization14	5.61	(2) 0.917	5.87	(2) 0.998	3.03	(1) 0.436	3.65	(1) 0.279	3.43	3.52
11	Centralization12	5.57	(2) 0.982	5.85	(2) 0.976	3.33	(1) 0.183	3.97	(1) 0.296	3.59	3.65
12	Centralization8	5.38	(2) 0.734	5.53	(2) 0.541	3.44	(1) 0.117	4.06	(1) 0.210	4.08	4.18
13	Centralization6	5.48	(2) 0.882	5.47	(2) 0.434	4.33	(1) 0.000 ***	4.89	(1) 0.001 ***	3.57	3.65

Table 2: **Treatment differences:** Averages of respondents' mean forecasts, their average forecast uncertainties, and their disagreement, by treatment. beta: Statistics based on a smoothed response. See text for details. mass-at-midpoint: Statistics using mass-at-midpoint assumption. Uncertainty is a measure of spread of a respondent's density forecast and disagreement is the standard deviation of the means of the density forecasts. T-tests assume unequal variance and are one-sided, (1), when specified in the hypotheses, two-sided, (2), otherwise. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

procedure described in Engelberg *et al.* (2009) and Becker *et al.* (2022).[4] As can be seen in the figure, moving the intervals of the scale to the left and right in *Shift* also moves responses to the left and right. Similarly, compressing the scale in *Compression* also compresses the answers.

Table 2 provides an overview of the responses to the density forecast (Q1) and a first answer to the question raised in the title of this manuscript. Inflation expectations elicited via density forecasts change when the response scale is modified. Table 2 shows, for each of

[4] Engelberg *et al.* (2009) suggest to smooth the responses (the histograms) by fitting a parametric distribution from which important statistics such as mean, spread, or tail risk may be computed. The procedure assumes a generalized beta distribution when the respondent assigns positive probabilities to three or more intervals and a triangular distribution when the respondent uses one or two intervals. We denote statistics based on this procedure with the abbreviation "beta". The procedure requires us to make an assumption about the "width" of the open intervals. We assume that the open intervals have twice the width of the adjacent closed interval. Since most treatments only have small amounts of probability mass in the two open intervals, changing this assumption only leads to small changes of the results. The two important exceptions are *ShiftMinus12* and *Compression0.25* (depicted in gray in Figure 4), and care has to be taken when interpreting the figure for these two treatments. Becker *et al.* (2022) extend the original procedure to response scales with unequal interval widths.

the 13 treatments, the average mean forecast, the average uncertainty, and the disagreement of respondents. The statistics are calculated from a smoothed response (see footnote 4) and from a simple mass-at-midpoint measure.

In the *Shift* treatments, a clear movement of the mean forecasts is observed, in line with Hypothesis 1. Shifting the scale to the right, shifts the average mean forecasts to the right. Shifting the scale to the left, shifts the average mean forecasts to the left. The effect is substantial: The shift amounts to $-5.88/-6.40$ (beta/mass at midpoint) for *ShiftMinus12* and $-1.25/-1.23$ for *ShiftMinus4*. In the other direction we find 1.03/0.96 for *ShiftPlus4* and 2.65/2.52 for *ShiftPlus12*.

In the *Compression* treatments, the entire scale is compressed or expanded. We use uncertainty (a measure of the spread of the density forecast) to have a first glance at *Hypothesis 4*, which states that the responses compress or expand when we compress or expand the response scale. Compared to the *Baseline* where the uncertainty is 3.08/3.81 (beta, mass at midpoint), uncertainty clearly increases in "wider" treatments and decreases in "narrower" treatments. Uncertainty in *Compression4* is 9.83/11.09 and 6.08/7.61 in *Compression2*. In the other direction we find uncertainty of 1.84/2.16 in *Compression0.5* and 1.07/1.10 in *Compression0.25*. Since compressing and expanding the scale also leads to a shift of the responses, we also find a knock-on effect on the average mean forecast and on disagreement. When we compress the scale, the average mean forecast is closer to the midpoint of the scale (which is zero in all *Compression* treatments and in *Baseline*), while disagreement is reduced. When we expand the scale, we observe the opposite effect.

Finally, we observe lower uncertainty in treatments with a higher number of intervals at the center of the scale in the *Centralization* treatments (see *Hypothesis 5*). The effect is less pronounced than when compressing or expanding the scale. The uncertainty in treatments with a large number of intervals is lower (*Centralization14*: 3.03/3.65 and *Centralization12*: 3.33/3.97) and in treatments with a smaller number of intervals is larger (*Centralization8*: 3.44/4.06 and *Centralization6*: 4.33/4.89). In Section 3.3 we specifically study whether respondents assign different probability masses to a given range of inflation when we add or remove intervals for this range.

The statistics used in the discussion so far (beta, mass-at-midpoint) require us to make an assumption about the widths of the open intervals, see footnote 4. When we now formally test the hypotheses in Sections 3.1 to 3.3, our tests compare the probability masses respondents assign to given ranges of inflation rates (e.g., the probability mass assigned to deflation). In these tests, no assumption about the "width" of the open intervals is necessary. The results based on probability masses confirm what we have seen so far.

3.1 Shift treatments

Figure 5 depicts average densities assigned to each interval in the *Shift* treatments, relative to the *Baseline*.[5] As the histograms show, the probability mass over the entire scale shifts in the same direction that the scale is shifted. This effect can be more clearly illustrated by concentrating on one side of the scale. Panel A of Figure 6 depicts the average densities respondents put into the deflation part of the scale. The average probability mass respondents put into deflation ranges from 35.67 percent in the *ShiftMinus12* treatment to just 3.11 percent in the *ShiftPlus12* treatment. We test these differences in Table 3, which reports the probability masses in the deflation range in comparison to *Baseline*. One-sided MWU tests and T-tests confirm *Hypothesis 1* for *ShiftMinus12*, *ShiftMinus4* and *ShiftPlus12*. For *ShiftPlus4* the relocation of probability mass goes in the expected direction but is not significant at the 5% level.[6]

Result 1: *Shifting the scale leads to a shift of the responses in the same direction.*

It is worth mentioning that the shifts of the responses are not symmetric in the Plus and Minus treatments, even though the shifts of the response scale are symmetric (see Table

[5] As the histograms depict raw data, we omit the open intervals in order to avoid the need to specify their widths. The probability masses assigned to the open intervals are shown in the lower left and upper right corners of the histograms.

[6] See also Table A3 in the appendix.

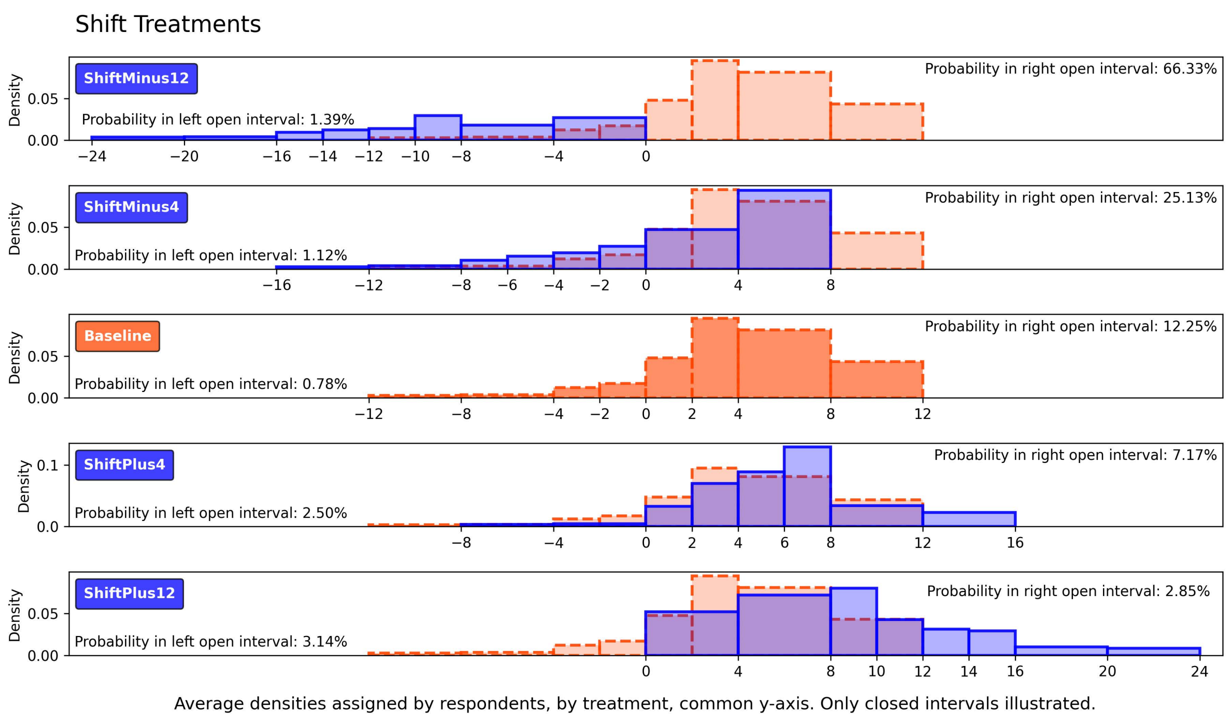

Figure 5: **Shift Treatments:** Histograms of the average densities assigned by the respondents in the Shift treatments. Baseline treatment in center and indicated by dashed bars in the other histograms. The average probabilities assigned to the open intervals are indicated on the left and the rights side of the histograms. Common y-axis.

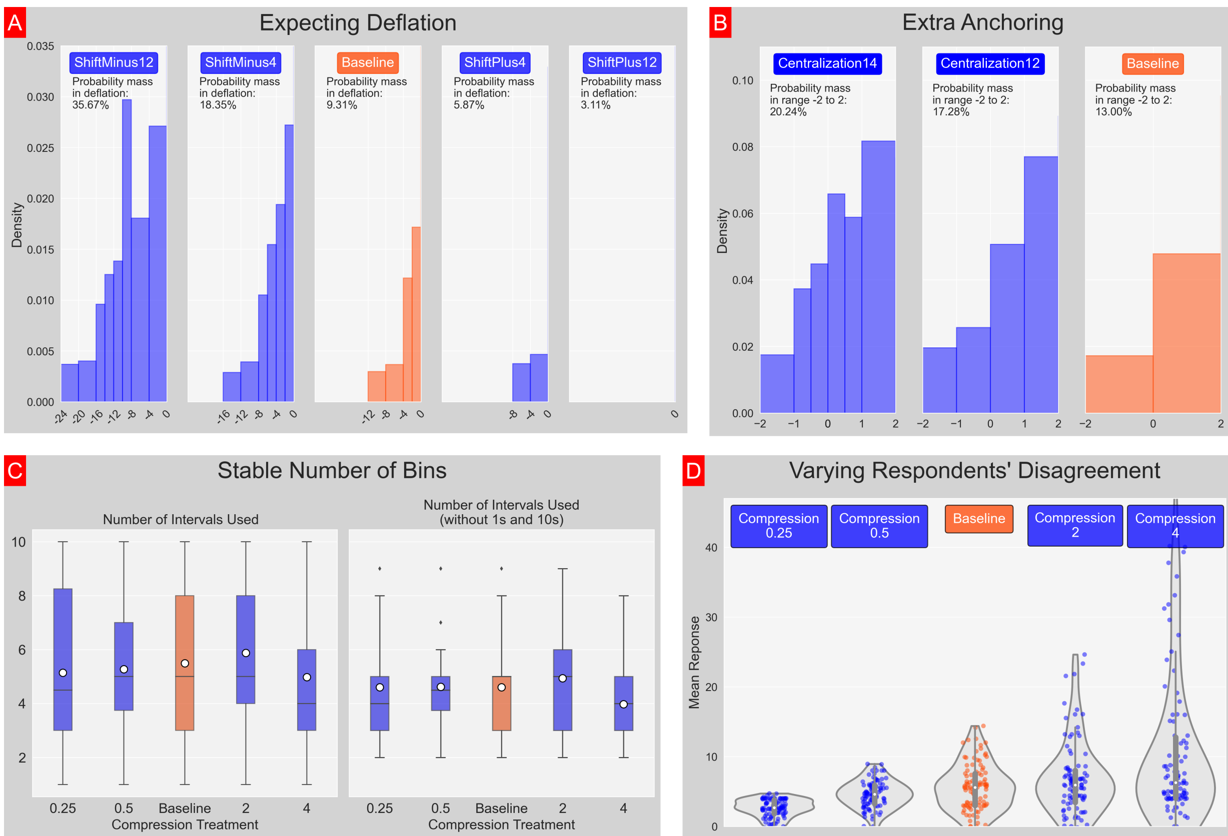

Figure 6: **Panel A**: Average densities assigned to negative inflation rates (deflation) in Shift treatments, by treatment, common y-axis. Only closed intervals illustrated. **Panel B**: Average densities assigned to intervals in the range from -2 to 2 in Centralization treatments, by treatment, common y-axis. **Panel C**: Boxplots of the number of intervals used by the respondent in the Compression treatments, by treatment. Large bright circles indicate averages. **Panel D**: Violin plots and scatterplots (jittered data) of respondents' disagreement in the Compression Treatments.

2). One possible reason for this is that respondents' true (ex-ante) expectations are in the range of positive values. Respondents could arrive at their stated answers by some process which combines their ex-ante expectations with the additional information given by seeing the scale.[7] In this case, their given answers (ex-post beliefs) could be differently influenced by shifting the center of the scale towards or away from their beliefs.

In our design, the density question is asked first, before two additional questions about inflation. The first additional inflation question is an up/down question about whether there will be inflation or deflation and the second question asks for a point forecast. The response scale used in the density question could potentially influence the answers to the succeeding questions. We use Hypotheses 2 and 3 to test this (see Section 2.4). However, there is no such effect: When testing the up/down question answers in the *Shifting* treatments vs *Baseline*, no treatment difference is significant at the 5% level, or higher.[8]

Result 2: *Shifting the response scale does not affect the responses of the succeeding up/down inflation question.*

Similarly, when testing the point forecasts (Q3) in the *Shifting* treatments vs *Baseline*, no treatment difference is significant at the 5% level, or higher.[9]

Result 3: *Shifting the response scale does not affect the responses of the succeeding question asking for point forecasts.*

3.2 Compression treatments

Figure 7 shows the average densities assigned to each interval in the *Compression* treatments relative to *Baseline*. The *Compression* treatments compress or expand the original scale by

[7] This process could be Bayesian updating after the arrival of new information in the form of the scale. Alternatively, respondents could simply be drawn towards the center of the scale and away from the extremes. See the discussion in Section 5.

[8] One-sided Fisher exact tests. One treatment difference is significant at the 10% level: *ShiftMinus4* vs *Baseline* (p=0.056, obs.=200).

[9] T-tests. One treatment difference is significant at the 10% level: *ShiftMinus4* vs *Baseline* (p=0.0854, obs.=200). When testing via Wilxocon-Mann-Whitney tests, no treatment difference is significant at the 10% level, or stronger.

Treatment		Test Range	Probability Mass			Tests (p-values)			
Nr.	Name		Baseline	Treatment	Ratio	MWU		T-Test	
2	ShiftMinus12	< 0	9.31	35.67	3.83	0.000	***	0.000	***
3	ShiftMinus4	< 0	9.31	18.35	1.97	0.014	**	0.002	***
4	ShiftPlus4	< 0	9.31	5.87	0.63	0.229		0.051	*
5	ShiftPlus12	< 0	9.31	3.11	0.33	0.042	**	0.001	***

Table 3: **Shift Treatments**: Average probability masses assigned to negative inflation rates (deflation) in the Shift treatments (the numbers in the table include the masses assigned to the open intervals). Tests for significant treatment difference (one-sided): MWU (Mann-Whitney two-sample statistic) tests, and t-tests (assumes unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

a factor of 0.25/0.5/2/4, keeping the center, the number of bins and their relative widths constant. For example, while the span of the closed intervals in the *Baseline* treatment is from −12% to +12%, the span increases to −48% to +48% in *Compression4*.

Compressing the scale has a strong effect on respondents answers and affects mean forecasts, uncertainty, and disagreement (see Table 2). When we now formally test *Hypothesis 4*, we test for changes in the probability mass respondents assign to given ranges of inflation. Since compressing the scale moves interval boundaries, the treatments require different test ranges. As a rule, we use the largest overlapping range consisting of closed intervals. Table 4 shows that compressing or expanding the scale significantly compresses and expands the stated responses in treatments *Compression4*, *Compression2*, and *Compression0.25*. In these treatments, respondents move probability mass into the center when the scale is compressed and away from the center when the scale is expanded.[10].

Result 4: *Compressing or expanding the response scale leads to compressed and expanded responses.*

Result 4 can be explained via the non-responsive use of intervals by the respondents. A responsive survey taker trying to "fit" a fixed distribution of inflation expectations to the scale would use different numbers of intervals in the different *Compression* treatments. As an example, consider a respondent who expects inflation to fall into the range from 0% to

[10]See also tables A4 to A6 in the appendix

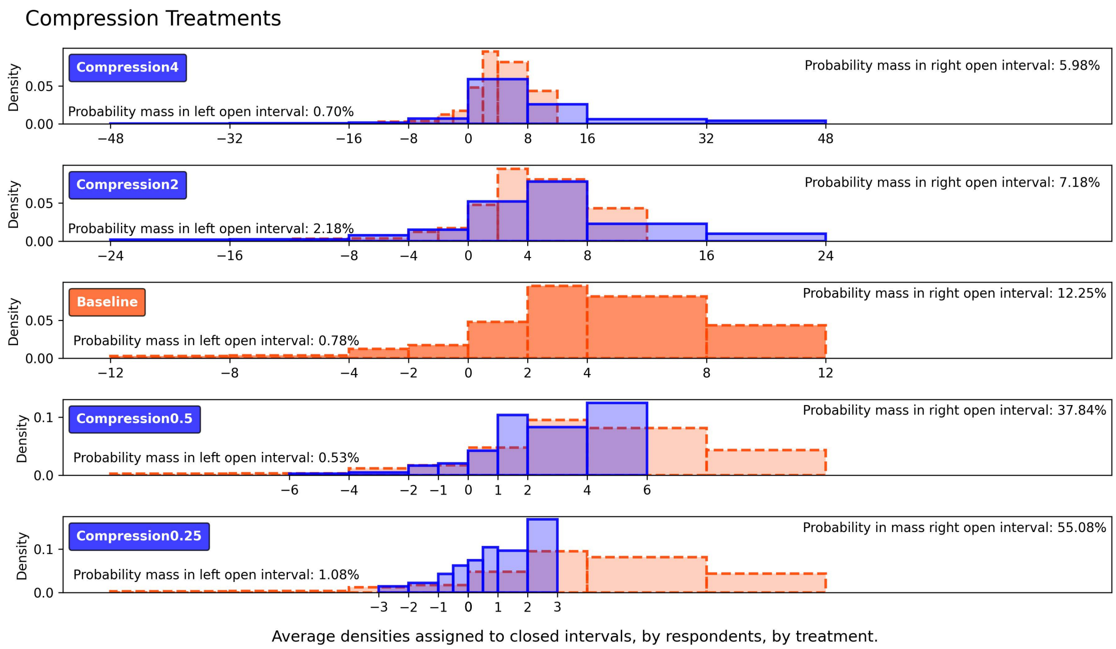

Figure 7: **Compression Treatments:** Histograms of the average densities assigned by the respondents in the Compression treatments. Baseline treatment in center and indicated by dashed bars in the other histograms. The average probabilities assigned to the open intervals are indicated on the left and the rights side of the histograms. Common y-axis.

+8%. In *Compression4*, this respondent needs only a single interval to express her subjective beliefs. In *Compression2*, the respondent needs two intervals and in *Baseline*, the respondent needs 3 intervals. Assuming a responsive use of intervals, one would expect the number of used intervals to decline as the scale gets expanded.

This is not what we find, however. The boxplots in Panel C of Figure 6 show that the average number of intervals respondents use is around 5 and does not vary much from treatment to treatment. The left sub-panel includes all data and the right sub-panel excludes respondents that use only a single interval or all ten intervals. The pattern is the same: Respondents tend to use roughly the same number of intervals independently of the width of the scale. This non-responsive use of intervals may also explain why disagreement declines when we compress the scale (Panel D of Figure 6)

Treatment		Test Range	Probability Mass			Tests (p-values)			
Nr.	Name		Baseline	Treatment	Ratio	MWU		T-Test	
6	Compression4	-8 to 8	68.46	52.71	0.77	0.000	***	0.000	***
7	Compression2	-8 to 8	68.46	61.23	0.89	0.029	**	0.045	**
8	Compression0.5	-4 to 4	34.48	36.07	1.05	0.287		0.355	
9	Compression0.25	-2 to 2	13.00	26.04	2.00	0.000	***	0.000	***

Table 4: **Compression Treatments**: Average probability masses assigned to overlapping ranges in the Compression treatments (the numbers in the table include the masses assigned to the open intervals). Tests for significant treatment difference (one-sided): MWU (Mann-Whitney two-sample statistic) tests, and t-tests (assumes unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

3.3 Centralization treatments

Figure 8 shows the average densities assigned to each interval in the *Centralization* treatments. *Centralization* provides a test of the "asymmetry" in the *Baseline* treatment where the four central intervals are narrower than the four other closed intervals. The *Centralization* treatments test, in particular, whether introducing additional intervals in a given range of inflation will increase the probability mass assigned to this range.

Table 5 shows the average probability masses assigned to the center ranges for which the intervals in *Baseline* and *Centralization* overlap ($[-2, 2]$ in *Centralization14* and *Centralization12*; $[-4, 4]$ in *Centralization8*; and $[-8, 8]$ in Centralization6). As an example, consider the range from -2 to 2 which is covered by the two central intervals in the *Baseline* treatment. *Centralization12* increases this number to four and *Centralization14* to six intervals. Table 5 shows that it is always the treatment with a higher number of intervals in the comparison range that attracts a higher probability mass. T-tests and MWU tests indicate that for *Centralization14*, *Centralization8*, and *Centralization6*, these treatment differences are significant at least at the 5% level. For *Centralization12*, they are significant at the 10% level.[11]. Panel B of Figure 6 illustrates this for *Centralization14* and *Centralization12*.

Result 5: *The probability mass assigned to a given range of the response scale increases with the number of bins used in this range of the scale.*

[11] See also tables A4 to A6 in the appendix

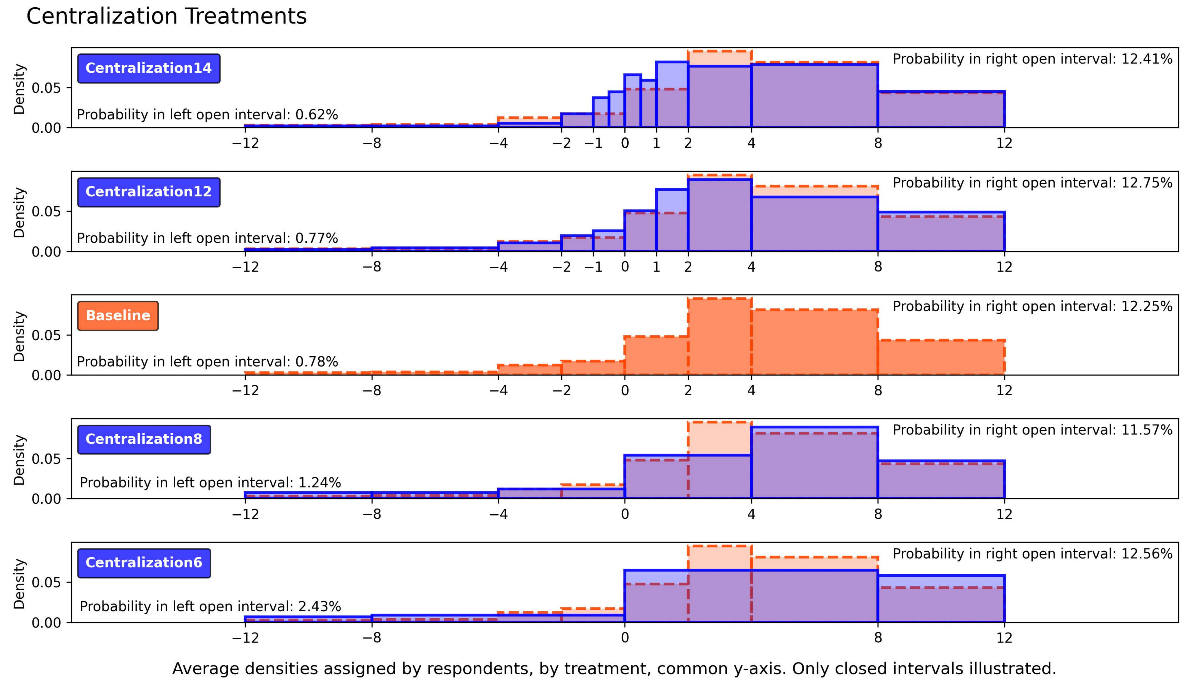

Figure 8: **Centralization Treatments:** Histograms of the average densities assigned by the respondents in the Centralization treatments. Baseline treatment in center and indicated by dashed bars in the other histograms. The average probabilities assigned to the open intervals are indicated on the left and the rights side of the histograms. Common y-axis.

Treatment		**Test Range**	**Probability Mass**			**Tests (p-values)**			
Nr.	Name		Baseline	Treatment	Ratio	MWU		T-Test	
10	Centralization14	-2 to 2	13.00	20.24	1.56	0.037	**	0.007	***
11	Centralization12	-2 to 2	13.00	17.28	1.33	0.077	*	0.055	*
12	Centralization8	-4 to 4	34.48	26.56	0.77	0.042	**	0.028	**
13	Centralization6	-8 to 8	68.46	58.90	0.86	0.003	***	0.011	**

Table 5: **Centralization Treatments**: Average probability masses assigned to overlapping ranges in the Centralization treatments (the numbers in the table include the masses assigned to the open intervals). Tests for significant treatment difference (one-sided): MWU (Mann-Whitney two-sample statistic) tests, and t-tests (assumes unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

The behavior we observe in the *Centralization* treatments has been described in the literature where it is referred to as unpacking bias (Tversky & Koehler, 1994; Sonnemann *et al.*, 2013) or partition dependence (Fox & Rottenstreich, 2003; Benjamin, 2019). A possible explanation for this behavior is that respondents are biased towards assigning roughly equal probability too each provided interval (Fox & Rottenstreich, 2003). Result 5 is especially important since the center of the scale in *Baseline* (and thus in the SCE), that features "additional" intervals, corresponds to the range of answers which is typically associated with "well-anchored" inflation expectations.

3.4 Respondents' internal consistency

In order to test *Hypothesis 6* we follow Engelberg *et al.* (2009) and construct nonparametric bounds on the mean and median of the histograms. The procedure does not impose specific distributional assumptions on the underlying densities. We then examine whether the reported point forecasts fall into the bounds.

For each respondent, we place the probability mass the respondent assigns to an interval at the interval's lower and upper limits. Doing this for each of the ten intervals of the response scale and summing up, we obtain lower and upper bounds on a respondent's mean. To construct the lower and upper bounds on the median, let $j \in \{1, 2, ..., 10\}$ denote the index of the response intervals whose lower bounds we denote θ_j and whose upper bounds we denote θ_{j+1}. With p_{ij} the probability assigned to interval j by respondent i, the median must fall with the interval $[\theta_k, \theta_{k+1}]$ where k is determined by $\sum_{s=1}^{k-1} p_{is} < 0.5$.

Table 6 shows the results of the consistency tests for all 13 treatments and for the SCE (December 2021). The consistency we observe in *Baseline* is comparable to the consistency we find for the SCE. When we compare the *Shift* treatments with *Baseline*, only *ShiftMinus12* shows a significant difference for the mean but not for the median. Table 6 also reports the results for the *Compression* and *Centralization* treatments.

Some caution should be used, however, when interpreting these results. Compressing,

Treatment		Mean	Fisher Exact		Median	Fisher Exact	
1	Baseline	0.624	p-value		0.634	p-value	
2	ShiftMinus12	0.455	0.023	**	0.515	0.115	
3	ShiftMinus4	0.596	0.772		0.636	1.000	
4	ShiftPlus4	0.551	0.316		0.531	0.153	
5	ShiftPlus12	0.545	0.315		0.626	1.000	
6	Compression4	0.727	0.133		0.727	0.174	
7	Compression2	0.616	1.000		0.657	0.769	
8	Compression0.5	0.583	0.662		0.615	0.883	
9	Compression0.25	0.390	0.001	***	0.510	0.088	*
10	Centralization14	0.660	0.657		0.639	1.000	
11	Centralization12	0.583	0.662		0.542	0.196	
12	Centralization8	0.677	0.461		0.707	0.295	
13	Centralization6	0.778	0.021	**	0.747	0.094	*
SCE December 2021		0.575	0.403		0.623	0.915	

Table 6: **Consistency**. The table shows the shares of point forecasts that fall within the bounds on the mean (column 2) or the median (column 4) of the density forecasts, by Treatment and using data from the SCE. All data from December 2021. Two-sided Fisher Exact tests compared to Baseline. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

expanding, shifting, splitting or combining intervals changes the "consistency target" (since not all intervals are equally wide). Wider targets (bounds) are easier to hit. It is therefore not surprising, for example, if the share of respondents with consistent answers grows in *Compression4* and *Centralization6*. In *Compression4*, only two intervals cover the entire range from 0% to 16%. In *Centralization6*, two intervals cover the range from 0% to 12%.

Result 6: *Between* 39.0% *and* 77.8% *of respondents report consistent answers.*

3.5 Financial literacy and answer certainty

After the questions about inflation expectations, we asked subjects to state their subjective certainty for each answer and let them complete a questionnaire with several financial literacy questions. In this section, we will take a look at the interactions of financial literacy and subjective answer certainty.

Subjective answer certainty was elicited via a 6-item Likert scale (*Certain,* ranging from

0=Very Uncertain to 5=Very Certain) asked directly after each inflation expectation question. Financial literacy was elicited in a questionnaire at the end of the experiment. We used the three item financial literacy test by Lusardi & Mitchell (2014), *Finance_lit,* ranging from 0 to 3 correct answers. In addition, we asked participants for their knowledge of the Federal Reserve Bank's inflation target (*Target_correct* dummy) and their level of education (*Education_high* dummy indicates a BA degree or higher).

Hypothesis 7 states that participants with better education, higher financial literacy, and knowledge of the inflation goal should be less affected by changes of the question scale. To evaluate this hypothesis, we regress on the answers in the density question, using the probability mass ranges established in Sections 3.1 to 3.3.[12] Explanatory variables are *Finance_lit, Target_correct, Education_high,* and their interaction terms with treatment dummies. Additional controls are *Certain*, *Female*, and *Age.* Specification (3) of Tables A3 to A6 in the appendix reports the results.

As can be seen, the financial literacy interaction is never significant at the 5% level or stronger for any *Shift* or *Centralization* treatment. It is significant at 5% level for three of the four *Compression* treatments, yet one of these significant results run into the opposite of the predicted direction. The interaction term for knowledge of the inflation target is never significant at 5% level or stronger for any treatment. Having high education leads to significant at 5% level or stronger interactions only for the treatments *ShiftMinus12* and *Centralization8*. Of these, only the former is in the predicted direction. Overall, we find little evidence for *Hypothesis 7*.

Result 7: *There is little evidence that higher educated or more knowledgeable participants are affected less by changes of the response scale of the density question.*

According to *Hypothesis 8*, participants with better education, higher financial literacy, and knowledge of the inflation goal should be more certain in their answers. We regress on

[12] Note that different treatments use different probability mass boundaries: All *Shift* treatments are evaluated via the range of deflation, $(-\infty, 0]$; *Compression4, Compression2,* and *Centralization6* via the range $[-8, 8]$; *Compression0.5* and *Centralization8* via the range $[-4, 4]$; and *Compression0.25* and *Centralization14* via the range $[-2, 2]$, see Section 3.

the *Certain* variable to test this in Table A7 in the appendix. For specifications (1)-(3), the *Certain* on the left side of the regression equation refers to the participants certainty answer after the density forecast, for specifications (4)-(6), it is the certainty answer after the up/down forecast, and for specifications (7)-(9), it is the certainty answer after the point prediction. Explanatory variables are the three measurements of interest, *Finance_lit, Target_correct*, and *Education_high*, as well as controls for *Female*, and *Age.* The final control is the participant's *Forecast* for this question. This variable depends on the question evaluated. For specifications (1)-(3), it is the mean of the fitted beta distribution, a dummy for predicting inflation for (4)-(6), and the value of the point prediction for (7)-(9).

As Table A7 shows, knowing the inflation target makes participants more certain of their answers in all three forecasts. However, for the point prediction, this becomes insignificant when controls are added. Instead, participants become less certain here with higher financial literacy. Education never has a significant influence on subjective certainty.

Result 8: *Respondents who know the Federal Reserve Bank's inflation target are more certain in their forecasts. Higher reported education or financial literacy have no effect on respondents' certainty.*

In general, participants predicting higher inflation with higher probability/inflation over deflation/a higher point prediction are more certain of their answers than participants with lower forecasts. Similarly, older participants are more certain than younger participants for the density forecast and the point prediction (but not significantly so for the up/down forecast), and females are always less certain than males.

4 Results of the German Survey

In addition to the data collected for the US via Prolific, we included two treatments, *Shift-Plus4* and *Centralization14*, in the Bundesbank Online Panel Households (BOP-HH).[13] The

[13]See Beckmann & Schmidt (2020) for a technical description of the BOP-HH Survey. Year-on-year CPI inflation in Germany in June 2022 was reported to be 8.2 percent.

Treatment		Statistics									
		Average Mean Forecast				Average Forecast Uncertainty				Disagreement	
Nr	Name	beta		mass-at-mid-point		beta		mass-at-mid-point		beta	mass-at-mid-point
1	Baseline	6.63	T-Test	6.72	T-Test	2.12	T-Test	2.18	T-Test	4.01	4.06
4	ShiftPlus4	7.22	(1) 0.000 ***	7.27	(1) 0.000 ***	1.79	(2) 0.000 ***	1.89	(2) 0.000 ***	3.56	3.60
10	Centralization14	6.42	(2) 0.073 *	6.50	(2) 0.081 *	2.04	(1) 0.081 *	2.07	(1) 0.066 *	3.87	3.86

Table 7: **Treatment differences for Bundesbank survey:** Averages of respondents' mean forecasts, their average forecast uncertainties, and their disagreement, by treatment. beta: Statistics based on a smoothed response. See text for details. mass-at-midpoint: Statistics using mass-at-midpoint assumption. Uncertainty is the standard deviation of a respondent's forecast and disagreement is the standard deviation of respondents' mean forecasts. T-tests assume unequal variance and are one-sided, (1), when specified in the hypotheses, two-sided, (2), otherwise. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

BOP-HH uses the same question as the SEC to elicit density inflation forecasts and, as such, is identical to our *Baseline* treatment. In June 2022, 4460 German households participated in Wave 30 of the BOP-HH. We removed observations from the sample whenever a household did not report probabilistic inflation expectations or if information for any of the socioeconomic characteristics is missing. We also exclude the response from one household which did not answer the question of whether she expects inflation or deflation. This leaves 4,094 observations in our sample for Wave 30. Of these, 1356 participated in the standard BOP-HH (*Baseline*) question, 1377 in *ShiftPlus4*, and 1361 in *Centralization14*.[14]

In Table 7, we replicate the analysis of Table 2 for this data set. As can be seen, the predicted treatment differences go in the same direction as in the US data. Regarding the *ShiftPlus4* treatment, the expected inflation is 7.22/7.27 (beta/mass-at-midpoint) versus the *Baseline* values of 6.63/6.72. Adding additional bins in the center of the scale in *Centralization14* leads to lower average forecast uncertainty, 2.04/2.07 (beta/mass-at-midpoint) versus 2.12/2.18 in *Baseline.* The differences are highly significant for t-tests of the *ShiftPlus4* treatment differences and weakly significant for the *Centralization14* treatment.

As in Section 3, we also employ tests that directly use the probability masses assigned to

[14] See our companion paper (Becker *et al.*, 2023) for a comparison with the preceding and the subsequent waves.

Treatment		Test Range	Probability Mass			Tests (p-values)			
Nr.	Name		Baseline	Treatment	Ratio	MWU		T-Test	
4	ShiftPlus4	< 0	7.10	3.72	0.52	0.000	***	0.000	***
10	Centralization14	-2 to 2	5.68	8.78	1.55	0.013	**	0.000	***

Table 8: **Shift and Centralization Treatments in Bundesbank Survey:** Average probability masses assigned by the respondents in baseline and treatment (includes probability masses assigned to open intervals). Tests for significant difference (one-sided) in average probability masses. MWU (Mann-Whitney two-sample statistic) and t-test (assumes unequal variances). * Significant at the 0.1 probability level. ** Significant at the 0.05 probability level. *** Significant at the 0.01 probability level.

the intervals. Table 8 repeats the analysis of Tables 3 and 5 for the German data. We find significant differences for both treatments. The size of the treatment effects is astonishingly similar to the US data results. The ratio of probability mass in the deflation region of *ShiftPlus4* is 0.52 times that of the probability mass in the deflation region of *Baseline* in Germany. In the US data, this ratio is 0.63. For *Centralization14*, the probability mass in the middle of the distribution is 1.55 times that of *Baseline* in Germany. In the US, this ratio is 1.56. Overall, despite running the treatments in a different country and at a different time, we find the same direction of treatment effects and very similar effect sizes.

5 Discussion

The results presented in Sections 3 and 4 have important policy implications. We begin the discussion by outlining two possible interpretations of the results. In Section 5.1, we interpret the behavior by participants in the experiment as stemming from behavioral biases (systematic deviations from rational behavior). In Section 5.2 treatment differences are instead interpreted as the result of rational considerations on the side of respondents. We find evidence in favor of both interpretations. Section 5.3 discusses implications for the design of surveys.

5.1 Interpretation 1: Biased answers

Underpinning the interpretation of the results as biased answers is the idea that respondents of the survey are not fully rational. Instead of maintaining coherent probability distributions over future events, such as inflation, and following the updating rules due to Bayes, respondents follow simpler heuristics which deviate from this rational behavior. While generally useful, these heuristics can lead to systematic deviations from rational behavior, which have been the subject of an extensive research program in both psychology and economics (Tversky & Kahneman, 1974). According to this interpretation, the treatment differences we find in *Result 1*, *Result 4* and *Result 5* can be explained via biases that are known in the literature from other settings. The shift of the entire distribution when the response scale shifts, observed in *Result 1*, could be due to the central tendency bias (Hollingworth, 1910; Duffy *et al.*, 2010), the propensity of respondents to prefer answers in the middle of the response scale. *Result 4* and *Result 5* show that respondents are influenced by the existence of additional intervals and seem to distribute their probabilities over a fixed number of intervals, regardless of the underlying interval limits. This is in line with what the literature calls unpacking bias (Tversky & Koehler, 1994; Sonnemann *et al.*, 2013) or partition dependence (Fox & Rottenstreich, 2003; Benjamin, 2019).

One piece of evidence favoring an explanation via behavioral biases is the lack of knock-on effects of the treatment intervention onto later forecasts elicited via different question formats, see *Result 2* and *Result 3*.

5.2 Interpretation 2: Rational updating

The interpretation in this section maintains the assumption that respondents possess a subjective probability distribution of inflation expectations and that they are able to report this distribution in surveys. This implies, for example, that respondents understand the following concepts: inflation, percentages, and percentage changes.

The treatment differences described above can be interpreted as the result of a rational cognitive process in which respondents use two sources of available information and weigh them up against each other before providing an answer. The first source of information is the respondent's prior knowledge about future inflation. Some respondents may possess such information or they may deduce it from information about past or current inflation, possibly combined with information about the macro-economic environment, and with information about the central bank's policy or inflation target.

The second source of information is what is called *context* in the survey research literature, see Schuman (1992) and Schwarz (2010). Context includes any information respondents obtain *due to participating in the survey.* In the case of density questions, the response scale is an important part of the question's context. It is not simply a measurement device, as it provides the respondent with a scale of reference.[15] In the case of inflation expectations, respondents may consider the response scale to reflect the central bank's own expectations of future inflation. By putting certain values of inflation in the center of the scale, the central bank signals that these values are more plausible than values in the peripheral intervals.

Evidence favoring the rational updating interpretation comes from the asymmetry of the treatment differences in Result 1. While a behavioral bias should work similarly in both the *ShiftPlus* and *ShiftMinus* treatments, updating can explain the asymmetry via priors that are not centered on zero.

5.3 Designing surveys

There is no perfect or optimal way to elicit inflation expectations. When designing surveys, researchers have a wide choice of possibilities, each having its advantages and drawbacks. In this section, that complements the discussion in Coibion *et al.* (2020) about survey designs, we first discuss two important trade-offs researchers face when using the density questions of the SCE. We then turn to the question about how best to respond when inflation and inflation expectations are high, and discuss two recent suggestion from the literature that

[15] Respondents may also extrapolate information from less obvious sources such as the wording of a question, the order of a question, or the affiliation of a researcher. Schwarz (2010) gives a good overview.

reduce the influence of the response scale.

Compared to point forecasts or qualitative questions, the advantage of the density questions is that they allow respondents to express the uncertainty they feel when answering the question (Manski (2004), Bruine de Bruin *et al.* (2011)). On the other hand, density questions always involve some framing since the response scale will affect how respondents answer. For central banks, measures of uncertainty are valuable and they may be willing to accept some framing in return for information about uncertainty.

A second trade-off concerns the "asymmetry" of the SCE response scale. The original idea of making the center intervals narrower is to allow respondents to give more precise answers in this range while keeping the overall number of intervals reasonably small. As the results in Section 3.3 show, this feature of the response scale has the consequence that respondents will assign more probability mass to the center intervals than they would otherwise do—giving the spurious impression that values of inflation in the narrow intervals are expected more often.

The unequal widths of the intervals have two other consequences. First, as we discuss in Section 3.4, the standard measures of consistency (Engelberg *et al.*, 2009; Zhao, 2022) will vary when the average inflation forecasts rise or fall. This may happen for example when higher inflation expectations move the responses away from the narrow to the wide intervals. Second, uni-modality is often a desirable property of the responses. The parametric analysis of Engelberg *et al.* (2009), for example, requires uni-modal forecasts. The SCE and our results here show that while the bar-chart of the assigned probabilities is often uni-modal, the implied histograms are not. In our US survey, 112 of the 1281 respondents supply uni-modal probabilities (uni-modal bar-charts), but the implied densities are bi-modal. It is possible that the bi-modal histograms are intentional but it seems more likely that the respondents have difficulties assigning probabilites to intervals with varying widths. Overall, the higher precision that is achieved from the narrow intervals comes at the cost of several distortions.

The SCE response scale was constructed for low inflation rates (inflation rates around the Fed's target of two percent). As we discussed above, for inflation rates around this target, respondents can provide fairly precise forecasts. For inflation expectations above 4 percent, the intervals are wider and for inflation expectations above 12 percent, there is only a single open interval. Higher inflation expectations may then require a shifting of the scale. However, the shifting itself can affect the responses, making a comparison across time difficult. A possible mitigating factor could be double surveys using the new and the old scale in parallel at least for some time.

There are some attempts in the literature to reduce the distortions from the scale. Crosetto & De Haan (2022) suggests a computerized question format in which respondents choose their own intervals. The challenge here is to keep the format simple enough for households to provide reasonable answers. An alternative idea is to center the response scale at the point forecasts.[16] In this design, the impact of the central tendency bias is reduced, likely minimizing the interaction with the other biases. Another advantage of this design is that the response scale shifts automatically when inflation expectations rise or fall.

[16] This design of ours is currently considered for testing by the Bundesbank.

References

Becker, Christoph, Duersch, Peter, Eife, Thomas A, & Glas, Alexander. 2022. Extending the Procedure of Engelberg et al.(2009) to surveys with varying interval-widths. *FAU Discussion Papers in Economics*.

Becker, Christoph, Duersch, Peter, Eife, Thomas A, & Glas, Alexander. 2023. Consistency of households' inflation expectations in high-inflation regimes. *Mimeo*.

Beckmann, Elisabeth, & Schmidt, Tobias. 2020. *Bundesbank online pilot survey on consumer expectations*. Technical Paper.

Benjamin, Daniel J. 2019. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, **2**, 69–186.

Bruine de Bruin, Wändi, Manski, Charles F, Topa, Giorgio, & Van Der Klaauw, Wilbert. 2011. Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics*, **26**(3), 454–478.

Bruine de Bruin, Wändi, Van der Klaauw, Wilbert, Topa, Giorgio, Downs, Julie S, Fischhoff, Baruch, & Armantier, Olivier. 2012. The effect of question wording on consumers' reported inflation expectations. *Journal of Economic Psychology*, **33**(4), 749–757.

Chen, Daniel L, Schonger, Martin, & Wickens, Chris. 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

Coibion, Olivier, Gorodnichenko, Yuriy, Kumar, Saten, & Pedemonte, Mathieu. 2020. Inflation expectations as a policy tool? *Journal of International Economics*, **124**, 103297.

Coibion, Olivier, Gorodnichenko, Yuriy, & Weber, Michael. 2022. Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy*, **130**(6).

Crosetto, Paolo, & De Haan, Thomas. 2022. Comparing input interfaces to elicit belief distributions. *GAEL Working Paper*.

Duffy, Sean, Huttenlocher, Janellen, Hedges, Larry V, & Elizabeth Crawford, L. 2010. Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review*, **17**(2), 224–230.

Engelberg, Joseph, Manski, Charles F, & Williams, Jared. 2009. Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, **27**(1), 30–41.

Fox, Craig R, & Rottenstreich, Yuval. 2003. Partition priming in judgment under uncertainty. *Psychological Science*, **14**(3), 195–200.

HOLLINGWORTH, HARRY LEVI. 1910. The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, **7**(17), 461–469.

KING, ROBERT G, LU, YANG K, & PASTÉN, ERNESTO S. 2008. Managing expectations. *Journal of Money, Credit and banking*, **40**(8), 1625–1666.

LUSARDI, ANNAMARIA, & MITCHELL, OLIVIA S. 2014. The economic importance of financial literacy: Theory and evidence. *Journal of economic literature*, **52**(1), 5–44.

MANSKI, CHARLES F. 2004. Measuring expectations. *Econometrica*, **72**(5), 1329–1376.

MANSKI, CHARLES F. 2018. Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise. *NBER Macroeconomics Annual*, **32**, 411–471.

PAYNE, STANLEY L. (ed). 1951. Princeton: Princeton University Press.

PHILLOT, MAXIME, & ROSENBLATT-WISCH, RINA. 2018. *Inflation expectations: The effect of question ordering on forecast inconsistencies*. Tech. rept. Swiss National Bank.

SCHUMAN, HOWARD. 1992. *Context Effects: State of the Past/State of the Art*. New York, NY: Springer New York. Pages 5–20.

SCHWARZ, NORBERT. 2010. *Measurement as Cooperative Communication: What Research Participants Learn from Questionnaires*. Los Angeles: Sage. Pages 43–61.

SCHWARZ, NORBERT, HIPPLER, HANS-J, DEUTSCH, BRIGITTE, & STRACK, FRITZ. 1985. Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, **49**(3), 388–395.

SONNEMANN, ULRICH, CAMERER, COLIN F., FOX, CRAIG R., & LANGER, THOMAS. 2013. How psychological framing affects economic market prices in the lab and field. *PNAS*, **110**(29), 11779–11784.

SUDMAN, SEYMOUR, & BRADBURN, NORMAN (eds). 1974. Chicago: Aldine.

TVERSKY, AMOS, & KAHNEMAN, DANIEL. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, **185**(4157), 1124–1131.

TVERSKY, AMOS, & KOEHLER, DEREK J. 1994. Support Theory: A Nonextensional Representation of Subjective Probability. *Psychological Review*, **101**(4), 547–567.

WOODFORD, MICHAEL. 2005. *Central bank communication and policy effectiveness*.

ZHAO, YONGCHEN. 2022. Internal consistency of household inflation expectations: Point forecasts vs. density forecasts. *International Journal of Forecasting*.

A Appendix: Instructions

Now we would like you to think about different things that may happen to inflation over the next 12 months. We realize that this question may take a little more effort. In your view, what would you say is the percent chance that, over the next 12 months,		
the rate of inflation will be 12% or higher		percent chance
the rate of inflation will be between 8% and 12%		percent chance
the rate of inflation will be between 4% and 8%		percent chance
the rate of inflation will be between 2% and 4%		percent chance
the rate of inflation will be between 0% and 2%		percent chance
the rate of deflation (opposite of inflation) will be between 0% and 2%		percent chance
the rate of deflation (opposite of inflation) will be between 2% and 4%		percent chance
the rate of deflation (opposite of inflation) will be between 4% and 8%		percent chance
the rate of deflation (opposite of inflation) will be between 8% and 12%		percent chance
the rate of deflation (opposite of inflation) will be 12% or higher		percent chance
		Total

Table A1: **Density question in the Baseline Treatment.**

B Appendix: Tables

Treatment		Response scale				Results									
						Forecast				Intervals				Unimodal	
Nr	Name	#	⊙	Span	obs	Point	Mean	Uncertainty	Disagreement	Used	Used open	Gaps	Single	Probability	Density
1	Baseline	10	0	24	101	8.30	5.56	3.08	3.48	5.50	0.96	0.149	0.020	0.86	0.81
2	ShiftMinus12	10	-12	24	99	7.52	-0.32	3.78	5.87	4.59	1.15	0.162	0.242	0.86	0.72
3	ShiftMinus4	10	-4	24	99	5.88	4.31	3.30	4.13	5.06	1.23	0.152	0.030	0.74	0.71
4	ShiftPlus4	10	4	24	98	7.50	6.59	3.38	3.51	6.39	0.87	0.204	0.010	0.76	0.70
5	ShiftPlus12	10	12	24	99	8.66	8.21	3.59	4.43	6.62	0.80	0.020	0.020	0.91	0.87
6	Compression4	10	0	96	99	11.83	10.98	8.85	11.55	4.98	0.63	0.030	0.010	0.87	0.84
7	Compression2	10	0	48	99	9.39	6.23	6.08	6.34	5.88	0.81	0.051	0.020	0.76	0.79
8	Compression0.5	10	0	12	96	6.50	4.55	1.84	2.00	5.27	1.14	0.125	0.052	0.82	0.67
9	Compression0.25	10	0	6	100	5.66	2.61	1.07	1.31	5.14	1.24	0.160	0.150	0.78	0.60
10	Centralization14	14	0	24	97	6.31	5.61	3.03	3.43	6.72	0.89	0.340	0.010	0.72	0.60
11	Centralization12	12	0	24	96	9.22	5.57	3.33	3.59	6.07	0.90	0.333	0.010	0.74	0.68
12	Centralization8	8	0	24	99	7.53	5.38	3.44	4.08	4.64	0.96	0.081	0.030	0.81	0.82
13	Centralization6	6	0	24	99	6.00	5.48	4.33	3.57	4.00	1.08	0.010	0.010	0.92	0.83
	Average	10	0		98.56	7.72	5.44	3.78	5.62	5.45	0.97	0.139	0.048	0.81	0.74

Table A2: **Descriptive Statistics by Treatment:** Number of intervals (#), center of response scale (⊙), span of the closed intervals, number of respondents (obs). Results report averages by treatment: Point forecast (point, Q3 in questionnaire), mean (Q1 in questionnaire, from fitted smoothed response (beta)), uncertainty (standard deviation of fitted smoothed response), disagreement (standard deviation of mean forecasts), number of intervals used (used), number of open intervals used (used open), number of gaps in response (gaps), share of responses using single interval, share of unimodal response (based on barchart of probabilities or on histogram).

Probability Mass in Deflation	(1)		(2)		(3)	
ShiftMinus12	26.36***	(0.000)	26.91***	(0.000)	31.02***	(0.004)
ShiftMinus4	9.047***	(0.006)	9.605***	(0.002)	27.58***	(0.004)
ShiftPlus4	-3.440	(0.300)	-4.758	(0.131)	-5.536	(0.601)
ShiftPlus12	-6.196*	(0.062)	-8.001**	(0.011)	-23.28**	(0.013)
Certain			-5.537***	(0.000)	-5.357***	(0.000)
Female			0.0117	(0.995)	-2.525	(0.226)
Age			-0.0960	(0.122)	-0.00999	(0.874)
Finance_lit					-5.521**	(0.032)
Finance_lit×ShiftMinus12					4.937	(0.264)
Finance_lit×ShiftMinus4					-4.698	(0.231)
Finance_lit×ShiftPlus4					1.722	(0.693)
Finance_lit×ShiftPlus12					7.072*	(0.051)
Target_correct=1					-3.772	(0.394)
Target_correct=1×ShiftMinus12					-9.885	(0.125)
Target_correct=1×ShiftMinus4					2.125	(0.742)
Target_correct=1×ShiftPlus4					2.833	(0.655)
Target_correct=1×ShiftPlus12					-0.286	(0.964)
Education_high=1					0.427	(0.927)
Education_high=1×ShiftMinus12					-14.94**	(0.026)
Education_high=1×ShiftMinus4					-10.89*	(0.099)
Education_high=1×ShiftPlus4					-6.781	(0.306)
Education_high=1×ShiftPlus12					-2.055	(0.751)
Constant	9.307***	(0.000)	29.71***	(0.000)	41.61***	(0.000)
Observations	496		495		493	
Adjusted R^2	0.197		0.267		0.327	

Table A3: **Regressions Shift Treatments:** OLS regressions on probability mass in intervals in range ≤ 0. p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Probability mass in range $[-8, 8]$	(1)		(2)		(3)	
Compression4	-15.75***	(0.000)	-14.46***	(0.001)	-46.55***	(0.000)
Compression2	-7.223*	(0.095)	-7.170*	(0.091)	-20.41*	(0.098)
Centralization6	-9.556**	(0.027)	-9.392**	(0.026)	-29.54**	(0.026)
Certain			-1.655	(0.179)	-1.866	(0.123)
Female			-12.78***	(0.000)	-9.459***	(0.003)
Age			0.222**	(0.017)	0.180*	(0.055)
Finance_lit					-2.433	(0.485)
Finance_lit×Compression4					14.71***	(0.003)
Finance_lit×Compression2					5.823	(0.234)
Finance_lit×Centralization6					10.10*	(0.051)
Target_correct=1					3.043	(0.612)
Target_correct=1×Compression4					-4.367	(0.605)
Target_correct=1×Compression2					-9.777	(0.259)
Target_correct=1×Centralization6					-3.399	(0.686)
Education_high=1					10.65*	(0.090)
Education_high=1×Compression4					-1.347	(0.879)
Education_high=1×Compression2					9.103	(0.306)
Education_high=1×Centralization6					-3.167	(0.715)
Constant	68.46***	(0.000)	69.30***	(0.000)	67.52***	(0.000)
Observations	398		398		398	
Adjusted R^2	0.026		0.074		0.138	

Table A4: **Regressions Compression4, Compression2, Centralization6 Treatments:** OLS regressions on probability mass in intervals in range $[-8, 8]$. p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Probability mass range $[-4, 4]$	(1)		(2)		(3)	
Compression0.5	1.598	(0.703)	2.291	(0.579)	-22.38*	(0.085)
Centralization8	-7.920*	(0.058)	-7.521*	(0.067)	-18.85	(0.128)
Certain			-5.103***	(0.001)	-5.522***	(0.000)
Female			-1.517	(0.659)	-1.879	(0.607)
Age			-0.0965	(0.361)	-0.0582	(0.595)
Finance_lit					-5.449	(0.120)
Finance_lit×Compression0.5					12.65**	(0.011)
Finance_lit×Centralization8					6.097	(0.228)
Target_correct=1					-2.072	(0.729)
Target_correct=1×Compression0.5					-5.725	(0.500)
Target_correct=1×Centralization8					15.11*	(0.075)
Education_high=1					5.217	(0.404)
Education_high=1×Compression0.5					-4.453	(0.609)
Education_high=1×Centralization8					-17.58**	(0.048)
Constant	34.48***	(0.000)	54.37***	(0.000)	64.93***	(0.000)
Observations	296		296		295	
Adjusted R^2	0.013		0.048		0.068	

Table A5: **Regressions Compression0.5, Centralization8 Treatments:** OLS regressions on probability mass in intervals in range $[-4, 4]$. p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Probability mass range $[-2, 2]$	(1)		(2)		(3)	
Compression0.25	13.04***	(0.000)	13.82***	(0.000)	16.73*	(0.073)
Centralization14	7.237**	(0.020)	6.474**	(0.026)	4.927	(0.599)
Centralization12	4.281	(0.168)	2.076	(0.476)	-3.653	(0.680)
Certain			-5.476***	(0.000)	-5.486***	(0.000)
Female			3.781*	(0.077)	1.620	(0.466)
Age			-0.160**	(0.013)	-0.109	(0.103)
Finance_lit					-4.803**	(0.049)
Finance_lit×Compression0.25					1.169	(0.755)
Finance_lit×Centralization14					2.554	(0.476)
Finance_lit×Centralization12					4.373	(0.213)
Target_correct=1					2.468	(0.554)
Target_correct=1×Compression0.25					-14.72**	(0.012)
Target_correct=1×Centralization14					-10.93*	(0.068)
Target_correct=1×Centralization12					-5.392	(0.359)
Education_high=1					-0.641	(0.883)
Education_high=1×Compression0.25					2.561	(0.668)
Education_high=1×Centralization14					2.912	(0.636)
Education_high=1×Centralization12					-3.301	(0.586)
Constant	13***	(0.000)	34.32***	(0.000)	43.49***	(0.000)
Observations	394		393		393	
Adjusted R^2	0.039		0.165		0.188	

Table A6: **Regressions Compression0.25, Centralization14, Centralization12 Treatments:** OLS regressions on probability mass in intervals in range $[-2, 2]$. p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

	Density			Up/down			Point prediction		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Target correct	0.384***	0.308***	0.295***	0.292***	0.238***	0.229***	0.140**	0.0916	0.100
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.027)	(0.142)	(0.108)
Finance lit.	0.0724	-0.0471	-0.0454	0.0520	0.0254	0.0267	-0.0477	-0.125***	-0.125***
	(0.110)	(0.306)	(0.315)	(0.192)	(0.538)	(0.513)	(0.226)	(0.002)	(0.002)
Education high	0.0326	0.0402	0.0394	0.0173	0.0269	0.0307	-0.0388	-0.0300	-0.0241
	(0.656)	(0.569)	(0.568)	(0.788)	(0.671)	(0.625)	(0.544)	(0.632)	(0.698)
Forecast		0.0302***	0.0488***		0.0341***	0.0450***		0.0249***	0.0318***
		(0.000)	(0.000)		(0.000)	(0.000)		(0.000)	(0.000)
Female		-0.513***	-0.530***		-0.203***	-0.212***		-0.302***	-0.317***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.000)	(0.000)
Age		0.0123***	0.0117***		0.00133	0.00137		0.00878***	0.00851***
		(0.000)	(0.000)		(0.493)	(0.476)		(0.000)	(0.000)
Constant	2.437***	2.305***	2.263***	3.518***	3.464***	3.700***	2.815***	2.646***	2.680***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Treatment dummies	no	no	yes	no	no	yes	no	no	yes
Adjusted R^2	0.027	0.096	0.135	0.019	0.053	0.076	0.002	0.045	0.055
Observations	1277	1276	1276	1278	1277	1277	1275	1274	1274

Table A7: **Regressions Certainty:** OLS regressions on certainty. Forecast is the mean of the beta distribution for (1)-(3), a dummy for predicting inflation for (4)-(6), and the value of the point prediction for (7)-(9). p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.