HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Measuring Inflation Expectations: How the Response Scale Shapes Density Forecasts

Christoph K. Becker

Peter Duersch

Thomas A. Eife

AWI DISCUSSION PAPER SERIES NO. 727
April 2023

Measuring Inflation Expectations: How the Response Scale Shapes Density Forecasts

Christoph K. Becker[*] Peter Duersch[†] Thomas A. Eife[‡]

April 24, 2023

Abstract

In density forecasts, respondents are asked to assign probabilities to a response scale with pre-specified ranges of inflation. In two large-scale experiments, one conducted in the US and one in Germany, we show how the specifics of the response scale determine the responses: Shifting, compressing, or expanding the scale leads to shifted, compressed, and expanded forecasts. Mean forecast, uncertainty, and disagreement vary by several percentage points. The findings have implications for survey design and for central banks' optimal adjustment of the response scales during times of high inflation.

JEL codes: C83, D84, E31
Keywords: Inflation, density forecast, experiment, survey

Acknowledgements

We thank Kenza Benhima, Muhammed Bulutay, Christian Conrad, Timo Dimitriadis, Zeno Enders, Frank Heinemann, Luba Petersen, Shyam Sunder, 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.

This RCT was registered as AEARCT Identification Number AEARCTR-0008716 (Becker, Duersch, Eife; 2021).

[*] Heidelberg University, Bergheimer Str. 58, 69115 Heidelberg, christoph.becker@awi.uni-heidelberg.de.
[†] University of Mannheim, L7, 3–5, 68161 Mannheim, duersch@xeeron.de.
[‡] Heidelberg University, Bergheimer Str. 58, 69115 Heidelberg, thomas.eife@awi.uni-heidelberg.de.

1 Introduction

Managing inflation expectations is an important part of modern central banking. When interest rates reached levels around zero after the financial crisis of 2008, central banks widely adopted this non-conventional policy tool. Managing expectations requires measuring expectations, and several new surveys have been established in the past decade, many using density questions. In this question format, respondents are given a response scale with prespecified intervals and are asked to assign probabilities to the intervals that best represent their beliefs about inflation.

The experimental results we present in this paper show how the specifics of the response scale determine the responses. Shifting or compressing the response scale causes respondents to shift or compress their answers. For example, we can vary respondents' mean inflation forecast from -0.32% to 8.15% simply by shifting the response scale. Similarly, we can double respondents' average uncertainty (the standard deviation of their response) from 3.08% to 6.08%, when we double the width of the 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) if the scale is not taken into account. Differencing inflation beliefs to obtain changes of expectations over time does not solve the problem since the distortion itself can change over time.

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 and Rosenblatt-Wisch (2018) discuss the effect of question ordering on respondents' forecast consistency. The effect of a question's wording is addressed in several papers. Bruine de Bruin et al. (2012) study whether asking for "prices in general", "inflation", or "prices you pay" affects the responses and conclude that inflation expectations were lower and less dispersed when asking for "inflation" (see also Manski, 2018; Bruine de Bruin et al., 2023). Asking for the "overall inflation rate" or for "prices overall in the economy" does not appear to systematically affect the results (Coibion et al., 2020). Providing additional information in the question (e.g., a newspaper article or a statement about the Federal Reserve's inflation target) affects households' responses (Coibion et al., 2022).

Our focus here is on variations of the response scale. In an 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 shift 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. Using the New York Fed's Survey of Consumer Expectations (SCE) question on inflation expectations as our baseline, we employ a battery of 12 treatments to systematically test whether and 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. The final four treatments study the consequences of the irregular spacing of the response scale of the SCE, where the four center intervals are narrower than the other closed intervals. That is, 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 (Germany's central bank) 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. Section 3 presents the results of the US survey and Section 4 the results of the German survey. Section 5 interprets the results, discusses possible improvements of density questions and how central banks can adjust the response scales during times of high inflation. Section 6 concludes.

2 Experimental design

We use three questions from the New York Fed Survey of Consumer Expectations (SCE) in our experiment. First, survey respondents are asked to provide a density forecast for 12-month ahead inflation (question Q9 in the SCE). Second, respondents report whether they expect inflation or deflation in a binary question (Q8v2 in the SCE). Third, we ask for a point forecast (Q8v2part2 in the SCE). Since we are primarily interested in the density forecast, our ordering of the questions differs from the ordering of the SCE. We move the density forecast in first place to prevent the other two questions (especially the point forecast) from confounding the responses of the density forecast. Several other surveys, such as the Bundesbank household survey, adopted the response scale of the SCE.[1]

We use the response scale of the SCE as our *Baseline* treatment. The scale has ten intervals, including two open outer intervals, and is centered at zero. The closed intervals range from -12% to 12%. The four central intervals are narrower than the outer closed intervals, see Figure 1. As in the SCE, the question asking for a point forecast varies

[1] Bruine de Bruin et al. (2023) discuss the history of eliciting expectations in economics and provide an overview of current surveys that use probabilistic (density) questions.

depending on whether respondents expect inflation or deflation in the binary question. When a respondent expects deflation in the binary question, the point forecast asks for a deflation rate. When a respondent expects inflation, the point forecast asks for an inflation rate.

After each of these three questions, respondents are asked to indicate how certain they feel about their answer on a 6-item Likert scale (ranging from Very Uncertain to Very Certain). Following these six main questions, respondents are asked to answer a short questionnaire about their age, gender, political orientation and state of residence. Additionally, the questionnaire includes three measures of potentially relevant knowledge: A question on highest education degree obtained, a question on their knowledge of the Fed's inflation target, and three questions on financial literacy taken from Lusardi and Mitchell (2014). Finally, the questionnaire includes a control question to test the attentiveness of the respondents.

In the survey, respondents face one of 13 different treatment conditions. The response scale in *Baseline* is identical to the SCE. The other 12 treatment conditions introduce different variations to the response scale. All other questions are the same in all treatments. Hence, our design allows us to isolate the effect of changing the scale of the density forecast on the forecast itself, but also on subsequent assessments of 12-month 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 (number of intervals and their relative widths) constant. This means that the center of the scale moves away from zero compared to *Baseline.* The *Shift* treatments allow us to test how respondents' forecasts are influenced by different positions of the scale on the number line. We implement both shifts in two different degrees, resulting in a total of four *Shift* treatments: *ShiftMinus12*, *ShiftMinus4*, *ShiftPlus4*, and *ShiftPlus12.* Figure 1 illustrates the four *Shift* treatments, with *Baseline* as a reference. In *ShiftMinus12* and *ShiftMinus4* we subtract 12 and 4 respectively from all interval limits. Conversely, in *ShiftPlus4* and *ShiftPlus12*, we add 4 and 12 to the interval limits.

2.2 Compression treatments

In the four *Compression* treatments, the interval limits of the response scale are multiplied by a constant factor, keeping the number and the relative size of the intervals unchanged. For factors below 1, this leads to a compression of the response scale around the center. Factors

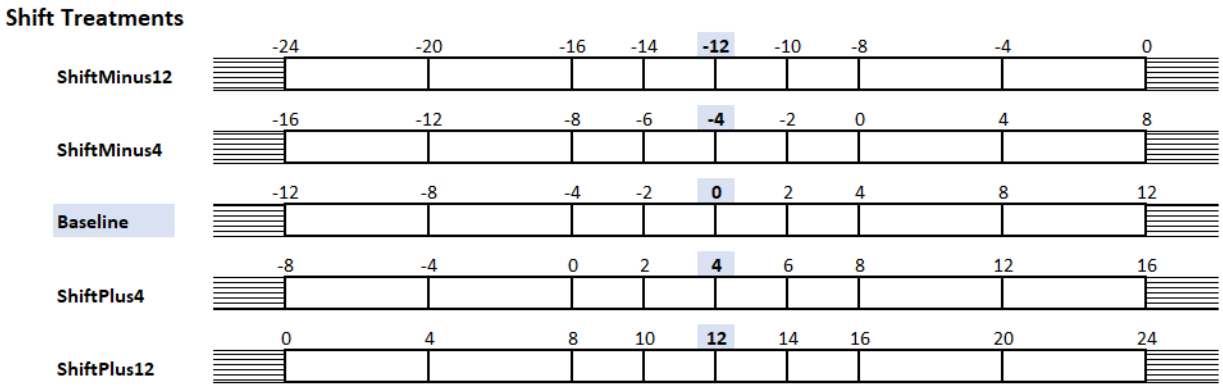


Figure 1: Response scales used in the **Shift treatments** (with Baseline for reference).

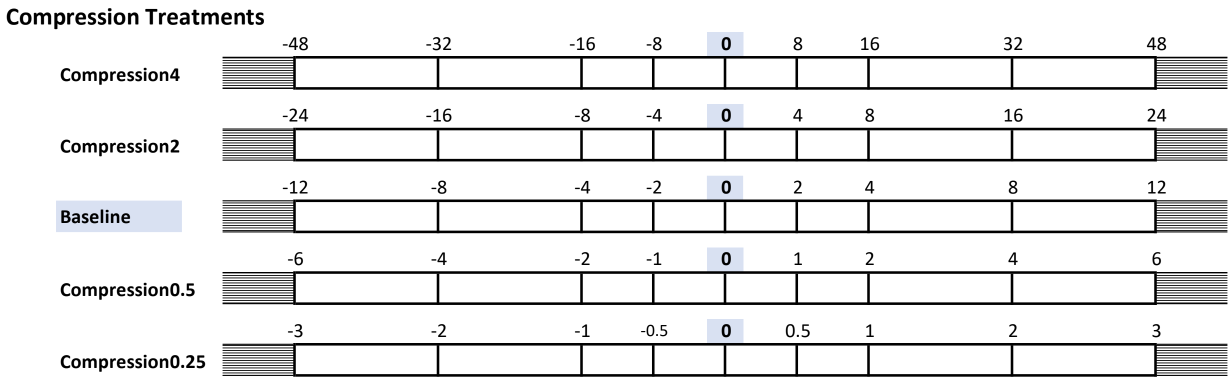


Figure 2: Response scales used in the **Compression treatments** (with Baseline for reference).

above 1 result in an expansion (decompression). As before, we implement both compression and decompression with two different degrees, giving us four *Compression* treatments: *Compression0.25*, *Compression0.5*, *Compression2*, and *Compression4*.

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 contrast, *Compression2* and *Compression4* widen the intervals. As Figure 2 illustrates, 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 respondents to better communicate beliefs about high inflation and high deflation rates, they also imply a coarser image of respondents' inflation beliefs around the center.

2.3 Centralization treatments

Finally, the four *Centralization* treatments vary the number of intervals around the center of the scale. Differently to the other two treatment categories, where the scale is either shifted

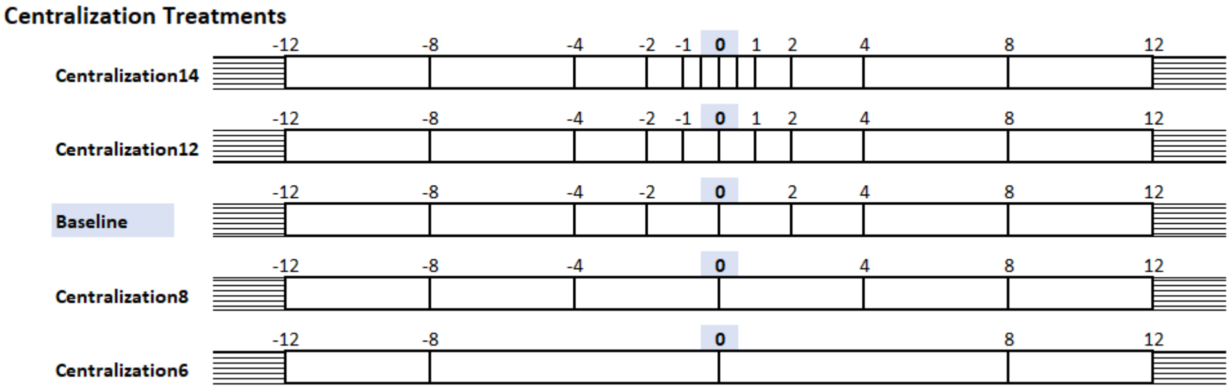


Figure 3: Response scales used in the **Centralization treatments** (with Baseline for reference).

or compressed, the overall span of the scale is identical to *Baseline*. Instead, we either split existing intervals around the center or we combine them, thus changing the number of intervals. Similarly to the *Compression* treatments, this allows for a finer or coarser image of respondents' inflation beliefs around the center, however, without changing the span 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 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. In *Centralization8* all closed intervals have the same width. Respondents in these treatments thus can only state very coarse beliefs. In *Centralization12* and *Centralization14*, on the other hand, we split the intervals around the center allowing respondents to more finely express beliefs in this range.

2.4 Hypotheses

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

2.4.1 Across treatment hypotheses

Regarding the *Shift* treatments, if respondents use the scale as a reference, we would expect the responses to shift in the same direction.

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

After the density forecast, respondents answer the binary inflation/deflation question and the point forecast. If the treatment interventions from the density forecast carry over to the two subsequent inflation questions, we expect the responses in the *Shift* treatments to differ from *Baseline*. The *Shift* treatments provide two intuitive predictions to test:

Hypothesis 2: *In the Shift 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 of the scale.*

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

In the *Compression* treatments we compress or expand the entire scale. If respondents use the scale as a reference, they should compress or expand their belief distribution. Thus, we expect the dispersion to differ from *Baseline.*

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

The *Centralization* treatments split or merge the intervals around the center of the scale. After splitting an interval into two smaller intervals, respondents can still provide the same response, but earlier literature has demonstrated that splitting and merging affects the responses. The sum of probabilities assigned to a subset of events typically exceeds the probability assigned to the overarching event (Tversky and Koehler, 1994; Sonnemann et al., 2013). Following this logic, splitting intervals around the center of the scale would lead to more probability mass being concentrated around the center. Accordingly, we would expect the dispersion being affected by the number of intervals around the center.

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

2.4.2 Within treatment hypotheses

In the within-treatment hypotheses, we study respondents' internal consistency and how our results are moderated by personal characteristics. We define consistency as a respondent's point forecast being compatible with the respondent's density forecast.

Hypothesis 6: *Subjects report consistent inflation forecasts.*

The effects of our treatment interventions might depend on a respondents' *proficiency* concerning monetary policy. As outlined above, we use three measures to capture different aspects of a respondents' proficiency. Respondents 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, respondents 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 from the target. Coibion et al. (2018), for example, show that managers' inflation expectations strongly react to receiving information about the central bank's inflation target. Additionally, such respondents might also feel surer that their answers are correct. In line with these deliberations, we test two further, directional hypotheses:

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

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

3 The US survey

Treatment	**Response scale**			**Demographics**				
	#	Center	Span	Obs	Avg. age	Share female	Share white	Share black
Baseline	10	0	24	101	45.25	0.48	0.71	0.13
ShiftMinus12	10	-12	24	99	44.45	0.45	0.74	0.11
ShiftMinus4	10	-4	24	99	47.09	0.41	0.78	0.12
ShiftPlus4	10	4	24	98	43.46	0.47	0.68	0.18
ShiftPlus12	10	12	24	98	43.64	0.48	0.77	0.15
Compression4	10	0	96	99	46.75	0.61	0.84	0.05
Compression2	10	0	96	99	43.70	0.51	0.72	0.15
Compression0.5	10	0	12	96	45.09	0.47	0.74	0.18
Compression0.25	10	0	6	100	45.80	0.61	0.82	0.09
Centralization14	14	0	24	96	44.90	0.45	0.73	0.15
Centralization12	12	0	24	96	43.85	0.48	0.71	0.14
Centralization8	8	0	24	99	46.08	0.56	0.85	0.11
Centralization6	6	0	24	99	44.18	0.46	0.75	0.14
Average				**98.4**	**44.95**	**0.49**	**0.76**	**0.13**

Table 1: **Descriptive statistics by treatment.** Number of intervals (#), center of response scale, span of the closed intervals, number of respondents (obs), average reported age, and percentage share of female. The last two columns show the share of people identifying as white or black as recorded by Prolific.

We conducted the survey in the US in December 2021. For this month, the Bureau of Labor Statistics reports year-on-year inflation of 7.2 percent. This is somewhat higher than in the preceding five months, where inflation averaged at around 6 percent. Especially energy prices had been increasing in the months before the survey. For November 2021, for example, the Bureau of Labor Statistics reports a year-on-year increase in the price of energy of more than 50 percent.

3.1 Implementation of the US survey

The US survey used all 13 treatments, was programmed in oTree (Chen et al., 2016), and was conducted on Prolific (`www.prolific.co`), a UK-based commercial subject pool.[2] On Prolific, we recruited a representative sample of the US population (stratified along sex, age and race). Data collection started on December 17th and finished on December 19th 2021.

In total, 1301 respondents completed our survey, with 100 respondents per treatment condition, except *Baseline*, which had 101 respondents. For the data analysis, we dropped 22 respondents: One failed the attention check, one appeared to reside outside the US and 20 provided beliefs in the density forecast that did not add up to 100 (see Table 1).[3] Respondents were paid a fixed amount of £1 (worth $1.33 at the time of the experiment) for completing the survey. On average, it took respondents 5:44 minutes to finish the survey. Based on our payment, respondents earned on average an hourly wage of $16.40, well above the average hourly earnings on Prolific.

3.2 Results of the US survey

Does the response scale affect the survey responses? Figure 4 gives a first impression. The figure shows the distribution of respondents' mean inflation expectations for all treatments. The *Shift* treatments are shown on the left, the *Compression* treatments in the center, and the *Centralization* treatments on the right. Moving the intervals of the scale to the left or right in *Shift* moves responses in the same direction. Similarly, compressing or expanding the scale in *Compression* also compresses and expands the answers.

Table 2 shows, for each of the 13 treatments, the average mean forecast, the average forecast uncertainty, and the disagreement of respondents. These statistics are calculated using a smoothed response ("beta") and a mass-at-midpoint measure ("m.a.m.").[4]

[2] See Appendix A for the instructions and Appendix B for screenshots of the three inflation questions.

[3] Respondents whose probabilities do not add up to 100 are prompted once to correct their answer. However, submitting an answer whose probabilities do not sum up to 100 was possible. For more detailed attrition and randomization checks, see Appendix D.

[4] Engelberg et al. (2009) suggest to smooth the responses (the histograms) by fitting a parametric distribution from which statistics such as mean, uncertainty, 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". This and the mass-at-midpoint procedure require 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 and when a respondent uses one or both of the open intervals,

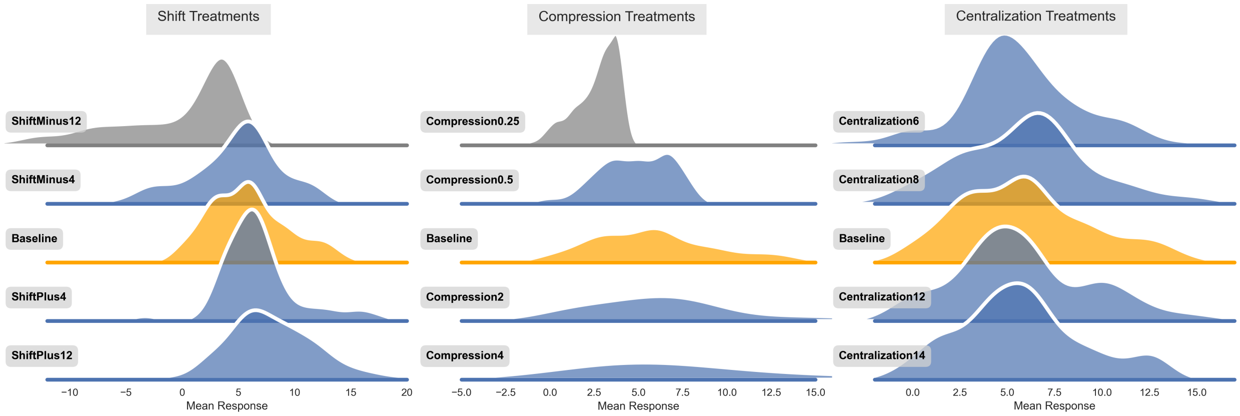


Figure 4: **Distribution of mean forecasts.** Kernel density estimates by treatment. *Shift* treatments in the left panel, *Compression* treatments in the center panel, and *Centralization* treatments in the right panel. Each panel uses a common y-axis with *Baseline* shown in orange in the center for comparison. Treatments with large probability mass in the open intervals in gray. Mean forecasts are calculated using a mass-at-midpoint assumption.

Treatment	**Statistics**									
	Mean Forecast				**Forecast Uncertainty**				**Disagree-ment**	
	beta		**m.a.m.**		**beta**		**m.a.m.**		**beta**	**m.a.m.**
Name	Avg.	P-value	Avg.	P-value	Avg.	P-value	Avg.	P-value		
Baseline	5.56		5.87		3.08		3.81		3.48	3.63
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
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
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
ShiftPlus12	8.15	(1) 0.000 ***	8.34	(1) 0.000 ***	3.54	(2) 0.035 **	4.07	(2) 0.181	4.42	4.46
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
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
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
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
Centralization14	5.50	(2) 0.906	5.76	(2) 0.831	3.05	(1) 0.457	3.68	(1) 0.325	3.28	3.38
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
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
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 for the US survey.** beta: Statistics based on a smoothed response (see footnote 4 for details). m.a.m: Statistics using mass-at-midpoint assumption. Uncertainty is the standard deviation of a respondent's density forecast and disagreement is the standard deviation of the (mass-at-midpoint) means of the density forecasts. The 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.

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 responses to the right. Shifting the scale to the left shifts the responses to the left. The effect is substantial: The shift amounts to −5.88/−6.40 (beta/mass-at-midpoint) percentage points for *ShiftMinus12* and −1.25/−1.23 for *ShiftMinus4*. In the other direction, we find 1.03/0.96 for *ShiftPlus4* and 2.59/2.47 for *ShiftPlus12*.

In the *Compression* treatments, the entire scale is compressed or expanded. We use forecast uncertainty (the standard deviation of a respondent's 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 *Baseline*, where the uncertainty is 3.08/3.81 (beta, mass at midpoint), uncertainty increases in wide treatments and decreases in narrow 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 find an indirect knock-on effect on the average mean forecast and on disagreement. When we compress the scale, the average mean forecast is closer to the center 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 find some support for Hypothesis 5 when looking at the uncertainty in treatments with a smaller number of intervals at the center of the scale in the *Centralization* treatments. The uncertainty is 3.44/4.06 in *Centralization8* and 4.33/4.89 in *Centralization6*, however, only the later is significantly different from *Baseline*. In *Centralization12* it is 3.33/3.97 and in *Centralization14* 3.05/3.68 by comparison.

The statistics used in the discussion so far (beta, mass-at-midpoint) have two weaknesses. First, they require us to make an assumption about the "width" of the open intervals, and second, they may mechanically skew some of the results.[5] In Sections 3.2.1 to 3.2.3 we test our hypotheses by comparing the probability masses the respondents assign to specific ranges of inflation (e.g., deflation). These tests avoid the two weaknesses and further strengthen our findings.

we follow Engelberg et al. (2009) and treat the limits of the beta distribution as parameters to be estimated. 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. Like Armantier et al. (2017), we allow the smooth responses to be bi-modal when respondents supply three or more intervals. See also the discussion about bi-modal responses in Section 5. We follow Becker et al. (2022) who extend the original procedure of Engelberg et al. (2009) to response scales with irregular spacing of the intervals.

[5] We discuss this point in detail in Appendix F.

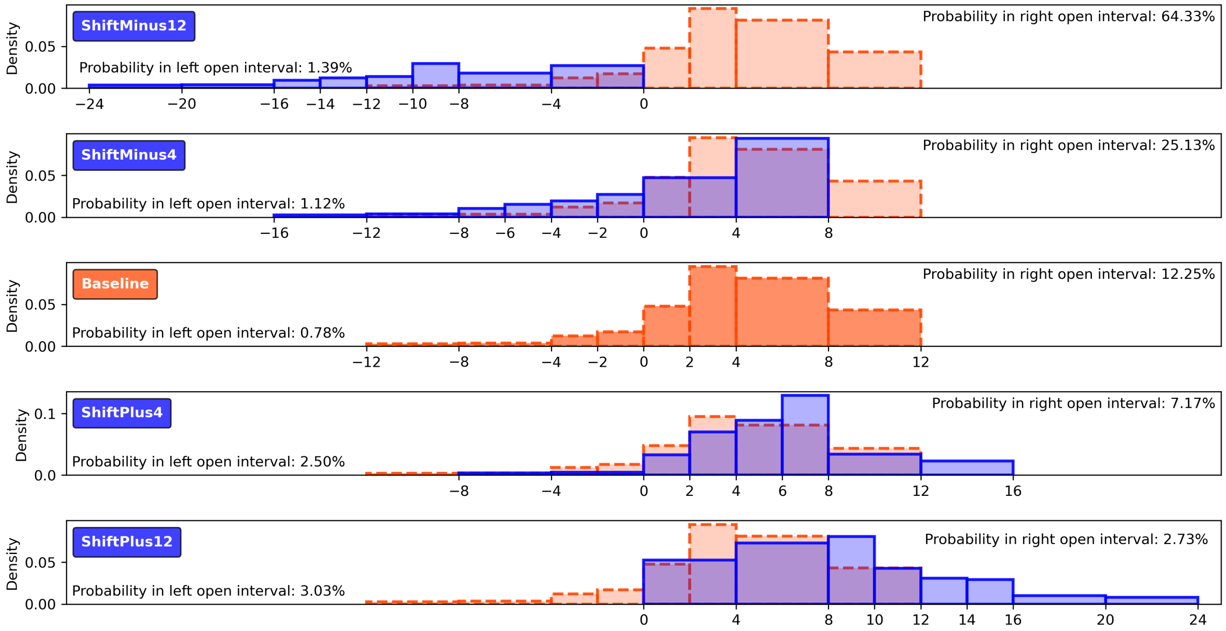


Figure 5: Histograms of the average densities assigned to the intervals in the **Shift treatments** (with Baseline in dashed bars for reference). Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

3.2.1 Shift treatments

Figure 5 depicts average densities assigned to each interval in the *Shift* treatments, relative to *Baseline*. As the histograms show, the probability mass over the entire scale shifts with the response scale.

This effect can be more clearly illustrated by focusing on one side of the scale. The average probability mass that respondents put into deflation, for example, decreases 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 Mann–Whitney–Wilcoxon (MWW) tests and t-tests confirm Hypothesis 1 for *ShiftMinus12*, *ShiftMinus4* and *ShiftPlus12*. For *ShiftPlus4* the relocation of probability mass goes in the hypothesized direction but is not significant at the 5% level.[6]

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

We do not find evidence that supports Hypothesis 2 and Hypothesis 3. The treatment interventions in the density forecast do not spill over to the binary inflation/deflation ques-

[6] See Table A3 in Appendix E for additional regressions supporting this pattern and Section 4 for significant results regarding treatment *ShiftPlus4* in the larger German survey.

Treatment	**Test Range**	**Probability Mass**			**Tests (p-values)**	
		Baseline	Treatment	Ratio	MWW	t-Test
ShiftMinus12	< 0	9.31	35.67	3.83	0.000 ***	0.000 ***
ShiftMinus4	< 0	9.31	18.35	1.97	0.014 **	0.002 ***
ShiftPlus4	< 0	9.31	5.87	0.63	0.230	0.051 *
ShiftPlus12	< 0	9.31	3.03	0.33	0.035 **	0.001 ***

Table 3: 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): MWW (Mann-Whitney-Wilcoxon two-sample statistic) tests, and t-tests (assuming unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

tion. When testing Hypothesis 2 for the *Shifting* treatments against *Baseline*, no treatment difference is significant at the 5 percent level.[7]

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

Similarly, when testing the point forecasts in the *Shifting* treatments versus *Baseline*, no treatment difference is significant at the 5% level.[8]

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

3.2.2 Compression treatments

Figure 6 shows the average densities assigned to each interval in the *Compression* treatments relative to *Baseline*. Compressing or expanding the scale has a strong effect on the responses and affects mean forecast, forecast uncertainty, and disagreement (see Table 2). We now test Hypothesis 4 via changes in the probability mass respondents assign to given ranges of inflation. Since compressing the scale moves interval boundaries, the treatment comparisons require different test ranges. As a rule, we use the largest overlapping range consisting of closed intervals.

[7] One-sided Fisher exact tests. One treatment difference is significant at the 10% level: *ShiftMinus4* versus *Baseline* ($p = 0.056$, obs.= 200). It should be noted that very few respondents expected deflation in the binary inflation/deflation question when we conducted the survey in December 2021. In *Baseline*, only a single respondent predicted prices to decline in the following 12 months. All other 100 participants expected prices to increase. In the other *Shift* treatments, the numbers are 95, 93, 94, and 94 (see Table A1 in the Appendix).

[8] T-tests: One treatment difference is significant at the 10% level: *ShiftMinus4* vs *Baseline* ($p = 0.085$, obs.= 200). When testing via Mann-Whitney-Wilcoxon (MWW) tests, no treatment difference is significant at the 10% level.

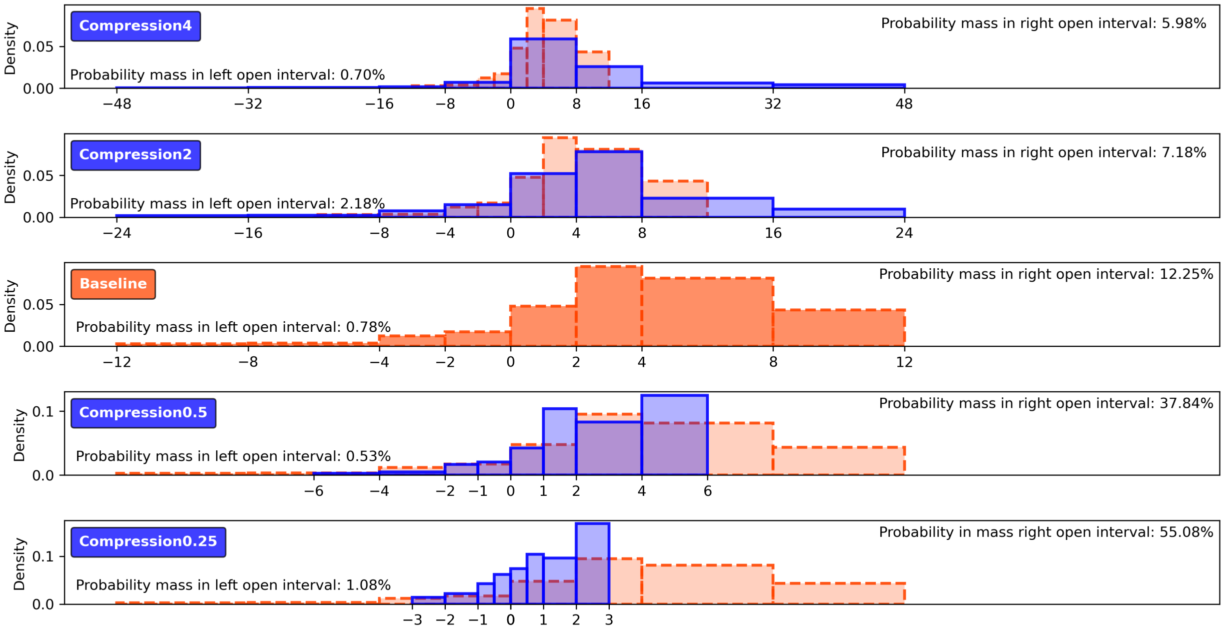


Figure 6: Histograms of average densities assigned to the intervals in the **Compression treatments** (with Baseline in dashed bars for reference). Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

Table 4 shows that compressing or expanding the scale significantly compresses and expands the stated responses in treatments *Compression4*, *Compression2*, and *Compression0.25*. When the scale is compressed, respondents move probability mass into intervals covering inflation rates close to zero. When the scale is expanded, respondents move probability mass away from intervals covering inflation rates close to zero.

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

Result 4 can be explained by a non-responsive use of intervals by the respondents. A

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

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

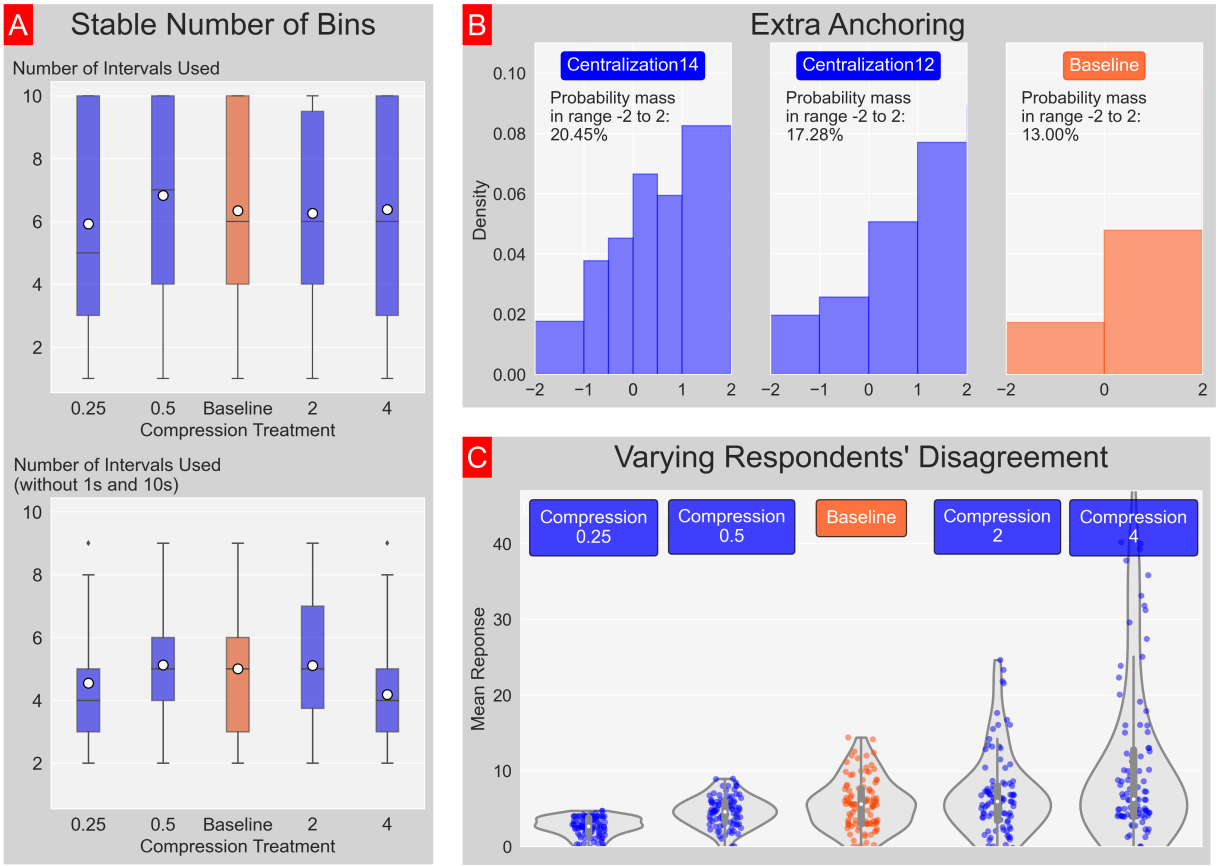


Figure 7: **Panel A.** Boxplots of the number of intervals used by the respondent in the *Compression* treatments, by treatment. Large bright circles indicate averages. Top: All data. Bottom: excluding respondents that use all ten intervals or only a single interval. **Panel B.** Average densities assigned to intervals in the range from -2 to 2 in *Centralization* treatments, by treatment, common axes. **Panel C.** Violin plots and scatterplots (jittered data) of respondents' mean forecasts (mass-at-midpoint) in the *Compression* treatments.

responsive participant who tries to accurately "copy" her subjective distribution of inflation expectations onto the response scale would use a different number of intervals in the different *Compression* treatments. As an example, consider a respondent who expects inflation to fall into the range from 0% to 8%. In *Compression4*, this respondent needs only a single interval to express her subjective beliefs. In *Compression2*, the respondent requires two intervals, and in *Baseline*, 3 intervals are needed. 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 A of Figure 7 show that the average number of intervals respondents use is around 6 and does not vary much between treatments. The upper part of the panel includes all data and the bottom part excludes respondents who either use a single interval or use all ten intervals for their answer. The

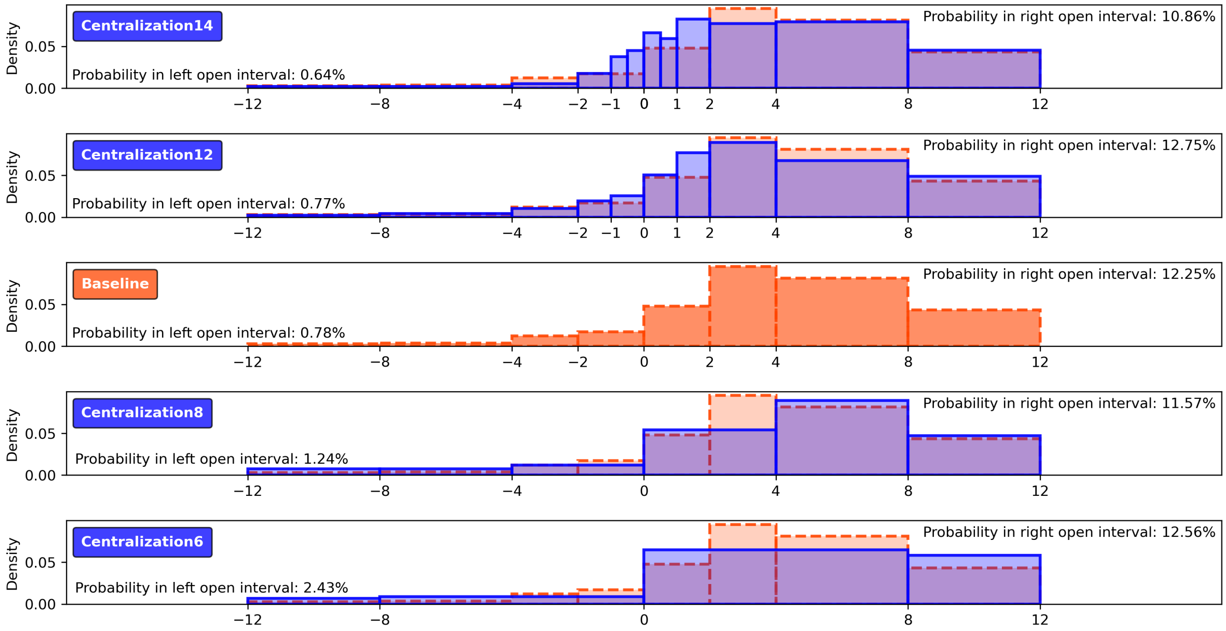


Figure 8: Histograms of average densities assigned to the intervals in the **Centralization treatments** (with Baseline in dashed bars for reference). Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

pattern is the same: Respondents tend to use roughly the same number of intervals, independent of the width of the scale. This non-responsive use of intervals may explain the strong treatment effect on respondents' uncertainty (Table 2) and may also explain why disagreement declines when we compress the scale (Panel C of Figure 7 and Table 2).[9]

3.2.3 Centralization treatments

Figure 8 shows the average densities assigned to each interval in the *Centralization* treatments. As before, we test Hypothesis 5 by comparing probability masses assigned to specific ranges of inflation. The rule we use to select these ranges is to take the smallest central range for which interval boundaries in *Baseline* coincide with the respective treatment boundaries. For *Centralization12* and *Centralization14* the range is from −2 to 2. For *Centralization8*, the range is from −4 to 4 and for *Centralization6* the range is from −8 to 8.

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 MWW tests indicate that for *Centralization14*, *Centralization8*, and *Centralization6*, these treatment differences

[9] Studying the Survey of Professional Forecasters (SPF), Glas and Hartmann (2022) find a related effect and report that forecasters do not automatically use twice as many intervals when the interval widths are cut in half. Instead, the forecasters only slightly increase the number of intervals inducing a noticeable drop in uncertainty.

Treatment	**Test Range**	**Probability Mass**			**Tests (p-values)**			
		Baseline	Treatment	Ratio	MWW		t-Test	
Centralization14	-2 to 2	13.00	20.45	1.57	0.029	**	0.006	***
Centralization12	-2 to 2	13.00	17.28	1.33	0.077	*	0.055	*
Centralization8	-4 to 4	34.48	26.56	0.77	0.042	**	0.028	**
Centralization6	-8 to 8	68.46	58.90	0.86	0.003	***	0.011	**

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

are significant at least at the 5% level. For *Centralization12*, they are significant at the 10% level.[10]

Result 5: *The probability mass assigned to a given range of inflation rates increases when the response scale provides more intervals in this range.*

The behavior we observe in the *Centralization* treatments has been described in the literature as unpacking bias or partition dependence, and is discussed in detail in Section 5. Result 5 highlights how the irregular layout of our response scale in *Baseline* (and in the SCE) allows the unpacking bias to reinforce the central tendency bias. The irregular layout moves probability mass towards the center of the scale, giving the impression that inflation expectations are anchored at low inflation rates. Panel B of Figure 7 illustrates this spurious "anchoring" for *Centralization14* and *Centralization12*.

3.2.4 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. We then examine whether the reported point forecasts fall into the bounds. The procedure does not impose specific distributional assumptions on the underlying densities.[11] Table A1 in Appendix C shows average point forecasts for all treatments.

For each respondent, we place the probability mass the respondent assigns to an interval at the interval's lower or upper limits. Doing this for each interval of the response scale and summing up, we obtain lower and upper bounds on a respondent's mean. If the point

[10] See Tables A4 to A6 in Appendix E for additional regression results supporting this pattern.

[11] Several papers study respondents' internal consistency by comparing point forecasts with measures of central tendency derived from the subjective probability distribution. For household surveys see Zhao (2022), Delavande and Rohwedder (2011), Bruine de Bruin et al. (2011), and for surveys of professionals see Engelberg et al. (2009) and Clements et al. (2023) among others.

Treatment	**Statistics**				
		Mean-consistent		Median-consistent	
	Observations	Share	P-value	Share	P-value
Baseline	101	0.62		0.63	
SCE (December 2021)	1283	0.58	0.403	0.62	0.915
ShiftMinus12	99	0.45	0.023**	0.52	0.115
ShiftMinus4	99	0.60	0.772	0.64	1.000
ShiftPlus4	98	0.55	0.316	0.53	0.153
ShiftPlus12	98	0.55	0.316	0.63	1.000
Compression4	99	0.73	0.133	0.73	0.174
Compression2	99	0.62	1.000	0.66	0.769
Compression0.5	96	0.58	0.662	0.61	0.883
Compression0.25	100	0.39	0.001***	0.51	0.088*
Centralization14	96	0.66	0.659	0.64	1.000
Centralization12	96	0.58	0.662	0.54	0.196
Centralization8	99	0.68	0.461	0.71	0.295
Centralization6	99	0.78	0.021**	0.75	0.094*

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.

forecast falls within those bounds, it is consistent with the mean. To construct the lower and upper bounds on the median, let $j \in \{1, 2, ..., N\}$ 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 point forecast must fall within the interval $[\theta_k, \theta_{k+1}]$, where k is determined by $\sum_{s=1}^{k} p_{is} \leq 0.5$ and $\sum_{s=1}^{k+1} p_{is} \geq 0.5$, to be consistent with the median.

As a reference, we also calculate the consistency measures for the SCE for the December 2021 wave. Table 6 shows the results of the consistency tests for all 13 treatments and for the SCE. Respondents in *Baseline* display the same consistency as the respondents of the SCE. When we compare the *Shift* treatments with *Baseline*, only *ShiftMinus12* shows a significant difference for the mean, though 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, 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 if the share of respondents with consistent answers grows in *Compression4* and *Centralization6*. In *Compression4*, for example, a single interval covers the entire range from

0% to 8%. Section 5 continues this discussion.

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

3.2.5 Impact of respondents' proficiency

After each question about inflation expectations, we ask subjects to state their subjective certainty for this answer. Respondents then complete a questionnaire with questions on financial literacy, highest obtained degree, knowledge of the Fed's inflation target. We refer to the collection of these measures as "proficiency". To evaluate whether respondents with higher proficiency are less affected by changes of the responses scale (Hypothesis 7), we use the inflation ranges established in Sections 3.2.1 to 3.2.3 and regress the probability mass assigned to the ranges on the proficiency variables, treatment dummies, interaction terms and several control variables. Specification (3) of Tables A3 to A6 in Appendix E reports the results.

The financial literacy interaction is never significant at the 5% level for any *Shift* or *Centralization* treatment. It is significant at the 1% level for *Compression4* and at the 5% level for *Compression0.5*. The interaction term for knowledge of the inflation target is never significant at the 1% level for any treatment and significant at the 5% level only for *Compression0.25*. Having high education leads to significant interactions at the 5% level only for the treatment *ShiftMinus12*. Overall, we find little evidence for Hypothesis 7.

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

According to Hypothesis 8, respondents with higher proficiency should be more certain in their answers. To test this, we regress respondents' certainty on the proficiency variables and other controls in Table A7 in Appendix E. Knowing the inflation target makes respondents more certain of their answer in all three forecasts (density forecast, point prediction, binary inflation/deflation forecast). However, for the point prediction, this becomes insignificant when controls are added. Instead, respondents become less certain here with higher financial literacy. Education never has a significant influence on subjective certainty. For all three questions, the higher a respondent's forecast, the higher their certainty. Women are always less certain, Republicans are always more certain.

Result 8: *Respondents who know the Federal Reserve Bank's inflation target are more certain in their forecasts. However, higher reported education or financial literacy do not increase respondents' certainty.*

4 The German survey

In addition to the data collected for the US via Prolific, we included two treatments, *ShiftPlus4* and *Centralization14*, in the Bundesbank Online Panel Households (BOP-HH) in June 2022. The BOP-HH closely follows the SCE in its design of the inflation density question, only the order of the intervals differs. The response scale of the BOP-HH starts with deflation whereas the response scale of the SCE starts with inflation (see question 1 in Appendix A).[12] Year-on-year CPI inflation in Germany was reported to be 7.1 percent in June 2022 and of similar magnitude in the preceding months (Bundesbank, 2022).

4.1 Implementation

In June 2022, 4460 German households participated in Wave 30 of the BOP-HH.[13] 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*.

4.2 Results of the German survey

In Table 7, we replicate the analysis of Table 2 for the German data. The predicted treatment differences go in the same direction as in the US data. 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 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 surprisingly similar to the US data results. The ratio of the 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 range from -2 to 2 is 1.54 times that of *Baseline* in Germany. In the US, this ratio is 1.57. Overall, despite running the treatments in a different country and at different times, we find the same direction of treatment effects and very similar effect sizes.

[12] For a technical description of the BOP-HH Survey see Beckmann and Schmidt (2020).

[13] In Becker et al. (2023) we use the panel structure of the survey and compare the June wave with the preceding and the subsequent waves.

Treatment	Statistics									
	Mean Forecast				**Forecast Uncertainty**				**Disagree-ment**	
	beta		**m.a.m.**		**beta**		**m.a.m.**		**beta**	**m.a.m.**
Name	Avg.	P-value	Avg.	P-value	Avg.	P-value	Avg.	P-value		
Baseline	6.63		6.72		2.12		2.18		4.01	4.06
ShiftPlus4	7.22	(1) 0.000 ***	7.28	(1) 0.000 ***	1.79	(2) 0.000 ***	1.89	(2) 0.000 ***	3.55	3.59
Centralization14	6.42	(2) 0.146	6.50	(2) 0.162	2.04	(1) 0.079 *	2.07	(1) 0.066 *	3.88	3.86

Table 7: **Treatment differences for the German survey.** beta: Statistics based on a smoothed response. See footnote 4 for details. m.a.m.: 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.

Treatment	**Test Range**	**Probability Mass**			**Tests (p-values)**			
		Baseline	Treatment	Ratio	MWW		t-Test	
ShiftPlus4	< 0	7.10	3.72	0.52	0.000	***	0.000	***
Centralization14	-2 to 2	5.68	8.78	1.54	0.007	***	0.000	***

Table 8: **Shift and centralization treatments in the German survey.** Tests for significant difference (one-sided) in average probability masses. MWW (Mann-Whitney-Wilcoxon two-sample statistic) and t-test (assumes unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

5 Discussion

One way to interpret the results of the two previous sections is to assume non-rational behavior via behavioral biases. An alternative interpretation that does not presuppose non-rationality can be found in Bayesian updating. In the first part of this section we describe the two interpretations in more detail. The second part outlines steps that could mitigate the problem, and discusses what to do in times of high inflation when respondents assign large probability masses to the open intervals. To keep the presentation simple, we refer to any discrepancy between a respondent's true or prior beliefs and the measured beliefs as "measurement bias".

5.1 Interpretation of the measurement bias

One way to interpret our results is to assume non-rational behavior. Instead of maintaining coherent probability distributions over future events (such as inflation) and following rational updating rules such as Bayes' law, respondents might follow simpler heuristics. Following this line of reasoning, the treatment differences we find in Results 1, 4 and 5 can be explained by behavioral biases that are known in the literature from other settings.

The central tendency bias (Hollingworth, 1910; Duffy et al., 2010) refers to respondents' propensity to prefer answers in the middle of the response scale. This could explain, as seen in Result 1, why respondents shift their reported probability distributions following a shift of the response scale. In Result 4 we observe that respondents tend to assign probability masses to the intervals without properly taking into account the compression of the scale. This is in line with support theory (Tversky and Koehler, 1994) and with partition dependence (Fox and Rottenstreich, 2003). Finally, we find that respondents tend to assign a larger amount of probability mass to a given range of inflation rates, the more intervals the scale uses to represent this range (Result 5). This is similar to behavior found in other studies where it is referred to as unpacking bias (Tversky and Koehler, 1994; Sonnemann et al., 2013). One piece of evidence favoring an explanation via behavioral biases is the lack of knock-on effects of the treatment intervention onto the binary inflation/deflation question and the point forecast (Results 2 and 3).

A second interpretation of our results is that the treatment differences described above are the result of a rational cognitive process in which respondents use two sources of information when providing an answer. The first source of information is the respondent's prior knowledge about future inflation, based on information about past or current inflation, possibly combined with information about the macro-economic environment and the central bank's policy. The second source of information is what is called *context* in the survey

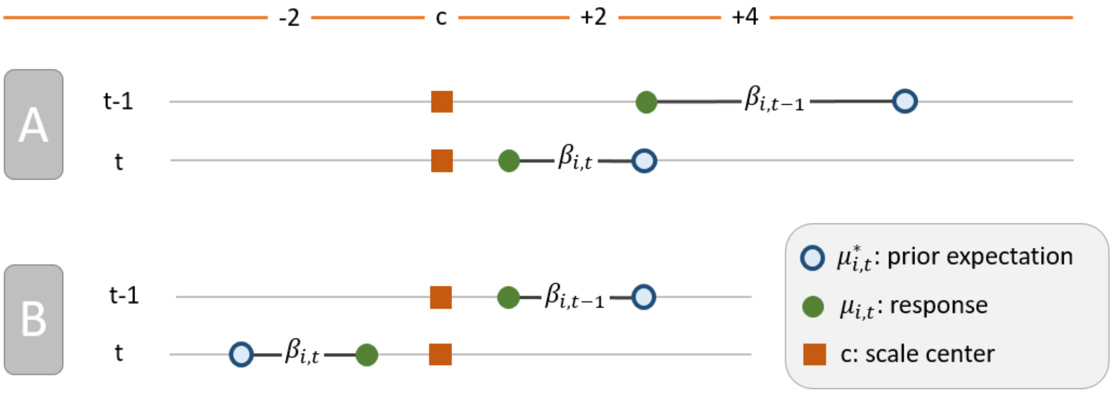


Figure 9: Illustration of an additive, time-varying measurement bias. In Panel B, $\beta_{i,t}$ reverses its sign from time $t-1$ to time t.

research literature (see Schuman, 1992; Schwarz, 2010). Context includes any information respondents obtain from participating in the survey. In the case of density questions, the response scale is an important part of the question's context. When asked about their inflation expectations, respondents may consider the response scale to reflect the surveyor's (i.e. the central bank's) own expectations. For example, 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.[14] Evidence favoring the rational updating interpretation comes from an asymmetry of the treatment differences in Result 1.[15] While a behavioral bias, if linear, would work similarly in both the *ShiftPlus* and *ShiftMinus* treatments, updating can explain an asymmetry via priors that are not centered on zero.

5.2 Mitigating the measurement bias

It is probably natural that taking time-differences comes to mind when looking for a way to mitigate the measurement bias. A measurement bias that is constant over time would cancel out when differencing. However, the assumption of a measurement bias that is constant over time seems improbable, as the example below illustrates.

Let $\mu_{i,t}$ be respondent i's mean forecast in period t and let $\mu^*_{i,t}$ be respondent i's prior (or

[14] In the US survey, we self-identified as researchers from the University of Heidelberg. The German survey was conducted by the Bundesbank. Apart from the response scale, respondents may also extrapolate information from the wording of a question (Schuman and Presser, 1996), the order of a question (Phillot and Rosenblatt-Wisch, 2018), or the affiliation of the surveying researcher (Schwarz, 2010).

[15] We use the fitted (beta) means of the *Shift* treatments to calculate the difference between individual means and the average of means in *Baseline*. Then we test whether the difference of *Baseline* and *ShiftPlus4* is different from the difference of *ShiftMinus4* and *Baseline* via t-tests (and similar for the *ShiftPlus12* and *ShiftMinus12* treatments). The differences for the +/-12 treatments are significantly different ($p \leq 0.001$, obs. = 197), but the differences for the +/-4 treatments are not ($p = 0.694$, obs. = 197).

true) inflation expectations in t. Assuming that the measurement bias ($\beta_{i,t}$) enters additively, we can express the change in inflation expectations as

$$(\mu_{i,t} - \mu_{i,t-1}) = \left(\mu^*_{i,t} - \mu^*_{i,t-1}\right) + (\beta_{i,t} - \beta_{i,t-1})\,.$$

To some extent, taking differences could alleviate the measurement bias, so that $(\beta_{i,t} - \beta_{i,t-1})$ is approximately zero. It seems possible, for example, that gender effects on $\beta_{i,t}$ are time invariant. But in general, $\beta_{i,t}$ is likely to vary with time because $\beta_{i,t}$ itself depends on the prior expectations $\mu^*_{i,t}$. Figure 9 illustrates this dependence. Result 1 shows that respondents are drawn towards the center of the scale, so the observed responses will typically lie between the prior expectations and the center of the scale. However, differencing mitigates the measurement bias only if the bias remains constant when the prior expectations vary over time. But when $\mu_{i,t}$ is bounded by the scale center, $\beta_{i,t}$ will decrease as $\mu^*_{i,t}$ approaches the scale center (Panel A). When the prior expectations happen to fall on the other side of the center (Panel B), the measurement bias will even change sign.[16] The assumption that the measurement bias is time-invariant is, therefore, not very convincing. Taking differences does not yield reliable estimates of a respondent's true changes in expectations.

As the measurement bias is in part introduced by the survey itself, a more promising approach to mitigate it is to modify the design of the question. One possible change is to use regularly-spaced response scales. Making the intervals narrow in some range is often motivated by the desire to give respondents the possibility to be more specific in some range while keeping the overall number of intervals reasonably small.[17] However, as the results in Sections 3 and 4 show, the narrow intervals attract additional probability masses, giving the spurious impression that values in the narrow intervals are expected more often.

The irregular spacing has other consequences. The first is that the consistency bounds are tighter when the intervals are narrow, making it more difficult for respondents to provide consistent responses (Zhao, 2022). In a survey with an irregularly spaced scale, such as *Baseline*, respondents expecting high inflation will then appear more consistent than respondents expecting low inflation.[18] A second consequence concerns the shape of the response. A response with a single mode (peak) is often a desirable property and, in fact, uni-modality is the "most basic assumption" in the parametric analysis of Engelberg et al. (2009, p. 36).

[16] The sign change of $\beta_{i,t}$ may open the possibility to identify bounds on the true expectations. We illustrate this idea in Appendix G.

[17] Here we are referring to an irregular spacing of the response scale in a range that respondents consider probable. There is a related problem with the two open intervals but the survey questions are typically designed in a way that keeps the probability mass in the open intervals small.

[18] For example, the bounds on the mean for a respondent in *Baseline* who expects inflation to fall into the range from 4 to 8 percent are twice as wide as the bounds for a respondent who expects inflation to fall into the range from 0 to 4.

Because of the irregular spacing, the subjective densities may be bi-modal even though the underlying probabilities are single peaked.[19]

A second promising design change is to give each respondent a personalized response scale centered on the respondent's point forecast.[20] This design minimizes the impact of the central tendency bias and reduces the need to provide very wide scales, rendering the irregular scales (introduced to achieve precise results near the assumed center of the distribution, while still allowing a broad range of expectations) unnecessary.

Such a design also eliminates the need to adjust the response scale in times of high inflation. These adjustments are necessary when respondents assign large probability masses in the open intervals. For example, the Survey of Professional Forecasters (SPF) of the Federal Reserve Bank of Philadelphia has regularly adjusted its response scale in the past decades by adding or removing intervals. The disadvantage of this approach is that any of these adjustments is likely to affect the responses. Responses from before and after the adjustment are, therefore, not directly comparable. A possible way to alleviate this problem could be to split the survey population and run two surveys (with the new and the old scale) in parallel for some time gathering data that could allow a chaining of the two series.[21]

6 Conclusion

In the past decade, several major central banks followed the New York Fed and started to elicit households' inflation expectations via density questions. An often cited advantage of density questions is that they allow us not only to quantify mean and median forecasts but also other variables that are valuable for central banking such as respondents' uncertainty or their perceived tail risk. Using the original question of the New York Fed as our baseline, the experiments in this paper provide a thorough test of how measured beliefs (the reported

[19] In the US survey, a large majority of the respondents supplies uni-modal responses (see Table A1 in Appendix C). But there are 112 (out of 1279) responses whose densities are bi-modal even though the bar-chart of the probabilities is uni-modal (the opposite occurs 22 times). As an example, consider respondent with id 659 who assigns single-peaked probabilities of 10, 15, 45, and 30 percent to the intervals 7 to 10 of the *Baseline* treatment. Since interval 7 is only half as wide as interval 8, the subjective histogram has two modes, and it is unclear whether this bi-modality is intentional.

[20] Dominitz and Manski (1997) use a scale determined by preliminary questions about subjective lowest and highest outcomes while studying household incomes, yet they warn against using these answers as minima and maxima of the scale. The Survey of Expectations of the Central Bank of the Republic of Turkey (CBRT) uses a regularly-spaced response scale centered on the respondents' point forecast, see e.g., Gülşen and Kara (2019). Crosetto and De Haan (2022) go a step further by letting respondents essentially construct their own scale via a click-and-drag interface.

[21] The SCE follows a different approach, using a comparatively wide response scale, and no change to the scale was considered necessary so far. Still, in March 2022 respondents assigned more than a fifth of the probability mass in the upper open interval (the average probability mass in the upper open interval between 2017 and 2019 is less than seven percent).

inflation forecasts) vary when we vary the response scale. The results show that shifting, compressing or expanding the scale leads to shifted, compressed and expanded forecasts. Beliefs measured using a density question systematically depend on the response scale. The resulting measurement bias is substantial, indicating that the quantitative nature of inflation density forecasts is deceptive. As such, inflation density forecasts can provide misleading information about how well respondents' expectations are anchored at a certain value. The measurement bias can vary over time so that even in differences, the forecasts are only suggestive.

However, the experiments also show that the measurement bias can be explained by well-known behavioral biases or even be rationalized. Understanding the underlying causes is a first step to control the measurement bias. Providing each respondent with a personalized response seems a promising way forward. Moreover, our experiments focused on households and it is possible that firms and especially professional forecasters are less affected by the behavioral biases than households, but it would be good to see more research in this direction.

References

Armantier, O., G. Topa, W. Van der Klaauw, and B. Zafar (2017). An overview of the survey of consumer expectations. *Economic Policy Review* (23-2), 51–72.

Becker, C., P. Duersch, T. A. Eife, and A. Glas (2022). Extending the procedure of engelberg et al.(2009) to surveys with varying interval-widths. *AWI Discussion Paper No. 707*.

Becker, C., P. Duersch, T. A. Eife, and A. Glas (2023). Households' probabilistic inflation expectations in high-inflation regimes. *FAU Discussion Papers in Economics 1_2023*.

Beckmann, E. and T. Schmidt (2020). Bundesbank online pilot survey on consumer expectations. Technical Paper.

Bruine de Bruin, W., A. Chin, J. Dominitz, and W. van der Klaauw (2023). Household surveys and probabilistic questions. In *Handbook of Economic Expectations*, pp. 3–31. Elsevier.

Bruine de Bruin, W., C. F. Manski, G. Topa, and W. Van Der Klaauw (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics 26*(3), 454–478.

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), 749–757.

Bundesbank (2022). Monthly report. *Deutsche Bundesbank* (06/2022).

Chen, D. L., M. Schonger, and C. Wickens (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Clements, M. P., R. W. Rich, and J. S. Tracy (2023). Surveys of professionals. In *Handbook of Economic Expectations*, pp. 71–106. Elsevier.

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*, 103297.

Coibion, O., Y. Gorodnichenko, and M. Weber (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy 130*(6).

Crosetto, P. and T. De Haan (2022). Comparing input interfaces to elicit belief distributions. *GAEL Working Paper* (01/2022).

Delavande, A. and S. Rohwedder (2011). Individuals' uncertainty about future social security benefits and portfolio choice. *Journal of Applied Econometrics 26*(3), 498–519.

Dominitz, J. and C. F. Manski (1997). Using expectations data to study subjective income expectations. *Journal of the American Statistical Association 92*(439), 855–867.

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

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

Fox, C. R. and Y. Rottenstreich (2003). Partition priming in judgment under uncertainty. *Psychological Science 14*(3), 195–200.

Glas, A. and M. Hartmann (2022). Uncertainty measures from partially rounded probabilistic forecast surveys. *Quantitative Economics 13*(3), 979–1022.

Gülşen, E. and H. Kara (2019). Measuring inflation uncertainty in turkey. *Central Bank Review 19*(2), 33–43.

Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods 7*(17), 461–469.

Lusardi, A. and O. S. Mitchell (2014). The economic importance of financial literacy: Theory and evidence. *Journal of economic literature 52*(1), 5–44.

Manski, C. F. (2018). Survey measurement of probabilistic macroeconomic expectations: Progress and promise. *NBER Macroeconomics Annual 32*, 411–471.

Payne, S. L. (Ed.) (1951). *The Art of Asking Questions.* Princeton: Princeton University Press.

Phillot, M. and R. Rosenblatt-Wisch (2018). Inflation expectations: The effect of question ordering on forecast inconsistencies. Technical report, Swiss National Bank.

Schuman, H. (1992). Context effects: State of the past/state of the art. In N. Schwarz and S. Sudman (Eds.), *Context Effects in Social and Psychological Research*, pp. 5–20. New York, NY: Springer New York.

Schuman, H. and S. Presser (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* Sage.

Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, and M. Viswanathan (Eds.), *The SAGE Handbook of Measurement*, pp. 43–61. Los Angeles: Sage.

Schwarz, N., H.-J. Hippler, B. Deutsch, and F. Strack (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly 49*(3), 388–395.

Sonnemann, U., C. F. Camerer, C. R. Fox, and T. Langer (2013). How psychological framing affects economic market prices in the lab and field. *PNAS 110*(29), 11779–11784.

Sudman, S. and N. Bradburn (Eds.) (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

Tversky, A. and D. J. Koehler (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review 101*(4), 547–567.

Zhao, Y. (2022). Internal consistency of household inflation expectations: Point forecasts vs. density forecasts. *International Journal of Forecasting*.

Appendix A Instructions

Annotations in italics represent comments on formatting/coding.

Introduction

Welcome!

You will take part in an academic survey conducted by the University of Heidelberg, Germany. We are interested in your personal views regarding the future inflation rate: it is therefore important that you answer honestly and read the questions very carefully before answering. This survey should take (on average) less than **8 minutes** to complete. For completing this survey, you will receive a fixed payment of **£1.00 (approximately $*[current value in US dollar]*)**.

Participation in this survey is entirely voluntary and you will remain anonymous throughout the survey. Results may include summary data, but you will never be identified. By continuing, you consent to the publication of survey results. Note that you cannot save and come back later to answer the survey. If you have any questions regarding this survey, you may contact us at **survey2021@awi.uni-heidelberg.de**.

If you understand and agree to the above information, please check "I consent, begin survey" below and click "Next" to begin. Otherwise, check "I do not consent" below and click "Next" to not take part in the survey.

O I consent, begin study

O I do not consent

[Next]

Instructions

We want to learn about your current outlook for future inflation in the United States. To do so, we will ask you a couple of questions. We are interested in your views and opinions. Your responses are confidential, and it helps us a great deal if you respond as carefully as possible. If you should come to any question that you can't or don't want to answer, just click on Next until the next question appears.

In some of the following questions, we will ask you to think about the percent chance of something happening in the future. Your answers can range from 0 to 100, where 0 means there is absolutely no chance, and 100 means that it is absolutely certain.

Thank you for your participation!

[Next]

Question 1

We would like you to think about the 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	□	**percent chance**

[Next]

Notes:

1. *Bin labels shown here are taken from the Baseline condition.*
2. *Page includes a running total that is updated as soon as a participant enters a value into one of the bins.*

Error messages for this page:

1. *Upon submitting an empty forecast (total of 0 percent chance):*
   Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click ***Next*** *to continue.*
2. *Upon submitting a forecast with a total of less than 100 percent:*
   Your total adds up to [percent sum]%. Please change the numbers in the table so they add up to 100%. Otherwise click ***Next*** *to continue.*

Question 2

How certain do you feel about your response to the previous question?

O Very Certain
O Certain
O Somewhat Certain
O Somewhat Uncertain
O Uncertain
O Very Uncertain

[Next]

Question 3

Over the next 12 months, do you think that there will be inflation or deflation? (Note: deflation is the opposite of inflation)

Please choose one.

O Inflation
O Deflation (the opposite of inflation)

[Next]

Error messages for this page:

1. *Upon trying to submit the page without providing an answer:*
   Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click ***Next*** *to continue.*

Question 4

How certain do you feel about your response to the previous question?
O Very Certain
O Certain
O Somewhat Certain
O Somewhat Uncertain
O Uncertain
O Very Uncertain
[Next]

Question 5

What do you expect the rate of *[inflation/deflation]* to be **over the next 12 months**? Please give your best guess.

Over the next 12 months, I expect the rate of *[inflation/deflation]* to be □ %

[Next]

Notes:

1. *Whether inflation or deflation is shown in the text depends on the participants' answer to Question 3. If participants answered in Question 3 that they expect inflation over the next 12 months, inflation was shown in Question 5 and deflation otherwise.*

Error messages for this page:

1. *Upon trying to submit the page without providing an answer:*
   Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click ***Next*** *to continue.*

Question 6

How certain do you feel about your response to the previous question?
O Very Certain
O Certain
O Somewhat Certain
O Somewhat Uncertain
O Uncertain
O Very Uncertain
[Next]

Questionnaire

To conclude the survey, we would like to ask you some questions about you and your household.

Age (leave blank if you prefer not to tell): □

Gender:
O Prefer not to answer
O Female
O Male
O Other

Highest educational degree:
O Prefer not to answer
O High school diploma
O Some college no degree
O Associate's degree occupational
O Associate's degree academic
O Bachelor's degree
O Master's degree
O Professional degree
O Doctoral degree

Please select "Squirrel". This question just helps us to screen out random clicking:
O Prefer not to answer
O Elephant
O Capybara
O Wolf
O Squirrel
O Mouse

The US Federal Reserve System (Fed) tries to control the inflation rate by keeping it close to a specific target value. What do you think is this target for the inflation rate?
O Prefer not to answer
O Positive inflation that averages 2% over time
O Negative inflation that averages -2% over time
O Positive inflation that averages 1% over time
O On average zero inflation over time
O Don't know

Your political orientation:
O Prefer not to answer
O Republican
O Democrat
O Independent
O Other

State of residence: □ *drop-down menu with list of states*

Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

O Prefer not to answer
O More than $102
O Exactly $102
O Less than $102
O Don't know

Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, with the money in this account, would you be able to buy...
O Prefer not to answer
O More than today
O Exactly the same as today
O Less than today
O Don't know

Do you think the following statement is true or false?
"Buying a single company stock usually provides a safer return than a stock mutual fund."
O Prefer not to answer
O True
O False
O Don't know
[Next]

Error messages for this page:

1. *When not submitting an answer to one of the questions:*
   Your answers are important to us. Please provide an answer or select "Refuse to answer".

End page

Thank you for your participation!

If you have any questions regarding this survey, you may contact us at **survey2021@awi.uni-heidelberg.de**.

Click here to confirm your participation and to return to Prolific. *[Sentence is hyperlink]*

No consent given page

As you do not wish to participate in this study, please return your submission on Prolific by selecting the 'Stop without completing' button.

If you have any questions regarding this study, you may contact us at **survey2021@awi.uni-heidelberg.de**.

You can close this window now.

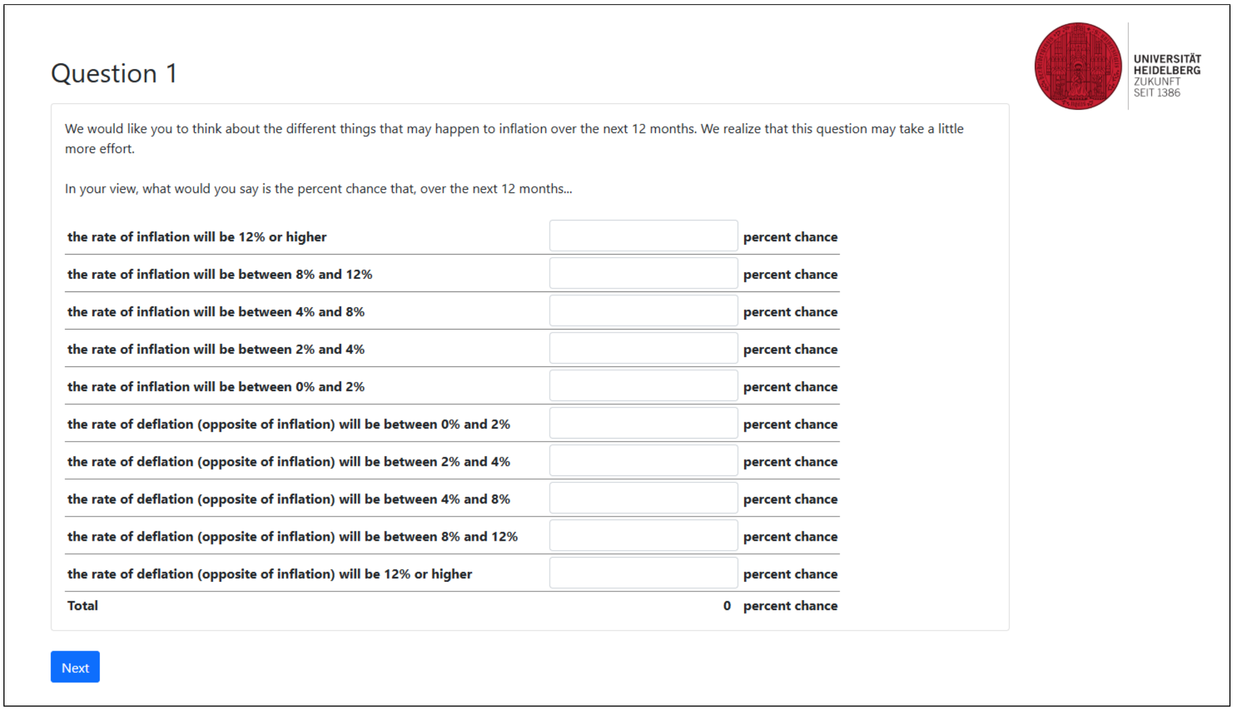
Timeout page

You did not complete the page in time. Thus you cannot finish this assignment.

If you have any questions regarding this study, you may contact us at **survey2021@awi.uni-heidelberg.de**.

You can close this window now.

Appendix B Screenshots



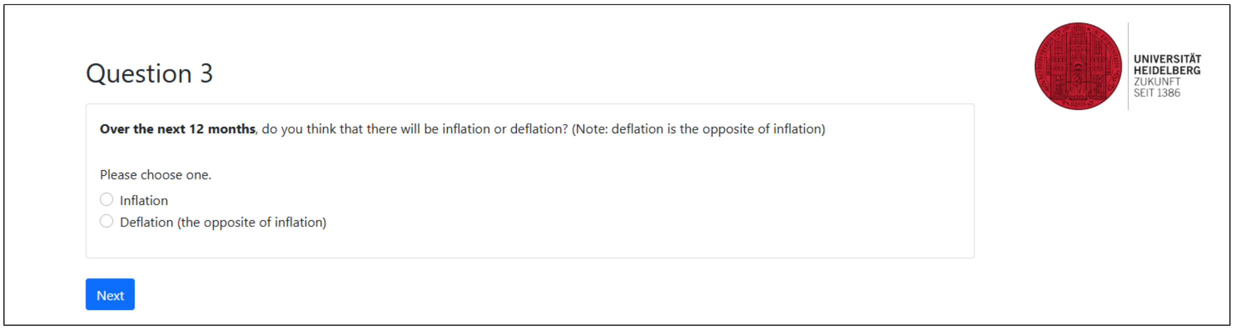
Question 1

We would like you to think about the 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	**0**	**percent chance**

Figure A1: Screenshot of the density question for the Baseline condition.



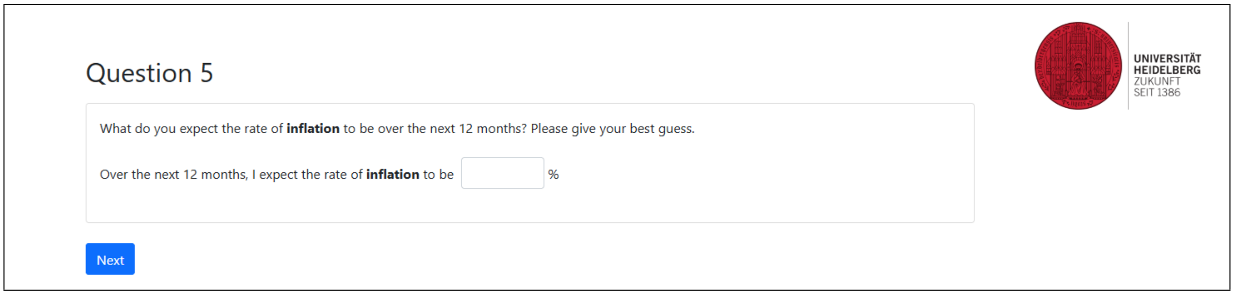
Question 3

Over the next 12 months, do you think that there will be inflation or deflation? (Note: deflation is the opposite of inflation)

Please choose one.

○ Inflation
○ Deflation (the opposite of inflation)

Figure A2: Screenshot of the inflation/deflation question.



Question 5

What do you expect the rate of **inflation** to be over the next 12 months? Please give your best guess.

Over the next 12 months, I expect the rate of **inflation** to be ______ %

Figure A3: Screenshot of the point forecast.

Appendix C Descriptive statistics

Treatment					Results											
		Response scale			Point forecast					Responses with				Uni-modality		
	obs	#	Center	Span	Mean	Trimmed mean	Median	Mean forecast	Intervals used	single interval	full set	open intervals	gaps	Probability	Density	Expecting inflation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Baseline	101	10	0	24	8.30	6.56	6.00	5.56	6.35	1	28	84	5	87	82	100
ShiftMinus12	99	10	-12	24	7.52	6.01	5.00	-0.32	5.28	19	27	90	3	85	71	95
ShiftMinus4	99	10	-4	24	5.88	6.01	6.00	4.31	6.20	2	37	98	4	73	70	93
ShiftPlus4	98	10	4	24	7.50	6.88	7.00	6.59	7.10	1	32	71	6	74	69	94
ShiftPlus12	98	10	12	24	8.72	7.38	7.00	8.16	7.51	2	25	74	0	90	86	94
Compression4	99	10	0	96	11.83	9.59	7.00	10.98	6.38	1	38	60	1	86	83	99
Compression2	99	10	0	48	9.39	8.04	7.00	6.23	6.26	2	25	63	1	75	78	96
Compression0.5	96	10	0	12	6.50	5.19	5.00	4.55	6.82	3	36	94	3	79	64	92
Compression0.25	100	10	0	6	5.66	4.49	4.40	2.61	5.92	12	33	100	4	78	60	97
Centralization14	96	14	0	24	6.16	5.93	6.00	5.50	8.64	0	28	76	9	69	57	90
Centralization12	96	12	0	24	9.22	6.77	6.00	5.57	7.58	1	27	77	8	71	65	94
Centralization8	99	8	0	24	7.53	6.66	7.00	5.37	5.51	0	34	85	3	80	81	95
Centralization6	99	6	0	24	6.00	6.70	7.00	5.48	4.27	1	40	79	0	91	82	91
Average	98.38				7.71	6.63	6.00	5.43	6.44	3.46	31.54	80.85	3.62	79.85	72.92	94.62

Table A1: **Descriptive statistics by treatment (US survey).** Number of respondents in column 1 (obs). Columns 2 to 4 give information about the response scale: Number of intervals (#), center of the response scale, span of the closed intervals. Columns 5 to 7 give mean, trimmed mean (trimmed at 10 percent), and median response for the point forecast (Q3). Columns 8 and 9 show average mean forecasts (beta), and the average number of intervals used by the respondents. Columns 10 to 13 give information about response attitudes: The number of respondents using a single interval (10), using the full set of intervals (11) which is generally 10 but varies in the *Centralization* treatments, using any of the two open intervals (12), and providing responses with gaps (13). Columns 14 and 15 report the number of responses with uni-modal (single-peaked) response: uni-modal bar-chart of probabilities (14), uni-modal histogram (15). Column (16) reports the number of respondents expecting inflation in the binary inflation/deflation question.

Appendix D Attrition and Randomization checks for the US survey

Table A2 shows that in the US survey, respondents are equally likely to drop out of any of the treatments. Columns 1 and 2 show information on the number of all surveys started, while columns 3 and 4 contain information on all surveys that were actually finished by respondents. Columns 5 and 6 depict the number of complete surveys, that is all finished surveys minus those respondents i) that did not pass the attention check (1 respondent), ii) that were living outside the US (1 respondent), and iii) that provided answers to the density forecast that do not add up to 100 (20 respondents). Attrition overall was very low.

	All surveys		Finished surveys		Complete surveys	
Treatment	Participants	Share	Participants	Share	Participants	Share
Baseline	105	7.74	101	7.76	101	7.90
ShiftMinus12	102	7.52	100	7.69	99	7.74
ShiftMinus4	107	7.89	100	7.69	99	7.74
ShiftPlus4	106	7.82	100	7.69	98	7.66
ShiftPlus12	103	7.60	100	7.69	98	7.66
Compression4	103	7.60	100	7.69	99	7.74
Compression2	102	7.52	100	7.69	99	7.74
Compression0.5	103	7.60	100	7.69	96	7.51
Compression0.25	105	7.74	100	7.69	100	7.82
Centralization14	107	7.89	100	7.69	96	7.51
Centralization12	104	7.67	100	7.69	96	7.51
Centralization8	106	7.82	100	7.69	99	7.74
Centralization6	103	7.60	100	7.69	99	7.74
Sum	1356	100	1301	100	1279	100

Table A2: **Attrition check.** Number of participants that started, finished, and completed the surveys (by treatment).

Additionally we compare the demographics shown in Table 1 against data from the US Census Bureau from 2021. We compare the mean age against the census mean age, accounting for the fact that respondents on Prolific has to be of age 18 or older. We also compare the shares for respondents identifying as female, black, or white against the shares reported in US census data. We find that our overall sample is representative of the US population in terms of the share of female, black, and white respondents. However, or sample is on average around 2 years younger than the average US citizen (diff$=-2.06$, $p < 0.001$, t-test). In terms of the individual treatments, we find differences in age at the 5% level for ShiftPlus4, ShiftPlus12, Compression2, and Centralization12 , as well as the 10% level for Centralization6 (t-tests). For share of females, we find differences at the 5% level for Compression4 and Compression0.25, and at the 10% level for ShiftMinus4 (proportions z-tests). For white respondents, we find differences at the 5% level for Compression2 and Centralization8, and at the 10% level for ShiftPlus4 (proportions z-tests). The share of black respondents is only significant at the 5% level in Compression4 (proportions z-tests). Overall, the one treatment showing strong differences is Compression4, with a significantly higher share of females (0.61), significantly higher share of white (0.84), and significantly lower share of black respondents (0.05). Looking more closely at this treatment, this is caused by Compression4 having a substantially larger number of white female respondents (52 out of 99 respondents). Note that if we control for multiple hypotheses testing using the Bonferroni-Holm method, only the aforementioned age difference for the overall sample would remain (strongly) significant.

Appendix E Respondents' proficiency regressions

Tables A3 to A6 in this appendix report regressions of the probability mass respondents' assign to certain ranges of inflation on treatment dummies, interaction terms with respondents' proficiency, and other controls. A respondent's proficiency refers to one of the following three measures: financial literacy, highest obtained degree, or knowledge of the Fed's inflation target. Financial literacy was elicited in a questionnaire at the end of the experiment. We used the three-item financial literacy test by Lusardi and Mitchell (2014), *Financial lit.*, ranging from 0 to 3 correct answers. In addition, we asked respondents 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).

As outlined in Sections 3.2.1 to 3.2.3, the different treatments require different test ranges: 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 $[-2, 2]$.

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. The regressions use age, gender, and political orientation. In the survey, we also elicited the state of residence from respondents and use this information to create region dummy variables based on the definition of the US Census Bureau (West, Midwest, South, Northeast, Territories). The regressions use these dummies to control for region of residence.

Probability Mass in Deflation	(1)		(2)		(3)	
ShiftMinus12	26.36***	(0.000)	27.58***	(0.000)	30.06***	(0.005)
ShiftMinus4	9.047***	(0.007)	9.588***	(0.003)	28.93***	(0.003)
ShiftPlus4	-3.440	(0.301)	-5.024	(0.110)	-5.779	(0.586)
ShiftPlus12	-6.276*	(0.059)	-8.427***	(0.008)	-24.97***	(0.009)
Certain			-5.707***	(0.000)	-5.467***	(0.000)
Financial lit.					-5.148**	(0.046)
Financial lit. × ShiftMinus12					5.140	(0.245)
Financial lit. × ShiftMinus4					-6.029	(0.128)
Financial lit. × ShiftPlus4					1.498	(0.733)
Financial lit. × ShiftPlus12					6.980*	(0.055)
Target correct=1					-4.779	(0.281)
Target correct=1 × ShiftMinus12					-8.278	(0.203)
Target correct=1 × ShiftMinus4					6.139	(0.350)
Target correct=1 × ShiftPlus4					4.218	(0.507)
Target correct=1 × ShiftPlus12					2.120	(0.738)
Education high=1					0.0854	(0.985)
Education high=1 × ShiftMinus12					-15.22**	(0.022)
Education high=1 × ShiftMinus4					-11.08*	(0.092)
Education high=1 × ShiftPlus4					-7.049	(0.287)
Education high=1 × ShiftPlus12					-2.042	(0.752)
Constant	9.307***	(0.000)	30.35***	(0.000)	42.41***	(0.000)
Controls	No		Yes		Yes	
Observations	495		494		492	
Adjusted R^2	0.204		0.294		0.372	

Table A3: **Shift Treatments.** OLS regressions of the probability mass assigned to deflation on respondents' proficiency and interactions. 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)	-12.54***	(0.003)	-41.97***	(0.000)
Compression2	-7.223*	(0.095)	-4.621	(0.270)	-16.40	(0.181)
Centralization6	-9.556**	(0.027)	-8.460**	(0.041)	-25.14*	(0.056)
Certain			-1.499	(0.217)	-1.773	(0.141)
Financial lit.					-2.078	(0.548)
Financial lit. × Compression4					13.74***	(0.004)
Financial lit. × Compression2					5.364	(0.270)
Financial lit. × Centralization6					8.766*	(0.088)
Target correct=1					5.070	(0.397)
Target correct=1 × Compression4					-5.534	(0.508)
Target correct=1 × Compression2					-12.56	(0.145)
Target correct=1 × Centralization6					-3.828	(0.646)
Education high=1					8.162	(0.192)
Education high=1 × Compression4					-1.379	(0.875)
Education high=1 × Compression2					10.21	(0.246)
Education high=1 × Centralization6					-3.316	(0.699)
Constant	68.46***	(0.000)	66.22***	(0.000)	63.88***	(0.000)
Controls	No		Yes		Yes	
Observations	398		398		398	
Adjusted R^2	0.026		0.113		0.161	

Table A4: **Compression4, Compression2, Centralization6 Treatments.** OLS regressions of the probability mass assigned to the intervals in the range $[-8, 8]$ on respondents' proficiency and interactions. 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)	3.324	(0.421)	-16.43	(0.212)
Centralization8	-7.920*	(0.058)	-6.462	(0.117)	-19.29	(0.121)
Certain			-4.498***	(0.004)	-4.997***	(0.001)
Financial lit.					-5.262	(0.134)
Financial lit. × Compression0.5					11.10**	(0.027)
Financial lit. × Centralization8					6.669	(0.188)
Target correct=1					-0.792	(0.895)
Target correct=1 × Compression0.5 1					-7.927	(0.352)
Target correct=1 × Centralization8					15.42*	(0.070)
Education high=1					3.975	(0.527)
Education high=1 × Compression0.5					-4.207	(0.629)
Education high=1 × Centralization8					-17.39*	(0.051)
Constant	34.48***	(0.000)	50.24***	(0.000)	59.92***	(0.000)
Controls	No		Yes		Yes	
Observations	296		296		295	
Adjusted R^2	0.013		0.058		0.080	

Table A5: **Compression0.5, Centralization8 Treatments.** OLS regressions of the probability mass assigned to the intervals in the range $[-4, 4]$ on respondents' proficiency and interactions. 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)	14.63***	(0.000)	18.86**	(0.044)
Centralization14	7.448**	(0.017)	6.915**	(0.017)	6.723	(0.471)
Centralization12	4.281	(0.168)	3.208	(0.269)	-2.418	(0.784)
Certain			-5.088***	(0.000)	-5.073***	(0.000)
Financial lit.					-4.371*	(0.072)
Financial lit. × Compression0.25					0.584	(0.877)
Financial lit. × Centralization14					1.909	(0.592)
Financial lit. × Centralization12					4.353	(0.214)
Target correct=1					2.467	(0.555)
Target correct=1 × Compression0.25					-14.17**	(0.016)
Target correct=1 × Centralization14					-10.01*	(0.099)
Target correct=1 × Centralization12					-5.742	(0.329)
Education high=1					-1.706	(0.696)
Education high=1 × Compression0.25					2.004	(0.736)
Education high=1 × Centralization14					2.255	(0.713)
Education high=1 × Centralization12					-3.307	(0.586)
Constant	13.00***	(0.000)	33.70***	(0.000)	42.71***	(0.000)
Controls	No		Yes		Yes	
Observations	393		392		392	
Adjusted R^2	0.040		0.183		0.204	

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

Table A7 shows the results of regressions of the responses on the three certainty questions on respondents' proficiency (i.e., *Financial lit.*, *Target correct*, and *Education high*), the corresponding forecast, as well as controls for age, gender, political orientation, and region. For specifications (1)-(3), the *Certain* variable on the left side of the regression equation refers to respondents' certainty answer after the density forecast, for specifications (4)-(6), it is the certainty answer after the binary inflation/deflation forecast, and for specifications (7)-(9), it is the certainty answer after the point prediction. *Forecast* is the respondent's forecasts preceding the certainty question: 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).

	Density forecast			**Binary forecast**			**Point forecast**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Target correct	0.381***	0.301***	0.290***	0.290***	0.237***	0.227***	0.138**	0.0914	0.0996
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.029)	(0.138)	(0.107)
Financial lit.	0.0710	-0.0407	-0.0372	0.0508	0.0387	0.0410	-0.0482	-0.109***	-0.109***
	(0.117)	(0.375)	(0.410)	(0.203)	(0.347)	(0.315)	(0.222)	(0.007)	(0.008)
Education high	0.0299	0.0642	0.0587	0.0155	0.0490	0.0532	-0.0408	-0.00562	-0.00181
	(0.683)	(0.365)	(0.398)	(0.810)	(0.441)	(0.398)	(0.524)	(0.928)	(0.977)
Mean forecast		0.0269***	0.0455***		0.0316***	0.0424***		0.0213***	0.0280***
		(0.000)	(0.000)		(0.000)	(0.000)		(0.000)	(0.000)
Female		-0.509***	-0.527***		-0.208***	-0.215***		-0.304***	-0.317***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.000)	(0.000)
Age		0.0108***	0.0105***		0.0000286	0.000156		0.00727***	0.00713***
		(0.000)	(0.000)		(0.988)	(0.936)		(0.000)	(0.000)
Democrat		-0.108	-0.0844		-0.0605	-0.0567		-0.0808	-0.0767
		(0.160)	(0.265)		(0.381)	(0.408)		(0.233)	(0.258)
Republican		0.296***	0.284***		0.262***	0.256***		0.315***	0.304***
		(0.003)	(0.004)		(0.003)	(0.004)		(0.000)	(0.000)
Living in region south		0.0590	0.0388		0.163**	0.166**		0.167**	0.149*
		(0.502)	(0.655)		(0.039)	(0.035)		(0.031)	(0.055)
Living in region west		-0.0173	-0.0807		-0.0660	-0.0823		-0.0808	-0.111
		(0.875)	(0.456)		(0.503)	(0.402)		(0.403)	(0.252)
Living in region midwest		-0.0245	-0.0354		0.0664	0.0953		0.0274	0.0222
		(0.805)	(0.719)		(0.457)	(0.285)		(0.754)	(0.801)
Constant	2.442***	2.344***	2.315***	3.523***	3.408***	3.648***	2.817***	2.601***	2.660***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Treatment dummies	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
R^2	0.029	0.115	0.159	0.021	0.075	0.107	0.005	0.076	0.092
Adjusted R^2	0.026	0.107	0.143	0.019	0.067	0.091	0.002	0.068	0.076
Observations	1275	1274	1274	1276	1275	1275	1273	1272	1272

Table A7: **Certainty.** OLS regressions of respondents' certainty on her proficiency, her corresponding forecast, and controls, for each of the three certainty questions. p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Appendix F Mechanical treatment effects

Distribution	$\mathcal{N}(0,4)$		$\mathcal{N}(0,9)$		$\mathcal{N}(4,4)$		$\mathcal{N}(4,9)$		$\mathcal{N}(8,4)$		$\mathcal{N}(8,9)$	
	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty
Theoretical	0.00	2.00	0.00	3.00	4.00	2.00	4.00	3.00	8.00	2.00	8.00	3.00
Baseline	0.00	2.18	0.00	3.24	4.23	2.20	4.14	3.17	8.07	2.42	8.22	3.47
ShiftMinus12	0.91	3.15	0.62	3.57	3.86	0.89	3.44	1.79	4.00	0.03	3.98	0.37
ShiftMinus4	0.23	2.20	0.14	3.17	4.07	2.42	4.22	3.47	8.91	3.15	8.62	3.57
ShiftPlus4	-0.23	2.20	-0.14	3.17	4.00	2.18	4.00	3.24	8.23	2.20	8.14	3.17
ShiftPlus12	-0.91	3.15	-0.62	3.57	3.93	2.42	3.78	3.47	7.77	2.20	7.86	3.17
Compression4	0.00	4.00	0.00	4.12	4.00	1.71	4.00	3.42	8.00	4.00	8.02	4.17
Compression2	0.00	2.34	0.00	3.26	4.05	2.47	4.17	3.55	8.91	3.15	8.65	3.64
Compression0.5	0.00	2.11	0.00	3.14	4.14	2.21	4.04	2.97	7.48	1.24	7.01	1.84
Compression0.25	0.00	2.02	0.00	2.62	3.28	1.21	2.86	1.78	3.99	0.14	3.89	0.54
Centralization14	0.00	2.21	0.00	3.25	4.26	2.16	4.17	3.15	8.07	2.42	8.22	3.46
Centralization12	0.00	2.20	0.00	3.25	4.26	2.16	4.17	3.15	8.07	2.42	8.22	3.46
Centralization8	0.00	2.34	0.00	3.22	4.00	2.34	4.01	3.24	8.05	2.47	8.17	3.55
Centralization6	0.00	4.00	0.00	4.08	3.95	1.51	3.84	3.08	7.14	3.26	7.52	3.97

Table A8: **Mechanical treatment effects.** The table shows mean and uncertainty under the assumption of normally distributed inflation expectations $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2. Mean and uncertainty calculated using a mass-at-midpoint assumption after binning the normally distributed data in the intervals of the response scales.

Modifying the response scale may affect mean and uncertainty in Table 2 independently of any behavioral biases.[22] Spurious treatment effects may show up when a large part of the probability mass is assigned to an open interval or when the intervals are "too wide". As an example of the latter case, imagine a respondent who expects inflation to fall into the narrow range from 0 to 2 percent. The response scale in *Baseline* has narrow intervals in this range and thus allows the respondent to provide a histogram that closely reflect her beliefs. But other response scales, such as *Compression4* where the corresponding interval is from 0 to 8 percent, would distort the respondent's beliefs and we would overestimate the respondent's mean and uncertainty.

To illustrate the magnitude of these effects, consider a hypothetical setting in which a household with fixed probabilistic expectations is confronted with the response scales of the 13 treatments. Assuming that the household's expectations are normally distributed, we calculate the probability mass assigned to each interval and compute mean and uncertainty using a simple mass-at-midpoint measure (the results are similar when we follow Engelberg et al. (2009) and Becker et al. (2022) and calculate a smoothed response instead). For the household's normally distributed beliefs, we assume means of 0, 4, and 8 and variances of 4 and 9 to capture settings with low and high inflation uncertainty. Table A8 presents the results. Regarding the mean, the mechanical treatment effects are typically small, except in cases where a large part of the probability mass is assigned to an open interval (i.e., *ShiftMinus12* and *Compression0.25*). Regarding uncertainty, we observe mechanical treatment effects when the open intervals contain a large part of the probability mass and when the intervals are comparatively wide (e.g., *Compression4*).

[22] In order to avoid these "mechanical treatment effects" the tests in Sections 3.2.1 to 3.2.3 compare the probability masses the respondents assign to specific ranges of inflation. The mechanical treatment effects we describe in this appendix are absent in these probability-mass-tests.

Appendix G Bounds on respondents' prior expectations

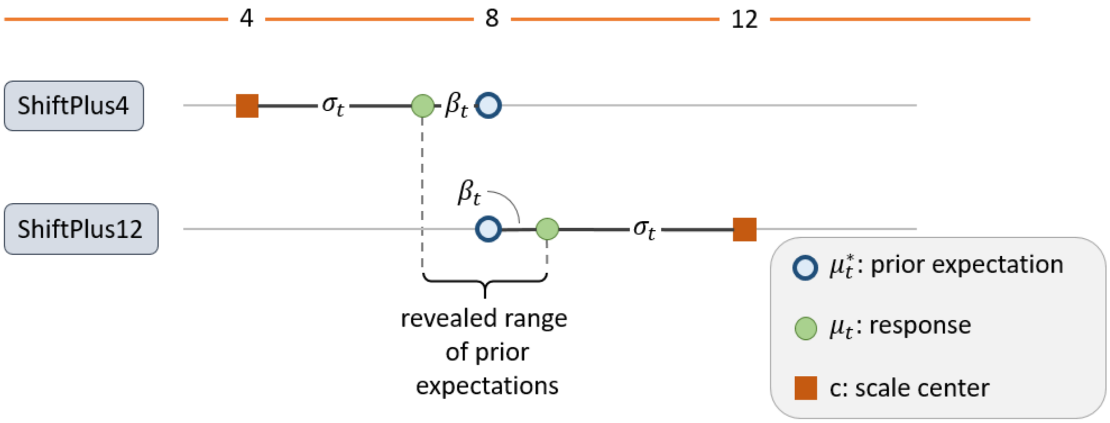


Figure A4: Constructing bounds on respondents' prior expectations μ_t^* under the assumption that the average mean response μ_t lies between the center of the response scale and respondents' prior expectations μ_t^*.

Under a comparatively strong assumption, the *Shift* treatments allow the identification of bounds on respondents' prior (or true) expectations. The prior expectations are a respondent's expectations before observing the response scale. This appendix sketches the idea. The central assumption is that the average mean response, denoted by μ_t, lies between the center of the scale, c, and respondents' prior expectations μ_t^*

$$c \leq \mu_t \leq \mu_t^* \quad \text{or} \quad c \geq \mu_t \geq \mu_t^*.$$

Figure A4 illustrates the scale center, the average mean response, and a possible location of the respondents' average prior distribution for *ShiftPlus4* and *ShiftPlus12*. $\beta_t = (\mu_t^* - \mu_t)$ denotes the average measurement bias and $\sigma_t = (\mu_t - c)$ the distance between the average mean response and the scale center. Using the numbers from Table 2, the average mean response in *ShiftPlus12* is 8.34 which is lower than the center of the response scale (12). Under our assumption, we may then conclude that 8.34 is an upper bound for μ_t^*. In *ShiftPlus4*, the average mean response (6.83) is larger than the center of the response scale (4) and we may conclude that 6.83 is a lower bound for μ_t^*. Notably, both the median and the trimmed mean of point forecast fall within these bounds, see Table A1.

How plausible is the assumption that the average mean response lies between the center of the scale and respondents' prior expectations? Respondents' tendency to move their answers towards the center of the scale (Result 1) is strong and it seems plausible that the majority of responses follow this pattern. However, violations of this assumption are possible. For example, even if we assume that such an ordering holds for every individual, it may be violated in the aggregate. In addition, the mechanical effects described in Appendix F may influence individual $\mu_{i,t}$ and therefore the ordering. Other violations are conceivable. We leave a rigorous treatment of this idea for future research.