HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Using point forecasts to anchor probabilistic survey scales

Christoph K. Becker

Peter Duersch

Thomas A. Eife

Alexander Glas

AWI DISCUSSION PAPER SERIES NO. 743
January 2024

Using point forecasts to anchor probabilistic survey scales

Christoph K. Becker* Peter Duersch†
Thomas A. Eife‡ Alexander Glas§

January 26, 2024

Abstract

We present the results of an experiment where a random subset of the participants in the Bundesbank's household panel receive personalized response scales, centered at each participant's point forecast. Personalized response scales offer two advantages over the standard scale which is centered at zero inflation: First, they mitigate the impact of the central tendency bias which leads respondents to assign greater probability mass to the center of the scale at zero. Second, they eliminate the need to adjust the scale when actual inflation falls outside the range for which the response scale was designed. Our results show that the personalized survey responses are of higher quality in three dimensions: (i) higher internal consistency, (ii) more uni-modal responses, and (iii) a significant reduction in the use of the (minimally informative) unbounded intervals of the response scale.

JEL codes: C83, D84, E31
Keywords: Inflation, density forecast, probabilistic forecast, experiment, survey design, personalized response scales

*Heidelberg University, Bergheimer Str. 58, 69115 Heidelberg, christoph.becker@awi.uni-heidelberg.de.
†Corresponding author, University of Mannheim, L7, 3–5, 68161 Mannheim, duersch@xeeron.de.
‡Heidelberg University, Bergheimer Str. 58, 69115 Heidelberg, thomas.eife@awi.uni-heidelberg.de.
§ZEW–Leibniz Centre for European Economic Research, L7, 1, 68161 Mannheim, alexander.glas@zew.de.

1 Introduction

In the past decade, several central banks have started new household surveys as part of their effort to manage inflation expectations. Most of these surveys include probabilistic questions where respondents are given a response scale with predefined intervals. The respondents are then asked to assign probabilities to the intervals that best represent their beliefs about inflation. It is common practice to provide the same response scale to all respondents.[1]

We report the results of an experiment in which the respondents receive a personalized response scale that is centered on the respondents' point forecast. All other characteristics of the scale (e.g., the number and widths of the intervals) are unchanged. The point forecasts are elicited in the question directly preceding the probabilistic question. Compared to fixed-center scales, personalized scales have two main advantages. First, since point forecast and scale center coincide by construction, the impact of the central tendency bias is mitigated. This bias refers to respondents' tendency to assign more probability mass to intervals in the middle of the scale and is a well-known phenomenon in survey research (e.g., Schwarz et al., 1985). Becker, Duersch, and Eife (2023) discuss how this bias distorts responses in the context of inflation expectations and Becker, Duersch, Eife, and Glas (2023) show that it exists in the Bundesbank Online Panel Households (BOP-HH). Second, personalized scales reduce the problem that actual inflation may fall outside the range of inflation rates covered by the fixed-center scale. The extreme outer intervals are typically unbounded and thus provide no means to signal upper or lower bounds on respondents' beliefs (see Figure 1). In order to minimize the use of the unbounded intervals, the ECB's Survey of Professional Forecasters (SPF) routinely adjusts its scale in an ad hoc fashion, aligning the center with the presumed inflation expectations of the respondents.[2] Personalized scales avoid these situations by automatically adjusting to changes of respondents' inflation expectations.

Our results show that personalized scales lead to higher-quality responses. Respondents' internal consistency (i.e., the consistency between their point forecast and their probabilistic forecast) is higher and the usage of unbounded intervals is reduced. In addition, we observe considerably more uni-modal responses. Responses with two more modes are generally treated as a signal that the respondent may have had difficulties in understanding or in answering the question. Our findings support this view. Dropout rates in both treatments are equally low, and respondents report that they do not find questions with personalized

[1] D'Acunto et al. (2023) and Dräger and Lamla (2023) provide recent overviews. Gülşen and Kara (2019) report that the Turkish central bank uses personalized response scales similar to what we discuss here. We are not aware of a systematic analysis studying the viability of personalized response scales.

[2] This does not prevent the possibility of surprise expectations. In the 2022Q2 wave, 66% of SPF respondents assigned positive probability to the upper, unbounded interval. Among those, the probability mass assigned to the interval was on average 26.19%.

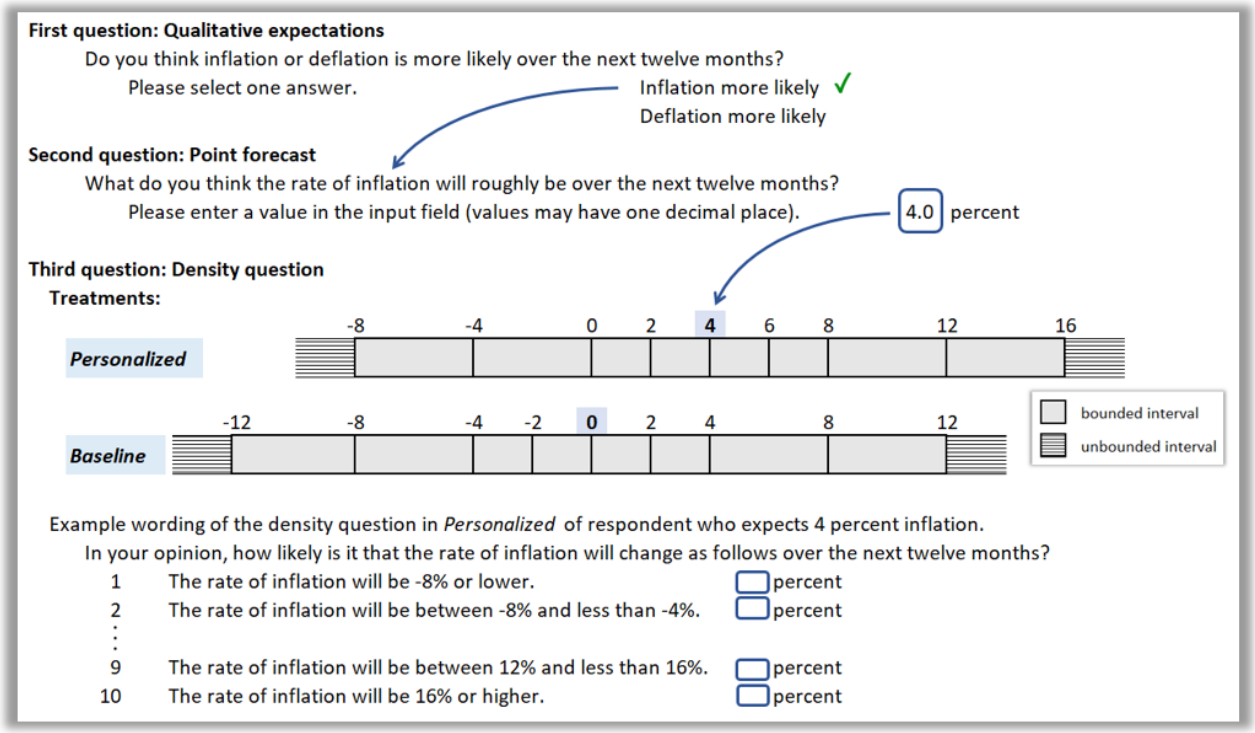

Figure 1: Illustration of the inflation expectations questions in the BOP-HH survey. The center of the personalized scale corresponds to the point forecast and may include a decimal place.

scales to be more difficult to answer. In fact, there is weak evidence that the respondents consider personalized scales to be less demanding.

2 Experiment

Our treatment design was included in Wave 38 of the Bundesbank's monthly BOP-HH. The survey has three questions on inflation expectations that are illustrated in Figure 1. All respondents receive the first (qualitative) question, the answer to which determines whether respondents are asked about the rate of inflation or of deflation in the second question. The answer to the second question then determines the center of the response scale for respondents in the *Personalized* treatment. The respondents in the other treatment (*Baseline*) receive the standard scale which is centered at zero inflation. Treatment allocation was random with 40% of the respondents assigned to *Personalized* and 60% to *Baseline*. Table 1 gives details.

Descriptive statistics and tests			**Treatment**		**Test**		
			Baseline	*Personalized*	*p*-value		Type
Responses	Number of respondents		2433	1632			
	Number of non-responses		88	68	0.378		t-test
	Number expecting deflation		222	138	0.493		”
	Number analyzed		**2123**	**1426**			
	Perceived difficulty (1-very difficult, 5-very easy)		3.32	3.37	0.079	*	t-test
Point forecast	Mean		7.66	7.68	0.916		”
	Median		7.00	7.00	0.725		Median
Probabilistic forecast	Mass-at-midpoint	Avg. Mean	5.94	7.08	0.000	***	t-test
		Avg. Uncertainty	1.96	1.43	0.000	***	”
	Beta	Avg. Mean	5.86	7.08	0.000	***	”
		Avg. Uncertainty	1.69	1.21	0.000	***	”
	Intervals	Avg. used	3.23	3.17	0.441		”
		Single (share)	0.23	0.21	0.263		”
		With gaps (share)	0.12	0.03	0.000	***	”
		Unbounded (share)	0.32	0.10	0.000	***	”
	Uni-modality	Probability (share)	0.91	0.97	0.000	***	”
		Density (share)	0.89	0.97	0.000	***	”
Internal consistency (between point forecast and mass-at-midpoint mean)	Correlation	Full sample	0.29	0.93	0.000	***	Fisher
		Trimmed sample	0.47	0.77	0.000	***	”
	Avg. absolute distance	Full sample	2.41	1.35	0.000	***	t-test
		Trimmed sample	2.11	1.33	0.000	***	”
	Point forecast within bounds	on the median (share)	0.59	0.89	0.000	***	”
		on the mean (share)	0.73	0.73	0.898		”
	Avg. width of bounds	on the median	3.76	2.14	0.000	***	”
		on the mean	3.80	2.23	0.000	***	”

Table 1: Respondents expecting deflation cannot be analyzed because of a coding error on the side of the data provider. ‘Beta’: Statistics based on a smoothed response following Engelberg et al. (2009) and Becker et al. (2022). Intervals: ‘Single’ counts respondents assigning 100 percent to a single interval, ‘With gaps’ counts respondents assigning zero percent to one or more intervals between two intervals with positive probabilities, ‘Unbounded’ counts respondents assigning probabilities to oneunbounded intervals. Internal consistency: Pearson correlation, testing via Fisher’s z-procedure (Zou, 2007), trimming one percent of point forecast, bounds calculated following Engelberg et al. (2009). All tests two-sided; */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

3 Results

The treatment intervention occurs after the point forecast has been elicited. Table 1 shows that there is no significant difference between respondents at the qualitative inflation/deflation question, nor at the point-forecast stage, indicating successful randomization. For the probabilistic question, the data show significant differences with lower average means and higher average uncertainty (i.e., average standard deviation) in *Baseline*. In the following, we compare the two ways of eliciting probabilistic forecasts and argue that the responses in *Personalized* are qualitatively better along three important dimensions.

Our first measure of quality is the share of responses with positive probability mass in the unbounded intervals. Being unbounded, these intervals are not very informative as there is no possibility to pin down what respondents truly believed when providing the answer. In practice, researchers typically impose bounds on the unbounded intervals before analyzing the responses.In *Baseline*, every third response uses an unbounded interval. In contrast, the number in *Personalized* is only one in ten. Panel A of Figure 2 provides more details. In both treatments, respondents use on average slightly more than three intervals. This and the share of responses using a single interval does not differ between the treatments, as illustrated by Panel B of Figure 2.

A second measure of quality is respondents' tendency to supply uni-modal responses. Responses with two or more modes are generally considered flawed and interpreted as a sign that the respondent may have had difficulties in understanding or in answering the question. Engelberg et al. (2009, p. 36) call uni-modality the "most basic assumption" in their parametric analysis. In *Personalized*, around three percent of responses have two or more modes, whereas in *Baseline* this number is about three times as large, as shown in Panel A of Figure 3.[3] A possible explanation for this treatment difference is that respondents in *Baseline* are confronted with two points on the response scale they may consider "focal". First, their point forecast and, second, the center of the response scale. Given that the center of the response scale is zero in *Baseline* and respondents' mean point forecast is 7.66, it is not surprising that the average histogram mean in *Baseline* is significantly lower than in *Personalized* (and lower than the respondents' average point forecast).

A third way to measure the quality of the responses is the internal consistency of the responses, i.e., how well the point forecast and the histogram forecast align.[4] Panel B of

[3]The different widths of the intervals mean that we have to distinguish between uni-modal probabilities and uni-modal densities.

[4]Zhao (2024) uses exponential tilting to match the means (or medians) of households' histogram forecasts to their point forecasts. While this procedure ensures by construction that the point and histogram forecasts are internally consistent, our approach to ex-ante center the response scale at the point forecast is less intrusive and allows to work with the actual survey responses without manipulating them.

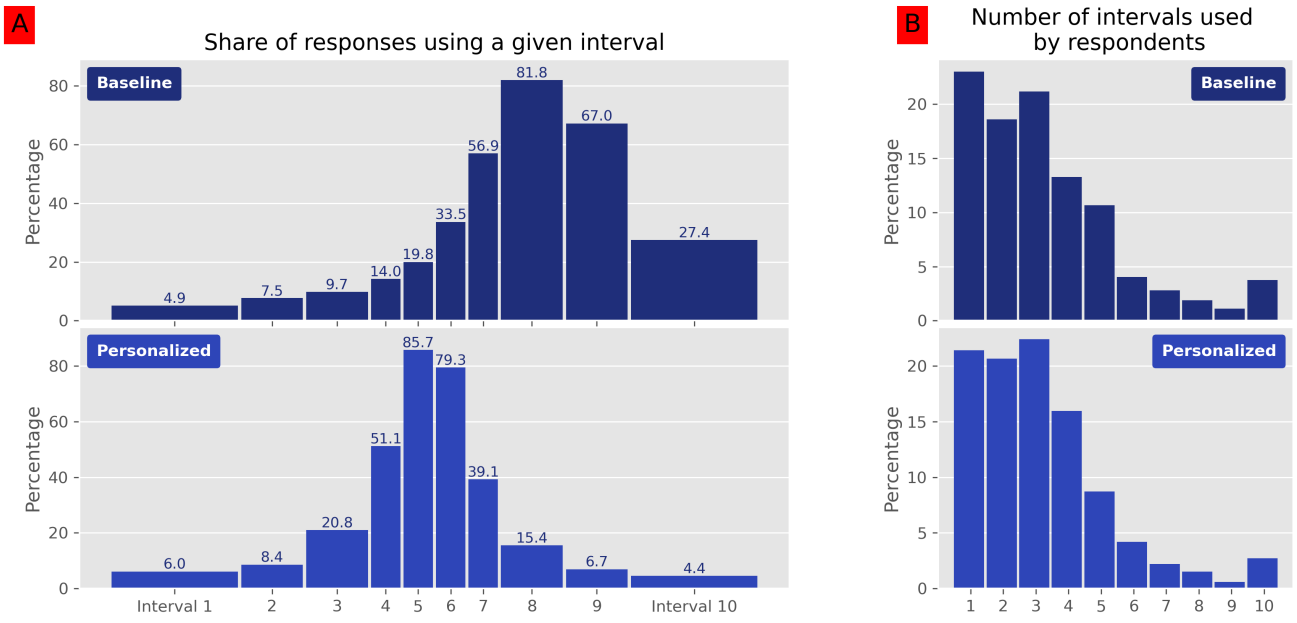

Figure 2: **Panel A**: Bar-plot indicating the share of the responses assigning positive probability to one of the ten intervals. Percentages do not sum up to 100 as responses may use more than one interval. The width of the bars corresponds to the with of the intervals. **Panel B**: Bar-plot of the number of intervals used.

Figure 3 illustrates the relationship between respondents' point forecast and their histogram means in the two treatments. Given that the question asking for respondents' point forecasts does not specify what measure of central tendency respondents should supply (e.g., mean, median, mode), some divergence is expected. The bottom rows of Table 1 present three ways of measuring consistency. The Pearson correlation between point forecast and histogram mean in *Personalized* is considerably higher than in *Baseline* (0.93 versus 0.29) and the difference remains significant even when we trim the sample and exclude the highest one percent of point forecasts. A second measure of consistency is the average absolute distance between point forecast and histogram mean. Again, the responses in *Baseline* are significantly less consistent than those *Personalized* according to this measure.

A third measure of consistency is proposed in Engelberg et al. (2009), who, for each respondent, weigh the interval limits with the interval's probability mass in order to construct upper and lower bounds on the mean. The upper and lower bounds on the median are given by the limits of the first interval in the cumulative histogram that has a probability equal to or above 50 percent. Table 1 reports these two measures but it should be noted that they need to be interpreted with care when analyzing surveys employing response scales with irregular interval widths. In *Personalized*, respondents tend to prefer the narrow center intervals (4, 5, 6, 7) whereas in *Baseline*, respondents tend to use the wide intervals 8 and 9 (see Panel A of Figure 2). As a consequence, the bounds suggested in Engelberg et al. (2009) are almost twice as wide in *Baseline* than in *Personalized*. Despite being significantly smaller and thus easier to "miss", we find a significant treatment difference for the median-bounds and no difference for the mean-bounds.

4 Acknowledgements

We are grateful to the Bundesbank for including our questions in the Bundesbank Online Panel Households. We would like to thank Christian Conrad, Zeno Enders, and Stefan Trautmann for helpful comments. Declaration of interest: none.

References

Becker, C., P. Duersch, and T. Eife (2023). Measuring inflation expectations: How the response scale shapes density forecasts. Available at SSRN: `http://dx.doi.org/10.2139/ssrn.4323706`.

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*.

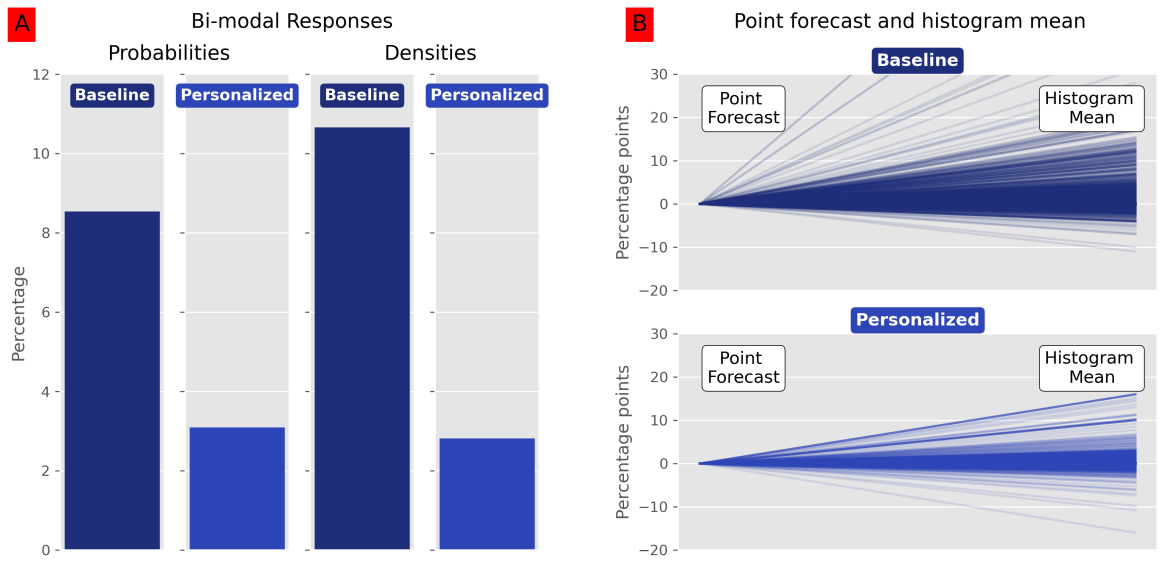

Figure 3: **Panel A** shows the share of bi-modal responses in the two treatments. Given the irregular widths of the intervals on the response scale, we have to distinguish between probabilities and densities. See text for details. **Panel B** illustrates the relationship between respondents' point forecast and their histogram means.

Becker, C., P. Duersch, T. A. Eife, and A. Glas (2023). Households' probabilistic inflation expectations in high-inflation regimes. *ZEW Discussion Paper No. 23-072*.

D'Acunto, F., U. Malmendier, and M. Weber (2023). What do the data tell us about inflation expectations? In *Handbook of economic expectations*, pp. 133–161. Elsevier.

Dräger, L. and M. J. Lamla (2023). Consumers' macroeconomic expectations. *Journal of Economic Surveys*.

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.

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

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.

Zhao, Y. (2024). Uncertainty of household inflation expectations: Reconciling point and density forecasts. *Economics Letters 234*, 111486.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods 12*(4), 399–413.