University of Heidelberg

Department of Economics

Discussion Paper Series | No. 664

Quantifying Subjective Uncertainty in Survey Expectations

Fabian Krüger and Lora Pavlova

May 2019

Quantifying Subjective Uncertainty in Survey Expectations

Fabian Krüger and Lora Pavlova*
Heidelberg University

May 29, 2019

Abstract

Several recent surveys ask for a person's subjective probabilities that the inflation rate falls into various outcome ranges. We provide a new measure of the uncertainty implicit in such probabilities. The measure has several advantages over existing methods: It is trivial to implement, requires no functional form assumptions, and is well-defined for all logically possible probabilities. From a theoretical viewpoint, the measure can be motivated as the entropy function of a strictly proper scoring rule. We demonstrate the measure's good performance in a simulation study based on empirical data from the Survey of Consumer Expectations.

1 Introduction

Economists have long used surveys to study expectations about future outcomes like the inflation rate or a person's wage. While surveys have traditionally focused on point expectations, Manski (2004), Delavande (2014) and Manski (2018) review a growing number of surveys in labor economics, development economics and macroeconomics that cover probabilistic expectations. The move from point predictions to probabilistic expectations seems natural since standard theories of choice indicate that probabilities matter for economic decision making.

In macroeconomics, the Survey of Professional Forecasters (SPF; Croushore, 1993) and its European counterpart (Garcia, 2003) are popular data sources covering experts' probabilistic forecasts. Furthermore, several recent surveys address the probabilistic expectations of firms and consumers. Examples include the Survey of Consumer Expectations (SCE) launched by the Federal Reserve Bank of New York (Armantier et al., 2017), a similar initiative by the Bank of Canada

*Financial support from the German Research Foundation (DFG) via grant KR 5214/1-1 is gratefully acknowledged.

(Gosselin and Khan, 2015), and the firm survey by Coibion et al. (2018). Given that expectations are potential drivers of economic decisions, these surveys are of interest to academics and policymakers alike (Trautmann and van de Kuilen, 2014; Armantier et al., 2015; Roth and Wohlfart, 2019; European Commission, 2019).

Figure 1 illustrates subjective probability distributions from the December 2017 wave of the SCE. Each survey participant provides probabilities for various outcome ranges of next year's inflation rate, as represented by the horizontal axis. In practice, each outcome range corresponds to a 'bin', to which participants are asked to assign a probability. The participant in the figure's top left panel assigns positive probability to a single bin, whereas the other participants use two (top right), three (middle left), four (middle right), five (bottom left) and ten (bottom right) bins. On average, SCE participants in the December 2017 wave use 4.3 bins, with 31 % of the participants using one or two bins.

The present paper considers methods for quantifying the uncertainty expressed by a histogram-type distribution as in Figure 1. Such uncertainty measures are an important input to studies that consider either the determinants or the consequences of subjective uncertainty. See, for example, Coibion et al. (2018) for an analysis of firms' expectations, Ben-David et al. (2019) for a household finance perspective (using the SCE data), and Rich and Tracy (2010) for a macroeconomic perspective (using expert forecasts from the Survey of Professional Forecasters, SPF).

Surveys like the SCE and the SPF feature a special type of censoring in that participants do not specify the probability distribution within each bin. This property precludes the direct computation of a standard measure of spread, such as the standard deviation or interquartile range, from the distribution. The popular approach by Engelberg et al. (2009) thus approximates a subjective histogram by a flexible generalized beta distribution. After this step, one can simply compute the desired measure of spread as implied by the fitted distribution. However, the approach is not feasible when all probability mass is concentrated on a small number of histogram bins. Engelberg et al. (2009) thus recommend to use the flexible approximation only if three or more bins contain nonzero probability, and to resort to a simple triangular distribution otherwise. This concern is important for the SCE, where about a third of the participants uses only one or two bins. Similar concerns apply to approximations based on other continuous distributions as reviewed by Glas (2018, Section 2) and Liu and Sheng (2019, Section 2).

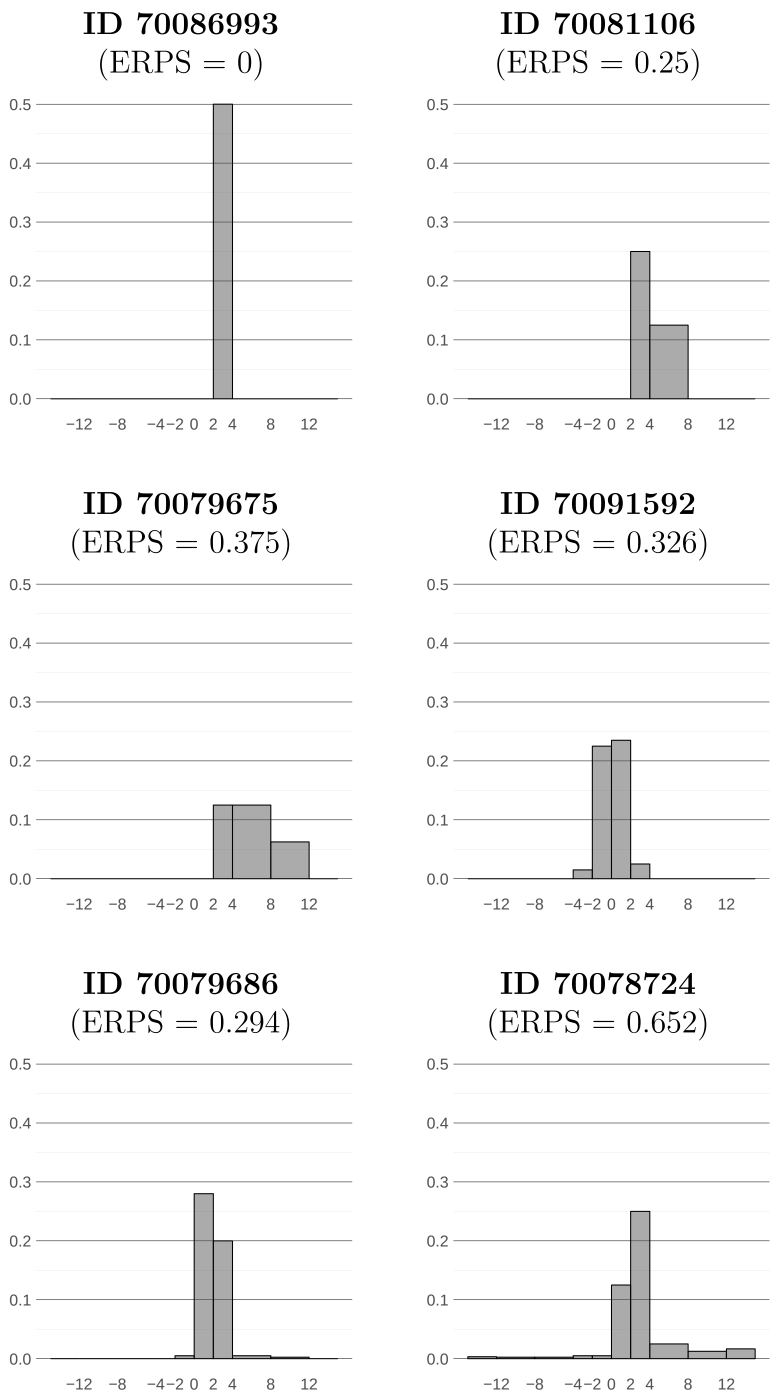

Figure 1: Illustration of probabilistic inflation expectations from the December 2017 wave of the SCE. The area of a rectangle corresponds to the subjective probability of the corresponding outcome range. For example, in the middle left panel the probability for an outcome between 4 and 8 equals $4 \times 1/8 = 1/2$.

Motivated by the properties of the SCE data, we propose a summary measure of uncertainty that is given by

$$\sum_{k=1}^{K} P_k \; (1 - P_k), \tag{1}$$

where $P_k = \sum_{j=1}^{k} p_j$ is the cumulative probability of the first k bins, and p_j indicates the probability of bin $j \in \{1, \ldots, K\}$. Note that the bins are arranged in ascending order: For example, the right endpoint of the first bin (-12% in the SCE) is smaller than the right endpoint of the second bin (-8% in the SCE). The measure at (1) has a number of advantages. First, it is transparent and trivial to implement. Second, it requires no information beyond that provided by survey respondents. By contrast, existing approaches require assumptions on the support of the subjective histogram, the distribution within each bin, or the functional form of the underlying continuous distribution. Third, it is well-defined even if all probability mass is concentrated on one or two bins. Finally, the measure can be theoretically motivated as the generalized entropy function of the ranked probability score (Epstein, 1969), a strictly proper scoring rule. We therefore refer to the measure at (1) as the ERPS, for Expected Ranked Probability Score.

To illustrate the ERPS, consider its assessment of the example distributions in Figure 1. The top left distribution attains the lowest possible score of zero, corresponding to no uncertainty. This assessment seems natural, given that all probability mass is contained in a single bin. The scores of the other subjective distributions are also shown in the figure. Intuitively, the scores reflect both the number of bins used by the respondents (ranging from one to ten in the figure) and the way the respondents distribute probability mass within these bins. For example, the middle right distribution yields a lower score than the middle left, as the majority of the mass is concentrated within the inner two adjacent bins, as compared to the more evenly distributed mass depicted in the middle left panel.

Along with the ERPS, we present a corresponding measure of objective uncertainty that is based on a recent proposal by Galvao and Mitchell (2019). Comparing subjective uncertainty (as measured by the ERPS) to a measure of objective uncertainty allows to assess whether survey participants' uncertainty can be rationalized from realized data. This comparison is of economic relevance since over– or underestimating objective uncertainty has possibly severe implications for decision making (see e.g. Moore et al., 2015, for a survey).

The remainder of this paper is structured as follows. Section 2 summarizes some stylized facts of the SCE data that motivate our methodology. Section 3 develops our proposal (the ERPS), detailing its advantages as mentioned above. Section 4 studies the behavior of the ERPS in Monte Carlo experiments based on the SCE data. In particular, we document that the ERPS is robust with respect to variations in the design of survey histograms, and to small changes in the subjective probabilities. Section 5 provides empirical illustrations. Section 6 sketches a corresponding measure of realized uncertainty, and Section 7 concludes. Proofs, implementation details and additional results are collected in appendices.

2 Background on the SCE data

As noted, the SCE data is a major motivation for the methods we propose in this paper. Here we briefly describe stylized facts that are relevant to our proposal, and review the Engelberg et al. (2009) quantification method. The latter method is used to derive uncertainty measures that are reported in official SCE publications such as Armantier et al. (2017).[1]

2.1 Design and stylized facts of probability questions

The SCE is conducted at a monthly frequency with a sample size of about 1,200 respondents per month. The core module of the SCE asks for subjective probabilities of various outcome ranges, covering three variables: The Consumer Price Index (CPI) at two different horizons, real estate prices, and the respondent's personal earnings. In the SCE questionnaire made available by Federal Reserve Bank of New York (2019), the relevant question codes are Q9 and Q9c (CPI inflation rate), C1 (growth rate of the average home price nationwide) and Q24 (growth rate of the respondent's personal earnings). The relevant outcome ranges (in percent), which are the same for all variables, can be represented by the intervals

$$(-\infty, -12]; (-12, -8]; (-8, -4]; (-4, -2]; (-2, 0); [0, 2); [2, 4); [4, 8); [8, 12); [12, \infty).$$

These outcome ranges are reflected in the horizontal axis labels of Figure 1. In the case of inflation, for example, the two rightmost intervals refer to an inflation rate between 8% and 12% and to an inflation rate of 12% or more.[2]

The upper panel of Table 1 presents summary statistics on the number of histogram bins used by SCE participants (that is, the number of bins containing strictly positive probability mass). We focus on the time period from January 2014 to December 2017 for comparability to the SPF (see below). For inflation and the average home price, around 30% of the participants uses one or two bins. For personal earnings, roughly half of the participants use one or two bins. The mean number of bins used is somewhat higher for inflation and the average home price ($4.2 - 4.4$), compared to personal earnings (3.3). Finally, more than a quarter of the participants use one or both of the outer bins that correspond to the intervals $(-\infty, -12]$ and $[12, \infty)$.

[1]These uncertainty measures are also included in the SCE data set that is available for public download (Federal Reserve Bank of New York, 2019).

[2]The inclusion (or exclusion) of interval limits is not specified by the SCE survey questions. For example, the survey question leaves it unspecified whether an inflation rate of exactly 12% belongs to the last or penultimate bin. Our choice of half-open intervals (with the exception of the $(-2, 0)$ interval) are arbitrary – as is any choice in that regard – but seem unlikely to be of empirical relevance.

	n	Share of respondents using one bin	two bins	outer bin(s)	Mean nr. of bins
			SCE		
Average Home Price	54220	16.0	15.8	40.1	4.2
Inflation (one-year)	61762	12.4	17.1	39.6	4.4
Inflation (three-year)	61874	13.0	17.5	39.4	4.4
Personal Wage	41837	26.4	23.9	28.5	3.3
			SPF		
Inflation (GDP def.)	569	2.5	14.9	19.2	4.5
GDP	592	3.9	17.7	7.1	4.5
Inflation (CPI)	571	1.6	14.0	14.2	4.6
Inflation (PCE)	539	1.1	15.2	13.9	4.6
Unemployment	567	8.6	22.2	29.8	3.4

Table 1: Summary statistics on the number of bins used in the SCE (January 2014 to December 2017 waves) and SPF (2014:Q1 to 2017:Q4 waves); n denotes the total number of responses. We exclude histograms that do not sum to one (less than 0.35% of responses in both surveys).

The lower panel of Table 1 presents analogous statistics for the SPF. The SPF histograms are similar in design to those of the SCE, except that the two surveys use different numerical ranges for the histogram bins.[3] The number of bins (ten) is the same as in the SCE, except for GDP (eleven). While the share of participants using two bins and the mean number of bins used are comparable to the SCE, there are two major differences to the SCE: First, the SPF features a much smaller share of participants who use a single bin. For example, this share is about ten percentage points lower for the inflation variables. Second, the share of participants using at least one outer bin is much smaller in the SPF. For example, this share is more than 20 percentage points lower for the inflation variables. Based on these statistics, we will argue that the common strategy of fitting a continuous distribution is less promising in the case of the SCE data than in the case of the SPF data.

2.2 The Engelberg et al. (2009) quantification method

The histogram probabilities do not specify a full probability distribution since the endpoints of the histogram's support as well as the distribution within each bin are unknown. Based on the raw probabilities alone, it is hence impossible to compute each participant's subjective mean or variance. Following similar

[3]While the SPF's bin definitions have been adapted over time (see the documentation by Federal Reserve Bank of Philadelphia, 2019, for details), they are constant over the time period reported in Table 1.

methods for income expectations (Dominitz and Manski, 1997), Engelberg et al. (2009) propose to fit a continuous distribution to the histogram probabilities. The choice of continuous distribution depends on the number of histogram bins that are being used: The authors propose to fit a simple triangular distribution in case a forecaster uses one or two bins, and to fit a flexible generalized Beta distribution if the forecaster uses three or more bins. We provide formal details in Appendix B. Figure 4 in the appendix provides examples for the method's practical application to the SCE data.

A major advantage of the quantification methodology is that it provides a full analytical distribution from which any feature of interest (such as measures of location or spread and probabilities for 'extreme' events) can be computed. However, this wealth of information comes at a cost: First, the choice of a particular parametric distribution is potentially restrictive, and seems hard to justify if the histogram uses only one or two bins. Second, the approach entails a discontinuity when moving from two bins (approximated via a triangle distribution) to three bins (approximated via a generalized beta distribution). We provide a numerical evaluation of this discontinuity in Section 4.2. Finally, practical implementation requires judgmental choices pertaining, e.g., to parameter limits imposed in numerical optimization, or to the handling of certain cases that are not covered by Engelberg et al.'s original proposal, simply because they did not occur in their original data set (see Appendix B for details). These choices may reasonably be made differently by different authors. Full reproducibility hence requires careful documentation of all implementation choices.

As noted above, the Engelberg et al. (2009) quantification method seems more appealing in the context of the SPF than for the SCE which features a higher share of single-bin histograms and a more pronounced use of outer bins. Furthermore, the larger sample size of the SCE calls for simple and robust methods that apply to any possible histogram. We next propose a method that satisfies these criteria.

3 A new approach to quantifying uncertainty in survey histograms

3.1 General idea: Quantifying uncertainty via entropy

Suppose a survey participant issues a vector of probabilities $\underline{p} := \big(p_1, \quad p_2, \quad \ldots, \quad p_K\big)'$, where p_k denotes the subjective probability that the inflation rate is within the interval r_k that defines the range of bin k. See Section 2.1 for the intervals characterizing the SCE's bin definitions. In practice, the intervals $\{r_k\}_{k=1}^K$ are disjoint and their union is the real line. Hence the probabilities $\underline{p}$ form a subjective survey histogram as in Figure 1.

Our proposed measures of uncertainty are based on the concept of entropy.

Informally, if the entropy of distribution $\underline{p}$ is large, then a forecaster with subjective distribution $\underline{p}$ places a high probability on making large forecast errors. In that sense, $\underline{p}$ is associated with little predictability and high uncertainty. Vice versa, under a low-entropy distribution $\underline{p}$, large forecast errors are unlikely, and hence low entropy is associated with high predictability and low uncertainty.

More formally, the concept of entropy relates to strictly proper scoring rules. The latter are a standard tool in statistical decision theory. In economics, they are commonly used for eliciting beliefs in experiments (Schotter and Trevino, 2014) and for evaluating probabilistic forecasts (Gneiting and Katzfuss, 2014). Scoring rules are functions of the form $\mathrm{S}(\underline{p}, k^*)$ that measure the performance of the probabilistic forecast $\underline{p}$ if the outcome k^* realizes. The integer $k^* \in \{1, 2, \ldots, K\}$ indicates the histogram bin that contains the realization. We consider specific choices of S below. For each of these choices, a smaller value of S indicates a better forecast. A scoring rule S is called strictly proper if a forecaster minimizes their expected score by stating what they think is the true probability distribution p (conditional on their information set); see Gneiting and Katzfuss (2014, Section 3.1.1) for a formal definition. The function

$$\mathrm{ES}(\underline{p}) \;=\; \sum_{k=1}^{K} p_k \; \mathrm{S}(\underline{p}, k) \tag{2}$$

is called the entropy function associated with the scoring rule S (e.g. Gneiting and Raftery, 2007, Section 2.2). We propose to use this function in order to measure the subjective uncertainty in a probabilistic survey forecast $\underline{p}$. To obtain intuition, consider a very confident forecaster who places a probability of one on a single bin l, such that $\underline{p} := \underline{p}_l$ is a vector of zeros, except for the lth entry being equal to one. If the lth entry materializes (i.e., if $k^* = l$), then the forecast is perfect in retrospect. Indeed, both choices of S we consider below satisfy $\mathrm{S}(\underline{p}_l, l) = 0$, which implies that $\mathrm{ES}(\underline{p}_l) = 0$. This implication is plausible: Since the forecaster places all probability mass on a single bin, they implicitly state that the only possible situation is one in which this bin realizes, such that the forecast is perfect in retrospect. By contrast, consider a forecaster who is very uncertain, such that the probability forecast $\underline{p} := \underline{p}_u$ is a discrete uniform distribution across all K bins. The uncertain forecaster hence places considerable probability mass on outcomes k for which $\mathrm{S}(\underline{p}, k)$ is 'large', i.e. the forecast is considered poor in retrospect (after all, the forecaster placed a probability of $1 - 1/K$ on outcomes other than the one that realized). Accordingly, $\mathrm{ES}(\underline{p}_u)$ is 'large' as well.

3.2 Expected Ranked Probability Score (ERPS)

Define

$$P_k = \sum_{j=1}^{k} p_j$$

to be the cumulative probability of the first k bins. As a first choice of scoring rule S, we consider the ranked probability score (RPS; Epstein, 1969) given by

$$\text{RPS}(\underline{p}, k^*) = \begin{cases} \sum_{k=1}^{K}(1-P_k)^2 & \text{if } k^* = 1 \\ \sum_{k=1}^{k^*-1}(P_k)^2 + \sum_{k=k^*}^{K}(1-P_k)^2 & \text{if } k^* \in \{2, 3, \ldots, K\} \end{cases};$$

The score rewards forecasters who put much probability mass into bins that are equal or close to the realizing bin k^*. The entropy function for the RPS is given by

$$\begin{aligned} \text{ERPS}(\underline{p}) &= \sum_{k=1}^{K} p_k \ \text{RPS}(\underline{p}, k) \\ &= \sum_{k=1}^{K} P_k(1-P_k). \end{aligned} \quad (3)$$

As its name suggests, the RPS is designed for ranked categorical variables. That is, the RPS treats the realizing bin $k^* \in \{1, \ldots, K\}$ as an ordinal variable, with $k^* = 1$ representing a smaller outcome than $k^* = 2$. However, the RPS does not attach any numerical interpretation to the outcomes.[4] In the context of survey histograms, this means that the RPS is invariant to re-definitions of the bin ranges, as long as their ordering is preserved and the probability in each bin remains unchanged.

As we show in Appendix A, the maximal ERPS is attained for the vector

$$\underline{p}^* = \big(1/2, \quad 0, \quad \ldots, \quad 0, \quad 1/2\big)'$$

that places probability one half on each of the two outer bins. The intuition for this solution is that under $\underline{p}^*$, it is certain that one of the two outer bins will materialize. Both outcomes produce a large score $\text{RPS}(\underline{p}^*, k)$, since $\underline{p}^*$ places no probability mass on the neighboring bins.

3.3 Expected Brier Score (EBS)

The Brier score (BS; Brier, 1950) is given by

$$\text{BS}(\underline{p}, k^*) = \sum_{k=1}^{K} (\mathbb{I}_{k^*=k} - p_k)^2,$$

[4] The continuous counterpart to the RPS, the Continuous Ranked Probability Score (CRPS Matheson and Winkler, 1976) is a popular strictly proper scoring rule in the econometric literature on probabilistic forecasting; see Krüger et al. (2019) and the references therein.

where $\mathbb{I}_{k^*=k}$ is an indicator function that equals one if $k^* = k$, and equals zero otherwise. The entropy function for the BS is given by

$$\begin{aligned} \text{EBS}(\underline{p}) &= \sum_{k=1}^{K} p_k \ \text{BS}(\underline{p}, k) \\ &= \sum_{k=1}^{K} p_k(1 - p_k). \end{aligned} \tag{4}$$

The BS is designed for multinomial random variables, that is, the outcome categories $k^* \in \{1, \ldots, K\}$ are viewed as interchangeable labels. In the context of survey histograms, this means that the BS is invariant to permutations of the histogram probabilities.

López-Menéndez and Pérez-Suárez (2019) have recently considered the EBS for measuring uncertainty in tendency forecast surveys, where each participant states a deterministic prediction of whether the economy will go 'up', 'down' or remain 'unchanged'. They then construct an aggregate probability distribution where the probability of each category ('up', 'down' or 'unchanged') corresponds to the proportion of participants for that category. They use the EBS in order to quantify the uncertainty of their aggregate probability distribution at a given point in time. By contrast, probability information in the histograms allows us to study the uncertainty of any individual survey participant.

As noted by López-Menéndez and Pérez-Suárez (2019), the maximal EBS is attained for

$$\underline{p}^{**} = \tau \times (1/K),$$

where τ is a $K \times 1$ vector of ones (see Appendix A for details). Hence flat probabilities represent maximal uncertainty, as is standard in a multinomial setup.

3.4 Discussion

The idea of measuring uncertainty via entropy is standard, and has most famously been applied in the context of the logarithmic score given by $\text{LS}(\underline{p}, k^*) = -\log p_{k^*}$, for which the entropy function $\text{ELS}(\underline{p})$ is the Shannon entropy; see Gneiting and Raftery (2007, Section 2.2) and the references therein. In economics, Shannon entropy plays a key role in the theory of rational inattention (Sims, 2003; Caplin et al., 2017). Despite its general popularity, applying the logarithmic score to survey histograms seems problematic: Due to the logarithmic function, it is very sensitive to small changes in the probabilities whenever some of the elements in $\underline{p}$ are close to (but strictly greater than) zero. This sensitivity seems unjustified in a survey context, where small changes may occur due to rounding or practical details of the survey design (such as the number of digits that participants are allowed to enter). See Selten (1998) for further discussion, and Boero et al. (2011)

for empirical results on survey histograms.[5] Our use of the RPS and BS, both of which remain numerically stable in the presence of small nonzero probabilities, is in line with the recommendations of the latter paper.

As noted, the RPS is based on an ordinal interpretation of the histogram bins, whereas the BS is based on a multinomial interpretation. Accordingly, neither the ERPS nor the EBS utilize information on the numerical definition of each bin. This property is very useful, as it removes the need to make assumptions on the support of the histogram (i.e., the endpoints of the two outer bins) or the probability distribution within each bin. That said, the multinomial interpretation underlying the EBS implies that the latter is invariant to permutations of the probabilities $p_1, \ldots, p_K$. For example, for a hypothetical three-bin histogram, the probabilities $\underline{p}_a = (1/4, 1/2, 1/4)'$ yield the same EBS as the probabilities $\underline{p}_b = (1/2, 1/4, 1/4)'$. This assessment seems implausible, given that $\underline{p}_b$ is obtained from $\underline{p}_a$ by shifting probability mass from the central bin to the more extreme leftmost bin. Under the ERPS, which utilizes an ordinal interpretation, $\underline{p}_b$ is considered more uncertain than $\underline{p}_a$.

4 Simulation study

This section presents numerical evidence on the behavior of the uncertainty measures we consider.

4.1 Design of the histogram bins

We first consider various designs for the histogram bins, including the one used by the SCE. At present there seems to be no strong consensus on how the histogram bins should be defined; for example, the SCE's bin definitions differ widely from the SPF's, and the SPF's definitions have been updated over time. We hence contend that an uncertainty measure should be qualitatively robust across various 'reasonable' choices. To study the robustness of our proposed measure, we consider a continuous distribution that represents a respondent's true but unobservable beliefs. We then discretize the distribution according to four sets of histogram bins shown below. Each bin design yields a histogram from which we then recover a measure of forecast uncertainty. Ideally, the measures obtained for the four bin designs should be similar, given that the underlying continuous forecast distribution is the same in each case.

[5] More specifically, the sensitivity of the logarithmic score appears unproblematic in the context of the Shannon entropy $\text{ELS}(\underline{p})$, but is of considerable practical relevance when computing the realized score $\text{LS}(\underline{p}, k^*)$ for an outcome bin k^*. The realized score is central for comparing subjective and objective uncertainty, an issue we address in Section 6.

Specifically, we consider the following histogram bin designs:

$$(-\infty,-12];(-12,-8];(-8,-4];(-4,-2];(-2,0);[0,2);[2,4);[4,8);[8,12);[12,\infty) \qquad \mathbf{B1}$$

$$(-\infty,-6];(-6,-4];(-4,-2];(-2,0);[0,2);[2,4);[4,6);[6,\infty). \qquad \mathbf{B2}$$

$$(-\infty,-4];(-4,-2];(-2,0);[0,2);[2,4);[4,6);[6,8);[8,\infty). \qquad \mathbf{B3}$$

$$\begin{aligned}(-\infty,-4];(-4,-3];(-3,-2];(-2,-1];(-1,0);[0,1);\\ [1,2);[2,3);[3,4);[4,5);[5,6);[6,7);[7,8);[8,\infty)\end{aligned} \qquad \mathbf{B4}$$

The first design, **B1**, is the one used by the SCE. The bins in **B2** cover a smaller range than the ones in **B1**, but are otherwise identical. **B3** is obtained by shifting **B2** to the right, such that the bins are symmetric around two percent (in contrast to **B1** and **B2**, which are symmetric around zero). Finally, the bins in **B4** are finer than the ones in **B3**, possibly allowing for a more detailed assessment of subjective uncertainty.

Of course, differences between the four histogram designs depend on the choice of underlying continuous distribution. We therefore consider a large number of distributions that we estimate from $n = 1,210$ histograms for inflation (one year ahead) from the SCE's December 2017 wave.[6] For participant $i = 1, \ldots, n$, we run the following steps.

- Apply the Engelberg et al. (2009) quantification method to participant i's survey histogram. As described in Appendix B, we obtain a triangular distribution in case the histogram contains one or two (adjacent) bins, and we obtain a generalized beta distribution otherwise.
- Discretize the distribution according to the bin designs **B1** to **B4** introduced above. This step yields the histograms $\underline{p}_{i,b}$, where $b \in \{1, 2, 3, 4\}$ indicates the bin design.
- Compute the uncertainty measures $\text{ERPS}(\underline{p}_{i,b})$, $\text{EBS}(\underline{p}_{i,b})$, $\hat{\sigma}(\underline{p}_{i,b})$, and $\text{IQR}(\underline{p}_{i,b})$, whereby the latter two denote the standard deviation and interquartile range of Engelberg et al.'s continuous approximation to $\underline{p}_{i,b}$.

To assess the robustness of the ERPS measure across the bin designs **B1** to **B4**, we then estimate the correlation matrix of $\text{ERPS}(\underline{p}_{i,b})$ across b. Ideally, the entries of this matrix should all be close to one. We proceed analogously for EBS as well as the fitted standard deviation and IQR.[7]

[6] We drop nine participants whose histogram probabilities are negative or do not add to one, and 54 participants for which the Engelberg et al. (2009) quantification method is undefined; see Appendix B for details.

[7] By using correlation as a measure of similarity, we do not require the uncertainty measures to remain numerically identical across bin designs but only require close linear dependence. Our results are qualitatively unchanged when using Spearman rank correlation instead of standard (Pearson) correlation.

ERPS

	B1	B2	B3	B4
B1	1.000	0.970	0.966	0.960
B2	0.970	1.000	0.979	0.978
B3	0.966	0.979	1.000	0.995
B4	0.960	0.978	0.995	1.000

EBS

	B1	B2	B3	B4
B1	1.000	0.876	0.907	0.702
B2	0.876	1.000	0.969	0.866
B3	0.907	0.969	1.000	0.861
B4	0.702	0.866	0.861	1.000

$\hat{\sigma}$

	B1	B2	B3	B4
B1	1.000	0.890	0.953	0.955
B2	0.890	1.000	0.934	0.930
B3	0.953	0.934	1.000	0.988
B4	0.955	0.930	0.988	1.000

Interquartile Range

	B1	B2	B3	B4
B1	1.000	0.909	0.962	0.961
B2	0.909	1.000	0.947	0.942
B3	0.962	0.947	1.000	0.990
B4	0.961	0.942	0.990	1.000

Table 2: Correlation matrices of uncertainty measures across bin designs, based on $n = 1,210$ probability histograms in each setup. See text for details.

Table 2 summarizes the simulation results. The ERPS, standard deviation and IQR are generally robust across bin designs, with correlation coefficients exceeding 0.93 in all instances except one (standard deviation $\hat{\sigma}$, correlation between settings **B1** and **B2**). The EBS is somewhat less robust across bin designs, with four of the six pairwise correlations being below 0.9. The latter result can partly be explained by the fact that the EBS treats the histogram probabilities as multinomial. This means that shifting probability mass to a neighboring bin is equivalent (in terms of EBS) to shifting probability mass to a distant bin.

4.2 Discontinuity of the Engelberg et al. (2009) method

As noted, the Engelberg et al. (2009) method uses a triangle distribution if the participant uses at most two bins, and uses a generalized beta distribution if the participant uses three or more bins. This case distinction leads to a discontinuity. We next present data-based simulation experiments to assess the magnitude of this discontinuity in empirically plausible scenarios.

We use data from the January 2017 to December 2017 waves of the SCE. We focus on participants who use two adjacent bins, none of which is an outer bin in the SCE's histogram design shown at **B1**. We further require that the histogram probabilities sum to one and exceed a threshold of one percent, which is the magnitude of the perturbation we consider. These selection criteria leave us with 2450 two-bin histograms. For each of these histograms, we consider two simple perturbations: First, we move one percentage point of probability mass from the left bin to its left neighboring bin. For example, suppose that the original histogram allocates 50% probability to the two bins $[0, 2)$ and $[2, 4)$. The perturbed histogram then places probability 1%, 49% and 50% to the three bins $[-2, 0)$, $[0, 2)$ and $[2, 4)$ respectively. Second, we apply an analogous perturbation to the right histogram bin, such that the perturbed histogram contains one percent of probability mass in a third bin located to the right of the original histogram. A perturbation of one percent is the smallest size that seems empirically plausible, and is hence deemed appropriate for the present experiment.

As expected, both types of perturbation lead to a rightward shift in the distribution of individual standard deviations. For example, the median standard deviation increases from 0.739 (without perturbation) to 0.784 (under either type of perturbation). Table 3 further shows the correlation between the fitted standard deviation under each setup. The correlation between the baseline setup and two perturbed versions is at 0.807 and 0.643 for left and right perturbation, respectively. These correlations seem remarkably modest given the small magnitude of the perturbation. The results for IQR are similar, with the two correlations being equal to 0.825 and 0.706, respectively.

The results in the bottom panels of Table 3 also indicate that the impact of right perturbation is larger than the impact of left perturbation. This effect is due to the empirical pattern that many of the two-bin histograms focus on the bins $[2, 4)$ and $[4, 8)$. According to the SCE's bin design shown at **B1** above, the left neighbor of these bins is at $[0, 2)$, whereas the right neighbor is at $[4, 8)$. Hence left perturbation expands the support of the histogram by two units, whereas right perturbation expands the support by four units. This asymmetry matters here since the Engelberg et al. (2009) algorithm adopts the support of the histogram if only interior bins are used.

For ERPS and EBS, the impact of the perturbation can be described analytically. Let $\underline{p}$ denote a two-bin histogram, and $\underline{\tilde{p}}_L$ and $\underline{\tilde{p}}_R$ its perturbed version with probability mass shifted to the left and right neighboring bin, respectively. Let δ denote the size of the perturbation (with $\delta = 0.01$ in our simulation study). For the ERPS, Equation (3) yields that

$$\text{ERPS}(\underline{\tilde{p}}_L) = \text{ERPS}(\underline{\tilde{p}}_R) = \text{ERPS}(\underline{p}) + \delta \ (1 - \delta),$$

i.e. both perturbations lead to an additive increase in ERPS by $\delta \ (1 - \delta)$. This means in particular that perturbation affects all n histograms in exactly the same way, leading to correlations of one in Table 3. For the EBS, Equation (4) yields that

$$\text{EBS}(\underline{\tilde{p}}_j) = \text{ERPS}(\underline{p}) + 2 \ \delta \ (p_j - \delta) > \text{ERPS}(\underline{p}),$$

where $j \in \{L, R\}$, and p_j denotes the probability mass in the (left or right) bin from which one percentage point is removed. Hence perturbation leads to an additive increase in EBS by $2 \ \delta \ (p_j - \delta)$. Empirically, this increase is negligible, in the sense that the correlation of the EBS values across perturbation setups is very close to one (see Table 3).

ERPS

	N	L	R
N	1.000	1.000	1.000
L	1.000	1.000	1.000
R	1.000	1.000	1.000

EBS

	N	L	R
N	1.000	0.999	0.999
L	0.999	1.000	0.997
R	0.999	0.997	1.000

$\hat{\sigma}$

	N	L	R
N	1.000	0.807	0.643
L	0.807	1.000	0.523
R	0.643	0.523	1.000

Interquartile Range

	N	L	R
N	1.000	0.825	0.706
L	0.825	1.000	0.609
R	0.706	0.609	1.000

Table 3: Correlation matrices of uncertainty measures across perturbation setups (N = no perturbation, L = left perturbation, R = right perturbation), based on $n = 2,450$ probability histograms in each setup. See text for details.

4.3 Discussion

The simulation experiment of Section 4.1 indicates that ERPS and the measures based on the Engelberg et al. (2009) method are robust with respect to various design choices related to the number, width and support of the histogram bins. This evidence is reassuring since there are various plausible choices of these parameters, with little methodological guidance as to which choice is most appropriate.

The evidence in Section 4.2 indicates that uncertainty measures based on the Engelberg et al. (2009) can be sensitive to small perturbations of the probability histogram. This sensitivity is due to a case distinction which implies that the derived uncertainty measures are discontinuous functions of the vector $\underline{p}$ of probabilities, which seems economically implausible. The ERPS and EBS do not share this drawback. The robustness of the ERPS in the setup of Section 4.2 can be established via simple analytical calculations.

While this paper focuses on measures of subjective uncertainty, a measure of location (such as the mean of the subjective forecast distribution) is also of interest in many applications. The Engelberg et al. (2009) method allows to construct such a measure. In the simulation designs of Sections 4.1 and 4.2, the mean forecast constructed from Engelberg et al.'s method is remarkably robust across the variants we consider in each design; see Table 4 in the appendix for details. Hence the method's discontinuity is not reflected in its mean forecasts.

5 Empirical illustration

This section presents an empirical illustration of the ERPS as a measure of individual uncertainty using SCE data from the June 2013 to December 2017 waves. We consider respondents' probabilistic forecasts about the inflation rate, change in the average price of a home, and change in personal earnings. All of these expectations are one year ahead and are taken from the core module of the survey (see Section 2.1 for details on the data). Figure 2 plots the average ERPS across all respondents within each wave.[8] Interestingly, respondents are consistently more uncertain about the future development of the two macroeconomic outcomes (inflation rate and house prices, marked by black and green lines), as compared to uncertainty about personal earnings (red line). The figure further shows that average uncertainty about the three variables is remarkably constant over time. The latter result can perhaps be rationalized by the relatively long one-year horizon to be forecast (such that conditioning information may be of limited subjective relevance), and by the short sample period of the SCE.

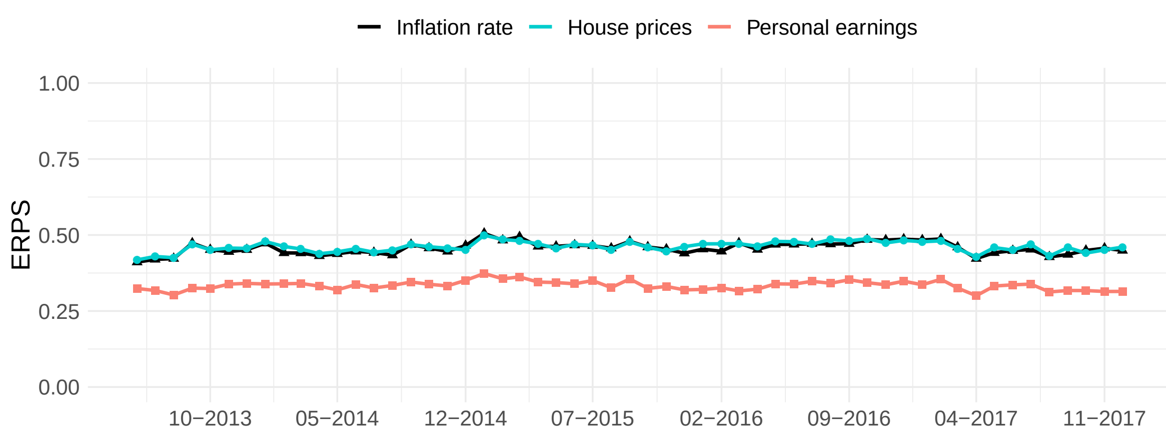

Figure 2: Subjective uncertainty (ERPS) across economic variables. For each variable and month, the figure shows the mean ERPS across survey respondents.

We next analyze individual heterogeneity in subjective uncertainty. Following Federal Reserve Bank of New York (2019), we distinguish respondents according to a number of demographic and socioeconomic characteristics: Age, educational attainment, household income, as well as financial literacy and numeracy skills.[9]

[8]For comparability, Figure 2 is based on respondents who provide expectations for all variables. This implies in particular that it covers only participants who are employed since earnings uncertainty is only available for these participants.

[9]The official SCE website (Federal Reserve Bank of New York, 2019) publishes graphical summaries of uncertainty for these demographic groups. Regarding their age, survey participants are classified into three groups: 'Under 40', '40 to 60' and 'Over 60' years. In terms of educational attainment, the SCE allows to distinguish between respondents with no college education, some college education and

Figure 3 plots the time trends in uncertainty by demographics, focusing on inflation uncertainty for brevity. The figure indicates that older, richer, more highly educated and more financially literate survey participants experience lower inflation uncertainty. These patterns are fairly consistent, in that the lines in Figure 3 rarely cross. Furthermore, we again observe little time variation in subjective uncertainty.

The differences across demographic and socioeconomic groups just reported are broadly similar for inflation at a three-year horizon as well as for house prices and personal earnings at a one-year horizon; see Figures 5 to 7 in the appendix. Furthermore, our findings in Figures 2 and 3 are qualitatively very similar to the ones reported by Federal Reserve Bank of New York (2019), which are based on the Engelberg et al. (2009) method.

6 Comparing subjective and objective uncertainty

It is often relevant to ask whether a person's subjective uncertainty is in line with an objective measure of uncertainty. In particular, miscalibrated probabilistic expectations (with subjective uncertainty exceeding objective uncertainty or vice versa) may lead to suboptimal decisions in a wide range of situations (see e.g. Ben-David et al., 2013, and the references therein). In the macroeconomic literature, subjective and objective uncertainty ar often called 'ex ante uncertainty' and 'ex post uncertainty' (see e.g. Clements, 2014). This terminology reflects the fact that subjective uncertainty is typically based on forecasts, whereas objective uncertainty is based on subsequent realizations.

Following a recent proposal by Galvao and Mitchell (2019), comparing a forecaster's ERPS to their RPS (on average across several time periods) yields a simple and theoretically appealing comparison of ex ante and ex post uncertainty.[10] We next provide a formal treatment tailored to our setup. To this end, we consider a so-called prediction space setup (Gneiting and Ranjan, 2013) that models the joint distribution of expectations and realizations. We treat the K histogram probabilities $\underline{\mathbf{p}}$ as a random vector, and denote the bin containing the realization by the discrete random variable $\mathbf{k}^* \in \{1, \ldots, K\}$. The sample space

a fully accomplished college degree. Household income is reported in three categories: 'Under 50k', '50k to 100k' and 'Over 100k'. Finally, a measure of the respondents' numeracy and financial literacy is introduced, such that one can distinguish between respondents with high and low skills. Following a widely used approach, five questions in the survey aim to evaluate respondents knowledge of concepts used in financial decision-making such as interest compounding, understanding of inflation and risk diversification. Respondents who give a correct answer to four out of the five questions are categorized as having high numeracy and financial literacy skills.

[10] As Galvao and Mitchell note, measuring ex post uncertainty via a generic measure of forecast performance (such as RPS) generalizes existing approaches such as Clements (2014) that measure ex post uncertainty via the squared errors of mean forecasts.

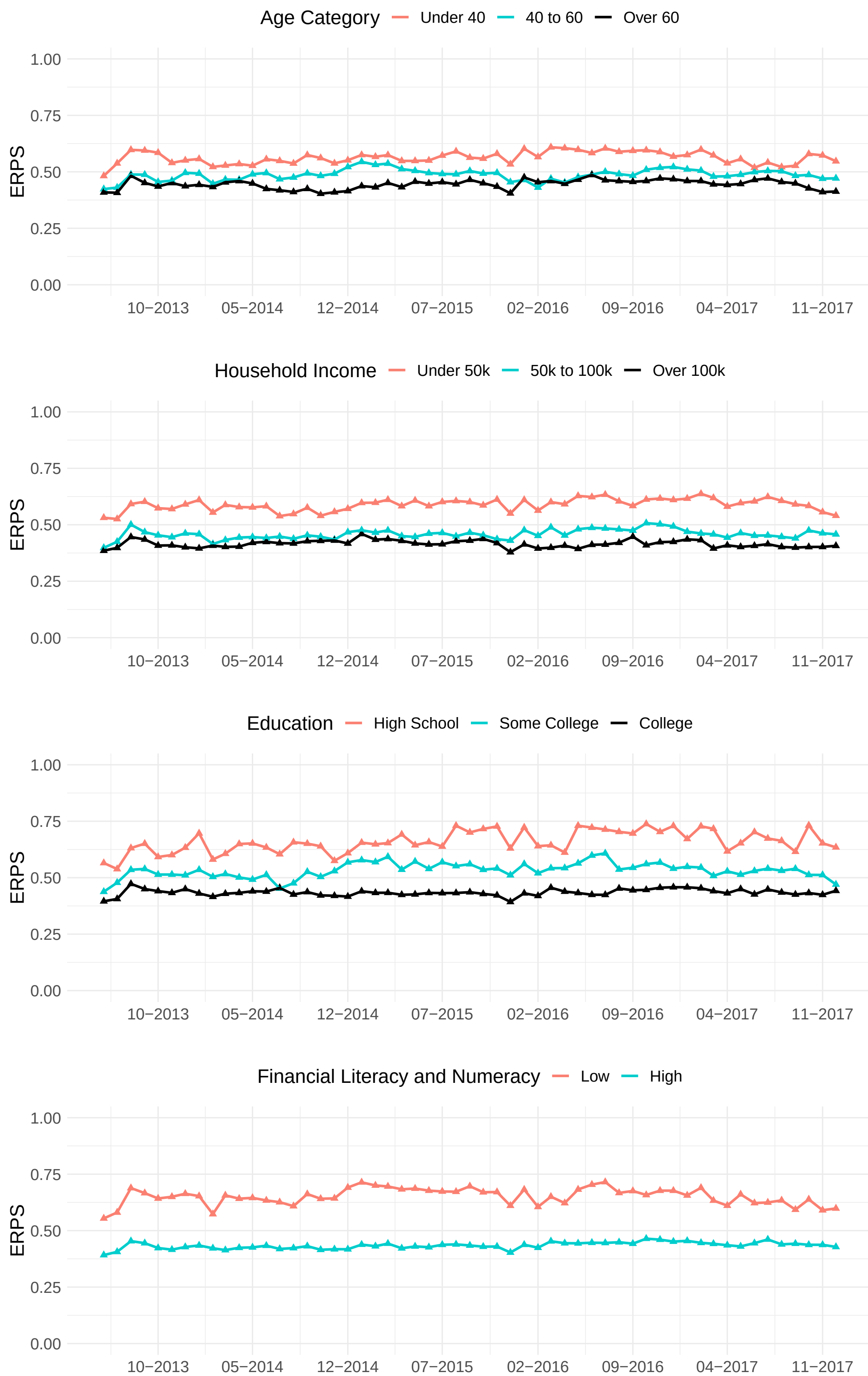

Figure 3: ERPS across sociodemographic groups (inflation, one year ahead).

of interest, Ω, consists of forecast-observation pairs $(\underline{\mathbf{p}}, \mathbf{k}^*)$. We omit time indexes for simplicity; to obtain an intuition, subsequent realizations of $(\underline{\mathbf{p}}, \mathbf{k}^*)$ can be thought of as independent (whereas one would expect contemporaneous dependence between $\underline{\mathbf{p}}$ and $\mathbf{k}^*$, of course).[11] As in Ehm et al. (2016, Section 3.1), let $\mathbb{Q}$ be a probability measure on $(\mathcal{A}, \Omega)$, where $\mathcal{A}$ is a σ-field on Ω. The following result then provides a formal condition under which ex ante uncertainty and ex post uncertainty coincide in expectation.

Assumption 1. *Assume that there is some information set $\mathcal{F} \subseteq \mathcal{A}$ such that*

$$\mathbb{Q}(\mathbf{k}^* = k|\mathcal{F}) = \mathbf{p}_k$$

holds almost surely for $k = 1, \ldots, K$, where $\mathbb{Q}(\mathbf{k}^* = k|\mathcal{F})$ is the true conditional probability that $\mathbf{k}^* = k$ (conditional on the information set $\mathcal{F}$), and $\mathbf{p}_k$ is the kth element of $\underline{\mathbf{p}}$.

Proposition 1. *Under Assumption 1, it holds that $\mathbb{E}(RPS(\underline{\mathbf{p}}, k^*)) = \mathbb{E}(ERPS(\underline{\mathbf{p}}))$.*

Proof. We have that

$$\begin{aligned} \mathbb{E}(\text{RPS}(\underline{\mathbf{p}}, \mathbf{k}^*)) &= \mathbb{E}(\mathbb{E}(\text{RPS}(\underline{\mathbf{p}}, \mathbf{k}^*)|\mathcal{F})) \\ &= \mathbb{E}(\sum_{k=1}^{K} \mathbf{p}_k \ \text{RPS}(\underline{\mathbf{p}}, k)) \\ &= \mathbb{E}(\text{ERPS}(\underline{\mathbf{p}})), \end{aligned}$$

where the first equality follows from the law of iterated expectations, the second equality follows from Assumption 1, and the final equality follows from the definition of ERPS. □

Assumption 1 requires that the probability forecast $\underline{\mathbf{p}}$ is correctly specified, in the sense that there is *some* information set relative to which the forecast is optimal. Under this assumption, Proposition 1 states that the RPS and ERPS of $\underline{\mathbf{p}}$ coincide in expectation. As a simple example (loosely following Gneiting et al., 2007, Table 1), let $Y = X + \varepsilon$, where both variables on the right are independently standard normal. Suppose for simplicity that there are only two outcome bins, $r_1 = (-\infty, 0]$ and $r_2 = (0, \infty)$. Consider forecaster A with $\mathbf{p}_1^{\text{A}} = \Phi(-X), \mathbf{p}_2^{\text{A}} = 1 - \Phi(-X) = \Phi(X)$. For forecaster A, Assumption 1 is satisfied with $\mathcal{F} = \sigma(X)$, the sigma algebra generated by X. In line with Proposition 1, it can be shown that the expected RPS and expected ERPS of forecaster A are both equal to 1/6.[12] For a second forecaster B with $\mathbf{p}_1^{\text{B}} = \mathbf{p}_2^{\text{B}} = 0.5$, Assumption 1 is satisfied with $\mathcal{F} = \emptyset$, the empty information set. The expected ERPS and expected RPS of forecaster B are both equal to 1/4, confirming the intuition that

[11] It can be shown that the methodology of comparing ERPS to RPS remains valid under serial dependence in the forecast-observation tuples, as long as their joint process is strictly stationary. See Strähl and Ziegel (2017) for a technical treatment of a prediction space under serial dependence.

[12] In the notation of Proposition 1, it holds that $\mathbb{E}(\text{RPS}(\underline{\mathbf{p}}^{\text{A}}, k^*)) = \mathbb{E}(\text{ERPS}(\underline{\mathbf{p}}^{\text{A}})) = 1/6$.

B's forecast is less informative than A's forecast.

We think that the ideas just sketched are a useful first step toward comparing subjective and objective uncertainty based on the (E)RPS. That said, further questions need to be addressed before applying the comparison to the SCE data, notably relating to the panel structure of the data (with many cross-sectional units and relatively few time periods). We leave these questions for future research.

7 Conclusion

In this paper, we propose a new measure of subjective uncertainty in histogram-type survey probabilities and demonstrate its conceptional and practical advantages. Our methodology is motivated by the Survey of Consumer Expectations (SCE), a novel micro data set covering consumers' probabilistic expectations of macroeconomic variables. In addition to its usefulness in the SCE, we think that the benefits of our measure are relevant in many other applications of histogram-type survey probabilities in labor economics, development economics, and macroeconomics.

References

Armantier, O., W. Bruine de Bruin, G. Topa, W. Van Der Klaauw, and B. Zafar (2015): "Inflation expectations and behavior: Do survey respondents act on their beliefs?" *International Economic Review*, 56, 505–536.

Armantier, O., G. Topa, W. van der Klaauw, and B. Zafar (2017): "An Overview of the Survey of Consumer Expectations," *Federal Reserve Bank of New York Economic Policy Review*, 23, 51–72.

Ben-David, I., E. Fermand, C. M. Kuhnen, and G. Li (2019): "Expectations Uncertainty and Household Economic Behavior," Working Paper, available at `http://public.kenan-flagler.unc.edu/faculty/kuhnenc/RESEARCH/bfkl_v6.pdf` (last accessed: May 28, 2019).

Ben-David, I., J. R. Graham, and C. R. Harvey (2013): "Managerial miscalibration," *The Quarterly Journal of Economics*, 128, 1547–1584.

Boero, G., J. Smith, and K. F. Wallis (2011): "Scoring rules and survey density forecasts," *International Journal of Forecasting*, 27, 379–393.

Brier, G. W. (1950): "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, 78, 1–3.

Caplin, A., M. Dean, and J. Leahy (2017): "Rationally inattentive behavior: Characterizing and generalizing Shannon entropy," National Bureau of Economic Research working paper no. 23652.

Clements, M. P. (2014): "Forecast uncertainty-ex ante and ex post: US inflation and output growth," *Journal of Business & Economic Statistics*, 32, 206–216.

Coibion, O., Y. Gorodnichenko, and S. Kumar (2018): "How do firms form their expectations? New survey evidence," *American Economic Review*, 108, 2671–2713.

Croushore, D. D. (1993): "Introducing: The Survey of Professional Forecasters," *Federal Reserve Bank of Philadelphia Business Review*, 6, 3–15.

Delavande, A. (2014): "Probabilistic expectations in developing countries," *Annual Review of Economics*, 6, 1–20.

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

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016): "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 505–562.

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, 30–41.

Epstein, E. S. (1969): "A scoring system for probability forecasts of ranked categories," *Journal of Applied Meteorology*, 8, 985–987.

European Commission (2019): "European Business Cycle Indicators, 1st Quarter 2019," Technical Paper 031.

Federal Reserve Bank of New York (2019): "Survey of Consumer Expectations," `https://www.newyorkfed.org/microeconomics/sce` (last accessed: May 2, 2019).

Federal Reserve Bank of Philadelphia (2019): "Survey of Professional Forecasters," `https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/` (last accessed: May 28, 2019).

Galvao, A. B. and J. Mitchell (2019): "Measuring Data Uncertainty: An Application using the Bankof England's "Fan Charts" for Historical GDP Growth," Working Paper, available at `https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxhbmFiZ2FsdmFvfGd4OjY0MDBjMzA1YTc0NDE4OWU` (last accessed: May 22, 2019).

GARCIA, J. A. (2003): "An introduction to the ECB's Survey of Professional Forecasters," *ECB Occasional Paper no. 8*.

GLAS, A. (2018): "Five dimensions of the uncertainty-disagreement linkage," Working Paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3295125` (last accessed: May 28, 2019).

GNEITING, T., F. BALABDAOUI, AND A. E. RAFTERY (2007): "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 243–268.

GNEITING, T. AND M. KATZFUSS (2014): "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151.

GNEITING, T. AND A. E. RAFTERY (2007): "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

GNEITING, T. AND R. RANJAN (2013): "Combining predictive distributions," *Electronic Journal of Statistics*, 7, 1747–1782.

GOSSELIN, M.-A. AND M. KHAN (2015): "A Survey of Consumer Expectations for Canada," *Bank of Canada Review*, 2015, 14–23.

KRÜGER, F., S. LERCH, T. L. THORARINSDOTTIR, AND T. GNEITING (2019): "Predictive inference based on Markov chain Monte Carlo output," Working Paper, available at `https://arxiv.org/abs/1608.06802` (last accessed: May 22, 2019).

LIU, Y. AND X. S. SHENG (2019): "The measurement and transmission of macroeconomic uncertainty: Evidence from the U.S. and BRIC countries," *International Journal of Forecasting*, forthcoming.

LÓPEZ-MENÉNDEZ, A. J. AND R. PÉREZ-SUÁREZ (2019): "Acknowledging Uncertainty in Economic Forecasting. Some Insight from Confidence and Industrial Trend Surveys," *Entropy*, 21, 413.

MANSKI, C. F. (2004): "Measuring expectations," *Econometrica*, 72, 1329–1376.

——— (2018): "Survey measurement of probabilistic macroeconomic expectations: progress and promise," *NBER Macroeconomics Annual*, 32, 411–471.

MATHESON, J. E. AND R. L. WINKLER (1976): "Scoring rules for continuous probability distributions," *Management Science*, 22, 1087–1096.

MOORE, D. A., E. R. TENNEY, AND U. HARAN (2015): "Overprecision in judgment," *The Wiley Blackwell Handbook of Judgment and Decision Making*, 182–209.

Rich, R. and J. Tracy (2010): "The relationships among expected inflation, disagreement, and uncertainty: evidence from matched point and density forecasts," *Review of Economics and Statistics*, 92, 200–207.

Roth, C. and J. Wohlfart (2019): "How do expectations about the macroeconomy affect personal expectations and behavior?" Working Paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016052` (last accessed: May 28, 2019).

Schotter, A. and I. Trevino (2014): "Belief elicitation in the laboratory," *Annual Review of Economics*, 6, 103–128.

Selten, R. (1998): "Axiomatic characterization of the quadratic scoring rule," *Experimental Economics*, 1, 43–61.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Strähl, C. and Ziegel (2017): "Cross-calibration of probabilistic forecasts," *Electronic Journal of Statistics*, 11, 608–639.

Trautmann, S. T. and G. van de Kuilen (2014): "Belief elicitation: A horse race among truth serums," *The Economic Journal*, 125, 2116–2135.

A Maximal ERPS and EBS

The ERPS of a distribution $\underline{p}$ is given by

$$\text{ERPS}(\underline{p}) = \sum_{k=1}^{K} P_k(1 - P_k)$$

In matrix notation, let $\underline{p}$ be the $K \times 1$ vector with probabilities p_k, and $\underline{P}$ be the corresponding vector of cumulative probabilities P_k. We have that $\underline{P} = C'\underline{p}$, where C is a $K \times K$ upper triangular matrix with all elements above the main diagonal equal to one, and all diagonal elements equal to one. We can write

$$\text{ERPS}(\underline{p}) = \underline{P}'(\tau - \underline{P}) = \underline{p}'C\tau - \underline{p}'CC'\underline{p},$$

where τ is a $K \times 1$ vector of ones. To find the maximand of the ERPS, we solve the following problem:

$$\arg\max_{\underline{p}} \text{ERPS}(\underline{p}) \text{ such that } \underline{p}'\tau = 1;$$

note that the constraint that probabilities be nonnegative need not be enforced explicitly. Setting up the Langrangian and solving the resulting quadratic problem then shows that the maximand is given by

$$\underline{p}^* = \begin{bmatrix} 1/2, & 0, & \dots, & 0, & 1/2 \end{bmatrix}';$$

note that the second-order condition for a maximum is satisfied since CC' is strictly positive definite. Along similar lines, consider the expected Brier score given by

$$\text{EBS}(\underline{p}) = \sum_{k=1}^{K} p_k(1 - p_k) = \underline{p}'(\tau - \underline{p}).$$

Using standard constrained optimization as above yields that $\text{EBS}(\underline{p})$ is maximized by

$$\underline{p}^{**} = \tau \times (1/K);$$

the latter result is also noted by López-Menéndez and Pérez-Suárez (2019).

B Details on the Engelberg et al. (2009) quantification method

Here we provide details related to the informal discussion in Section 2.2.

Case A: Forecaster uses one or two bins

Following Engelberg et al., we construct isosceles triangles that are completely characterized by their support which we denote by $[a, b]$. The mode of the distribution is located at $c = (a + b)/2$.

In case a forecaster uses only one bin, we use a triangular distribution with support equal to the support of the bin used. This approach, which is recommended by Engelberg et al. (2009, Section 4.1.1), differs from the SCE, which assumes a uniform distribution over the support of the bin (Armantier et al., 2017, Footnote 28).

To discuss the two-bin case, suppose that a forecaster uses two adjacent bins, $[l, m)$ and $[m, r)$, with $l < m < r$, probability mass α in the left bin $[l, m)$, and probability mass $(1 - \alpha)$ in the right bin $[m, r)$.

If $\alpha < 1/2$, we set $b = r$ and

$$a = m - \frac{(r - m)\ (\alpha + \sqrt{2\alpha})}{2 - \alpha}.$$

Hence the triangular distribution satisfies $\mathcal{T}([a, m)) = \alpha$ and $\mathcal{T}([m, r)) = 1 - \alpha$, where the notation $\mathcal{T}(I)$ indicates the probability mass assigned to the interval I. In that sense, the triangular distribution matches the bin $[m, r)$ containing more probability mass. As $\alpha \to 1/2, a \to m - (r - m)$, such that the fitted triangle is symmetric around m, and the triangle's base length is $2\ (r - m)$.

If $\alpha \geq 1/2$, we set $a = l$ and

$$b = m + \frac{(m-l)\ \left(1 - \alpha + \sqrt{2\ (1-\alpha)}\right)}{1+\alpha}.$$

Hence it holds that $\mathcal{T}([l, m)) = \alpha$ and $\mathcal{T}([m, b)) = 1 - \alpha$, i.e. the triangular distribution matches the bin $[l, m)$ containing more probability mass. For $\alpha = 1/2$, the fitted triangle is symmetric around m, with base length equal to $2\ (m - l)$.

There are two scenarios that are not covered by the preceding description:

- The forecaster uses two non-adjacent bins such as $[0, 2)$ and $[4, 8)$.
- The forecaster uses one or two bins, including one of the outer bins (i.e., $p_1 > 0$ or $p_K > 0$).

The method by Engelberg et al. (2009) does not prescribe a solution for the former scenario. In the latter scenario, any solution would seem to depend on an arbitrary choice of support limit. In our simulation analysis of Section 4, we hence drop observations from either of the two scenarios in order not to distort our findings on the quantification method.

Case B: Forecaster uses three or more bins

If the forecaster uses three or more bins, Engelberg et al. propose to fit a generalized Beta distribution given by

$$\begin{aligned} F_{\text{gBeta}}(x; a, b, l, r) &= \begin{cases} 0 & x \leq l, \\ \frac{1}{B(a,b)} \int_l^x \frac{(u-l)^{a-1}(r-u)^{b-1}}{(r-l)^{a+b-1}} du & l < x \leq r, \\ 1 & x > r, \end{cases} \qquad (5) \\ B(a, b) &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \\ \Gamma(a) &= \int_0^\infty u^{a-1} \exp(-u)\ du. \end{aligned}$$

Instead of the limits 0 and 1 of the regular Beta distribution, F_{gBeta} entails flexible left and right limits $l, r \in \mathbb{R}$ with $l < r$. The two shape parameters $a, b \in \mathbb{R}_+$ play the same role as in regular Beta distributions. Engelberg et al. impose the constraint that $a > 1$ and $b > 1$ in order to obtain a unimodal shape, which seems plausible in the present context.

In order to fit the distribution at (5) to a vector of histogram probabilities $\underline{p}$, Engelberg et al. propose to fix the limits l and r at the endpoints of the bins that are being used. If one or both of the two outer bins are being used, the authors propose to treat the limits l and/or r as free parameters to be estimated. That is, l is a free parameter if $p_1 > 0$, and r is a free parameter if $p_K > 0$, where $K = 10$ in the case of the SCE. Following Armantier et al. (2017, Appendix C), we impose the constraint that $l > -38$ and that $r < 38$ when estimating l and/or r. We further impose that $l < -12$ and $r > 12$, as is logically required by the SCE's bin design. The shape parameters a and b are estimated in either case. In the most general case where l and r are both estimated, the fitting problem is thus given by

$$\max_{\substack{a>1, b>1,\\ -38<l<-12,\\ 12<r<38}} \sum_{k=1}^{K} \left[F_{\text{gBeta}}(x_k; a, b, l, r) - P_k\right]^2,$$

where x_k is the right endpoint of the kth histogram bin, and $P_k = \sum_{j=1}^{k} p_j$ is the cumulative probability of the first k bins.

C Additional simulation results

Design of Section 4.1

	B1	B2	B3	B4
B1	1.000	0.922	0.925	0.925
B2	0.922	1.000	0.959	0.953
B3	0.925	0.959	1.000	0.995
B4	0.925	0.953	0.995	1.000

Design of Section 4.2

	N	L	R
N	1.000	0.997	0.995
L	0.997	1.000	0.993
R	0.995	0.993	1.000

Table 4: Correlation of mean forecasts computed via the Engelberg et al. (2009) method across simulation designs. The top and bottom panel is analogous to Table 2 and 3, respectively.

D Additional figures

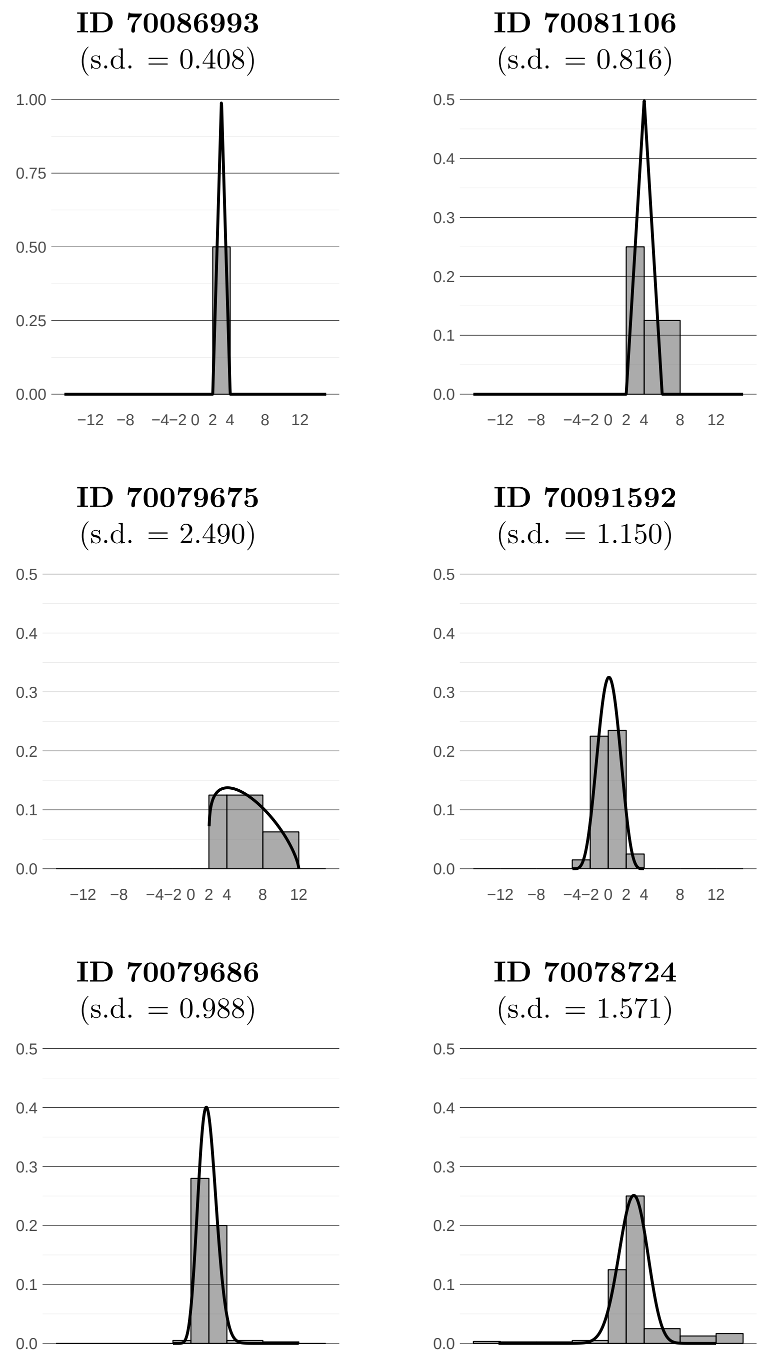

Figure 4: Illustration of probabilistic inflation expectations from the December 2017 wave of the SCE. The figure is analogous to Figure 1, except that it adds fitted continuous distributions (Engelberg et al., 2009) and indicates the standard deviation (s.d.) instead of the ERPS for each panel.

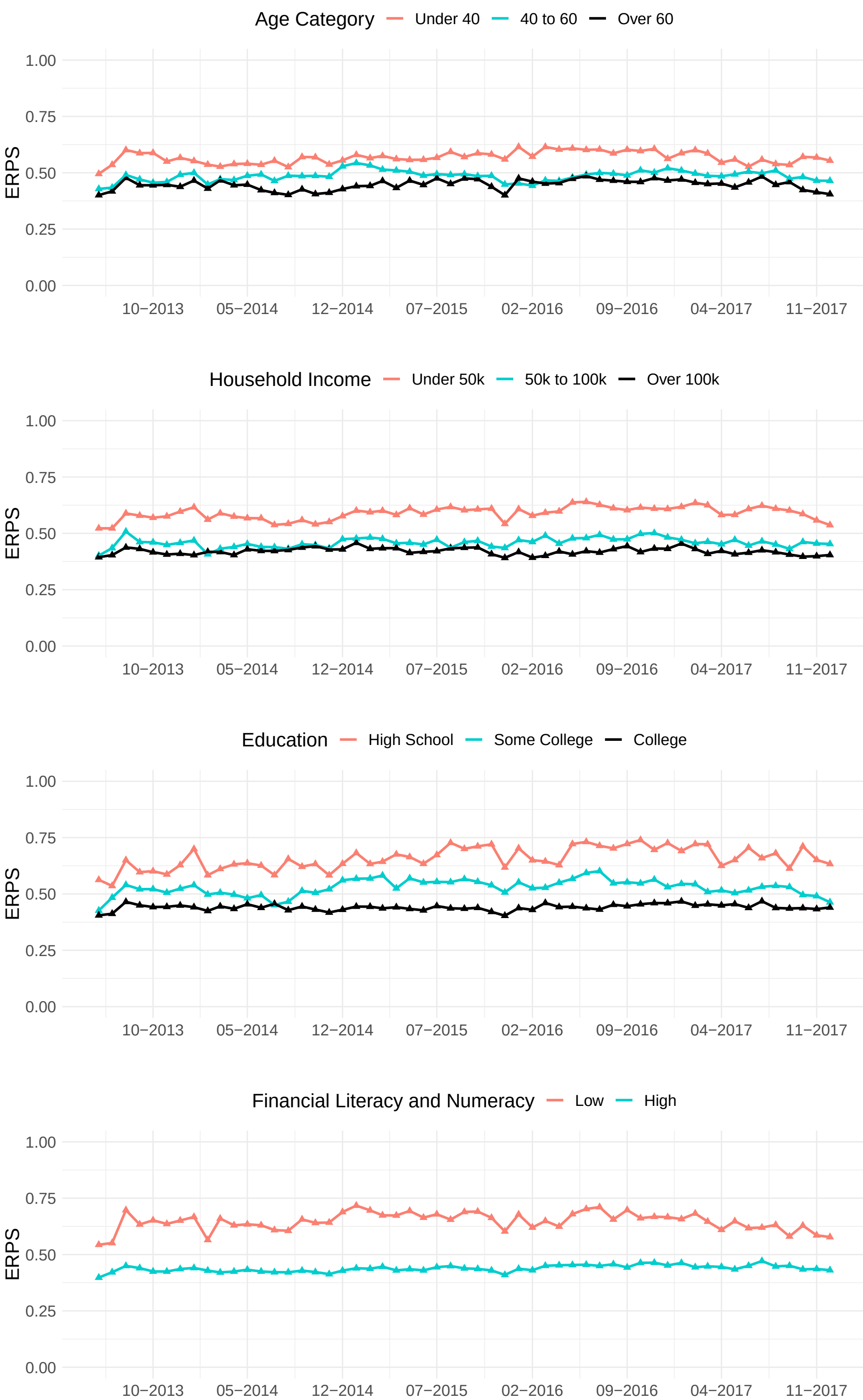

Table 5: ERPS across sociodemographic groups (inflation, three years ahead).

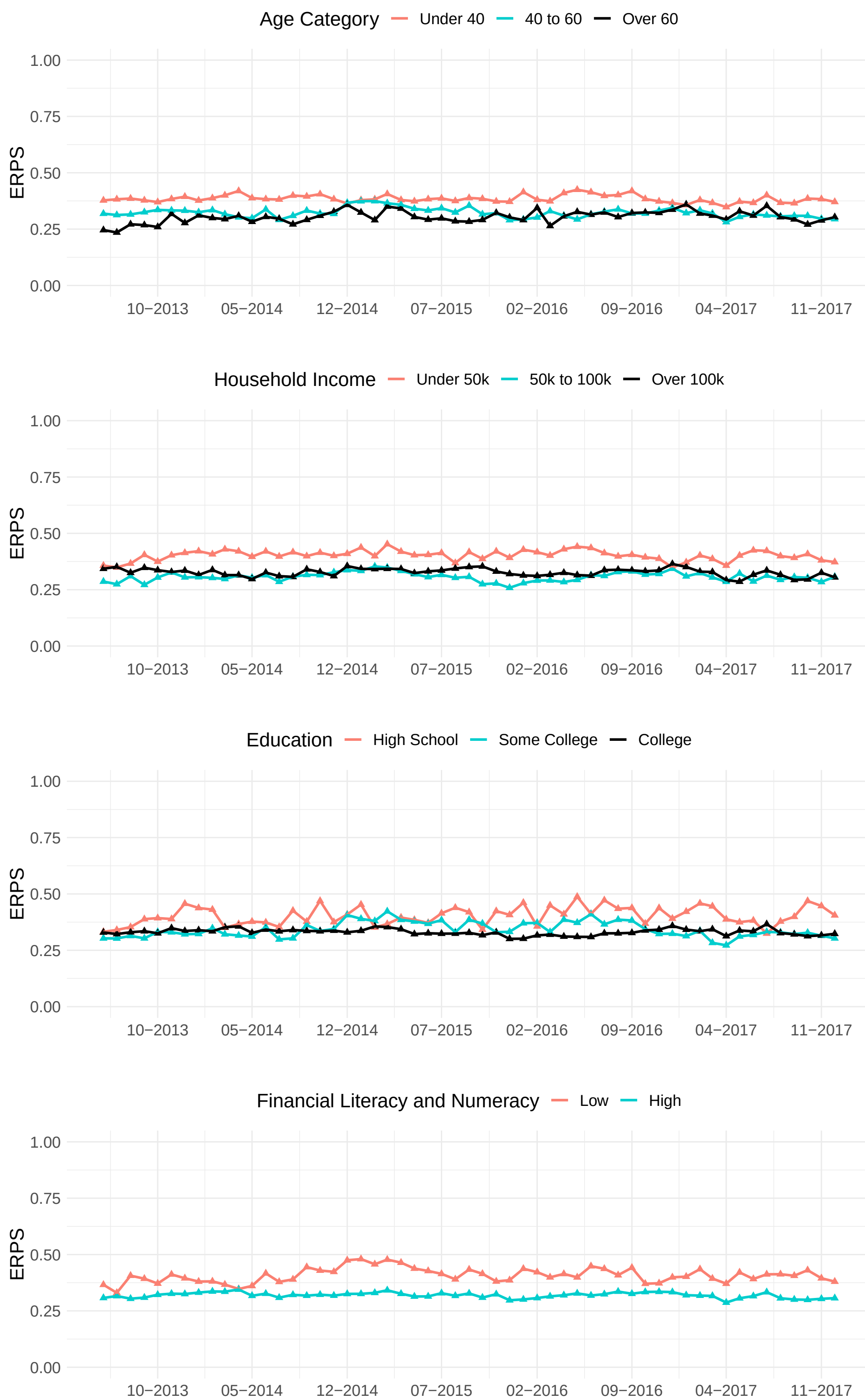

Table 6: ERPS across sociodemographic groups (pers. earnings, one year ahead).

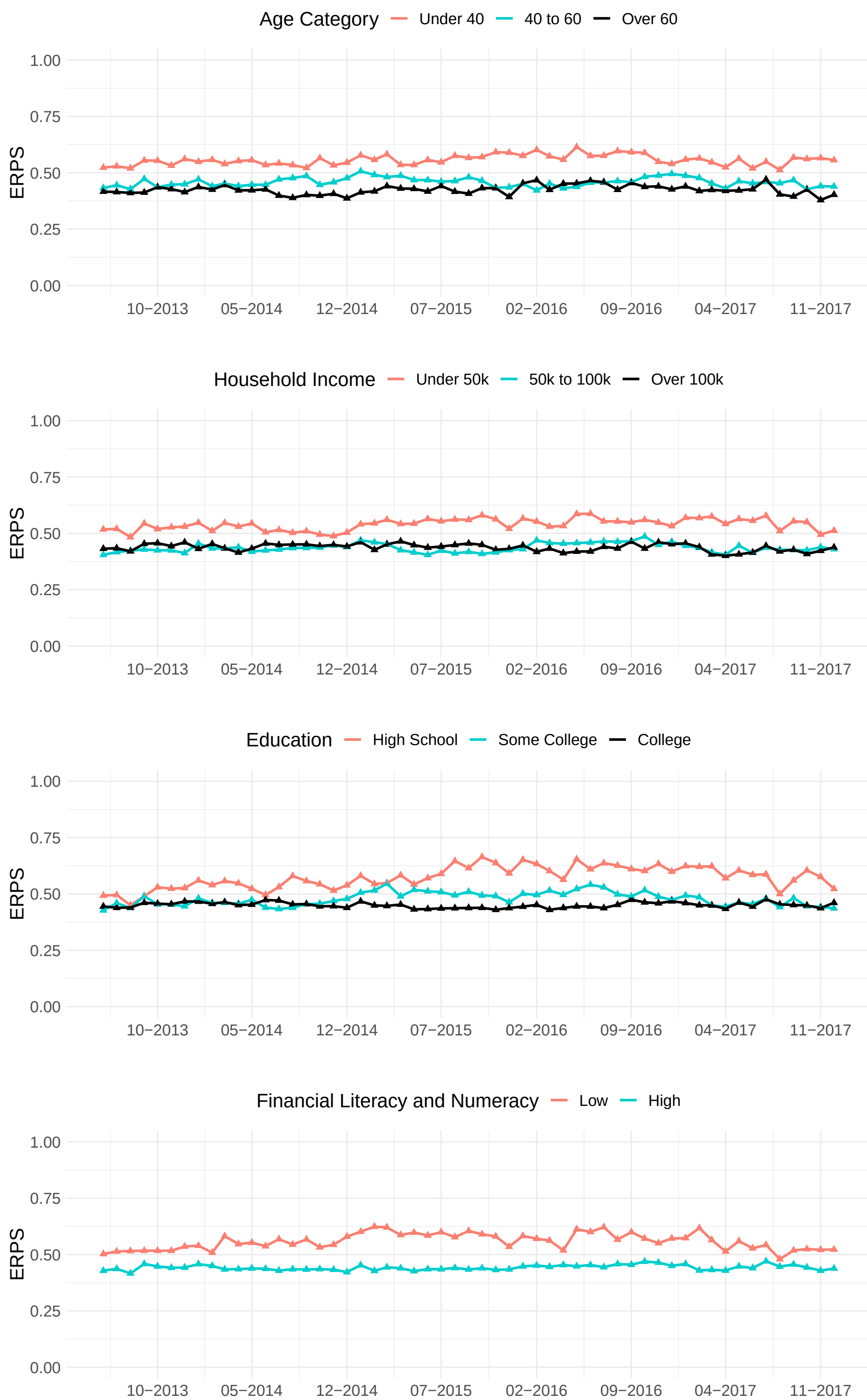

Table 7: ERPS across sociodemographic groups (house prices, one year ahead).