ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Quantifying subjective uncertainty in survey expectations

Fabian Krüger *, Lora Pavlova

Karlsruhe Institute of Technology, Germany

ARTICLE INFO

Keywords:
Survey expectations
Uncertainty
Forecasting
Proper scoring rules
Macroeconomics

ABSTRACT

An increasing number of household and firm surveys ask for 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 robust, trivial to implement, requires no functional form assumptions, and is well-defined for all logically possible probabilities. These advantages are particularly relevant when analyzing microdata from extensive consumer surveys. We illustrate the new measure using data from the Survey of Consumer Expectations.

1. Introduction

Expectation uncertainty matters in economics. Consumers who experience high inflation uncertainty, especially during economic turmoil, increase their savings (Armantier et al., 2021). Uncertain firms tend to respond less to monetary or fiscal policy (Bloom, 2009). Monitoring inflation expectations and the associated uncertainty may help recognize early signs of eroding central bank credibility or de-anchoring of inflation expectations (Grishchenko, Mouabbi, & Renne, 2019); central banks are paying increasing attention to consumer and firm expectations for this purpose (ECB, 2019). Subjective uncertainty also features prominently in theoretical models of expectation formation, such as rational inattention (Mackowiak & Wiederholt, 2009; Sims, 2003).

There is hence much interest in measuring uncertainty, both at the level of the aggregate economy (e.g. Baker, Bloom, & Davis, 2016; Carriero, Clark, & Marcellino, 2018) and at the level of individual persons or firms. In the present paper, we propose a new measure of individual-level uncertainty based on reported subjective probabilities. Such a measure is an important input to studies considering the determinants or the consequences of subjective uncertainty. See, for example, Coibion, Gorodnichenko, and Kumar (2018) for an analysis of firms' expectations, Ben-David, Fermand, Kuhnen, and Li (2019) for a household finance perspective, and Clements, Rich, and Tracy (2023) for an overview of macroeconomic expert forecasts.

Manski (2004, 2018) review many economic surveys in which participants assess the probability of a variable falling into various outcome ranges. In macroeconomics, the Survey of Professional Forecasters (SPF; Croushore, 1993) and its European counterpart (Garcia, 2003) are popular data sources covering expert forecasts. Furthermore, several surveys address the probabilistic expectations of consumers and firms. Examples include the Survey of Consumer Expectations (SCE) launched by the Federal Reserve Bank of New York (Armantier, Topa, van der Klaauw, & Zafar, 2017), a similar initiative by the Bank of Canada (Gosselin & Khan, 2015), and the firm survey by Coibion et al. (2018). These data on probabilistic expectations promise to shed new light on consumers' uncertainty, complementing more traditional surveys using point expectations. The latter do not contain direct information about uncertainty. However, Binder (2017) utilizes a rounding pattern in the point forecasts data, namely respondents reporting multiples of five, to construct a measure of individual uncertainty.

* Correspondence to: Department of Economics and Management, Karlsruhe Institute of Technology, Germany.
E-mail address: fabian.krueger@kit.edu (F. Krüger).

https://doi.org/10.1016/j.ijforecast.2023.06.001

Please cite this article as: F. Krüger and L. Pavlova, Quantifying subjective uncertainty in survey expectations. International Journal of Forecasting (2023), https://doi.org/10.1016/j.ijforecast.2023.06.001.

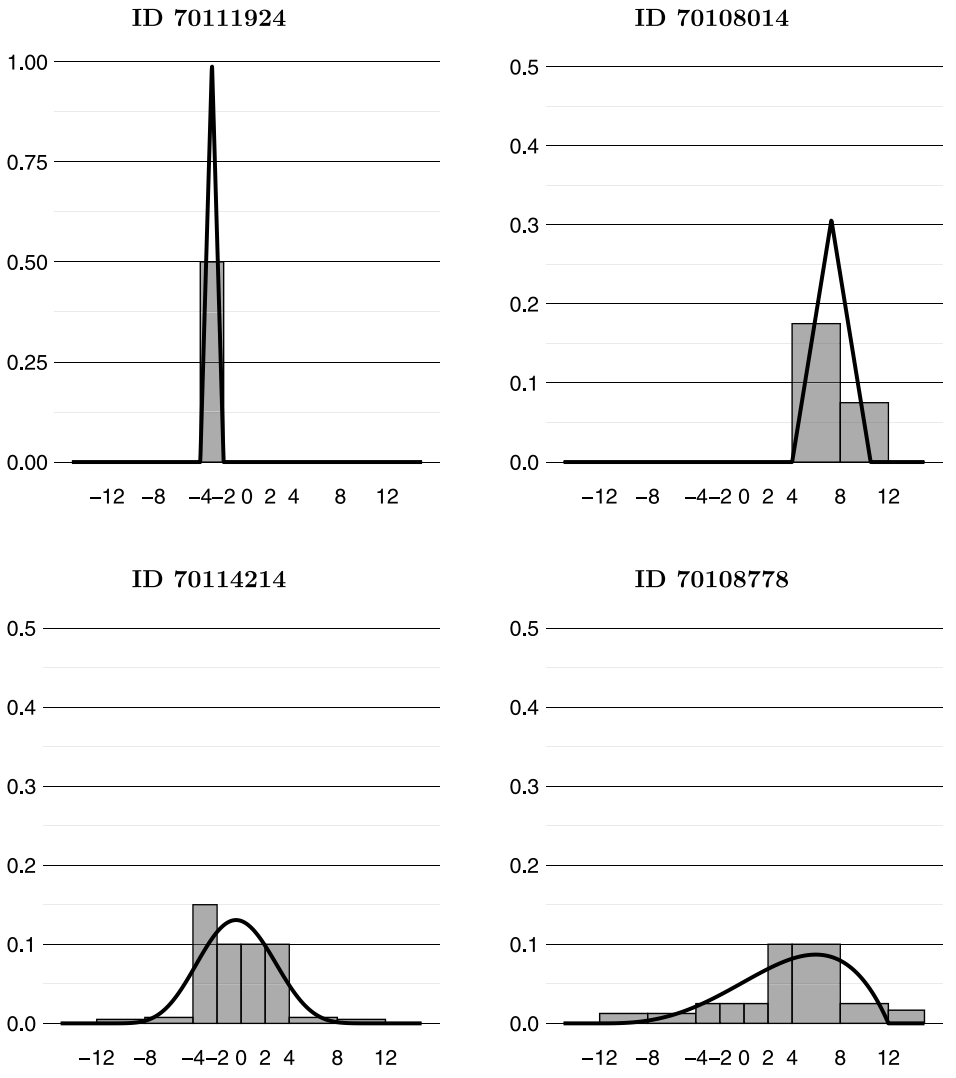

Fig. 1. Illustration of probabilistic inflation expectations from the April 2020 wave of the SCE. The area of a rectangle corresponds to the subjective probability of the corresponding outcome range. For example, in the bottom left panel, the probability for an outcome between 2 and 4 equals $2 \times 0.1 = 0.2$. Solid lines indicate fitted probability density functions via the EMW method.

Fig. 1 illustrates subjective probability distributions ('histograms') from the April 2020 wave of the SCE.[1] Each survey participant provides probabilities for various outcome ranges ('bins') of next year's inflation rate, as represented by the horizontal axis. The SCE contains a substantial share of responses using only one or two bins. Such responses, called 'sparse histograms,' are made by roughly a third of the SCE participants. Sparse histograms pose a challenge for existing measures of individual uncertainty (notably Engelberg, Manski, & Williams, 2009, henceforth EMW), which are based on fitting a parametric distribution. For sparse histograms, fitting a flexible distribution is not possible, and a simple triangular shape is commonly used instead (see the two examples in the top row of Fig. 1).

Motivated by the SCE data, we propose a new uncertainty measure that is transparent, trivial to implement, and well-defined even for sparse histograms. 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. Our proposed 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 new measure as ERPS, for Expected Ranked Probability Score.

The remainder of this paper is structured as follows. Section 2 summarizes some stylized facts of the SCE probabilities. Section 3 describes existing methods for quantifying uncertainty. Section 4 develops the ERPS, detailing

[1] Source: *Survey of Consumer Expectations*©,2013–2023 Federal Reserve Bank of New York (FRBNY). The *SCE* data are available without charge at http://www.newyorkfed/microeconomics/sce and may be used subject to license terms posted there. FRBNY disclaims any responsibility for this analysis and interpretation of *Survey of Consumer Expectations* data.

Table 1
Summary statistics on the number of bins used in the SCE (January 2014 to March 2020 waves) and SPF (2014:Q1 to 2020:Q1 waves); n denotes the total number of responses. We exclude histograms that do not sum to one (less than 0.4% of responses in both surveys).

	n	Share of respondents using one bin	two bins	outer bin(s)	Mean nr. of bins
	SCE				
Average Home Price	85155	16.3	16.0	39.1	4.2
Inflation (one-year)	97019	12.6	17.3	38.9	4.4
Inflation (three-year)	97213	13.1	17.9	38.5	4.4
Personal Wage	65240	26.8	24.1	28.3	3.3
	SPF				
Inflation (GDP def.)	843	2.0	13.6	15.5	4.5
GDP	875	3.1	19.3	6.5	4.5
Inflation (CPI)	842	1.3	14.4	12.7	4.6
Inflation (PCE)	803	0.9	15.4	13.4	4.6
Unemployment	834	8.9	34.4	52.3	3.1

its advantages as mentioned above. Sections 5 and 6 study the behavior of the ERPS for simulated and empirical data, respectively. Section 7 concludes. The appendix contains details, proofs, and additional results.

2. Subjective probabilities in the SCE data

The SCE is conducted monthly with a sample size of about 1,300 respondents per month. The core module of the SCE asks, among others, for subjective probabilities of various outcome ranges, covering three variables: the inflation rate 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 (2020), the relevant question codes are Q9 and Q9c (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 depicted in the horizontal axis labels of Fig. 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]

Table 1 compares the SCE to expert forecasts in the SPF in terms of response behavior. The table's upper panel presents summary statistics on the number of histogram bins SCE participants used (the number of bins containing strictly positive probability mass). We focus on the period from January 2014 to March 2020 for comparability to the SPF (see below). For inflation and the average home price, around 30% of the participants uses one or two bins ('sparse histograms'). For personal earnings, roughly half of the participants use one or two bins. The mean number of bins used is higher for inflation and the average home price (4.2 − 4.4) than personal earnings (3.3). Finally, over a quarter of the participants use one or both outer bins corresponding to the intervals $(-\infty, -12]$ and $(12, \infty)$.

The lower panel of Table 1 presents analogous statistics for the SPF. The SPF questions are similar in design to those of the SCE, except that the two surveys use different numerical ranges for the histogram bins. While the SPF's bin definitions have been adapted over time (Federal Reserve Bank of Philadelphia, 2022), they are constant over the period reported in Table 1. 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, there are some 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, for the inflation variables, the share of participants using at least one outer bin is much lower in the SPF than in the SCE. Unemployment is the only SPF variable in Table 1 for which participants actively use an outer bin. However, this finding appears quite distinctive and is driven by a mismatch between the SPF's bin definitions and the empirical unemployment rate during the sample period considered in Table 1.[3]

Given its large sample size and the empirical patterns just reported, the SCE necessarily contains some histograms with non-standard shapes that are hard to capture by parametric distributions. Examples include distributions with multiple modes, distributions with 'holes' (strictly positive probability assigned to non-adjacent bins), or substantial probability mass in one or both outer bins. These features call for simple and robust methods that quantify the uncertainty in any possible histogram.

[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 is arbitrary – as is any choice in that regard – but seems unlikely to be of empirical relevance.

[3] As documented by Federal Reserve Bank of Philadelphia (2022), the bins for unemployment range from 'less than 4%' to 'more than 9%' for the sample period in question. Given that the actual US unemployment rate was close to or below 4% during much of the second half of the sample period, survey participants' use of the left outer bin seems empirically plausible. In retrospect, the SPF's bin definitions seem at odds with the empirical unemployment rate. Indeed, the SPF bin definitions were changed from 2020:Q2 onwards, reflecting a wider range of unemployment outcomes.

3. Existing uncertainty measures

Survey probabilities, as in Fig. 1, 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 impossible to compute each participant's subjective mean or variance. In the following, we briefly review two methods that use parametric assumptions to account for the missing information.

3.1. Distribution fitting

Following earlier work by Dominitz and Manski (1997), Engelberg et al. (2009, EMW) propose to fit a continuous distribution to the histogram probabilities. Their choice of continuous distribution depends on the number of histogram bins being used: EMW propose fitting a simple triangular distribution if the histogram is sparse and fitting a flexible generalized Beta distribution if the forecaster uses three or more bins. If the forecaster uses the leftmost bin (left limit of $-\infty$) or rightmost bin (right limit of $+\infty$), EMW propose treating the limits of the distribution's support as a free parameter. We provide details on the EMW method in Appendix A.1. The method is used to derive uncertainty measures that are reported in official SCE publications such as Armantier et al. (2017), and are made available for download by Federal Reserve Bank of New York (2020).

The EMW method provides a full analytical distribution from which any feature of interest (such as subjective measures of location, spread, or tail risk) can be computed. However, this wealth of information comes at a cost. First, choosing a particular parametric distribution seems hard to justify for sparse histograms and is potentially restrictive even for dense histograms. For example, the generalized Beta distribution cannot accommodate multimodal histograms, which may be empirically relevant in some situations. (In principle, the generalized Beta distribution could accommodate two modes at the left and right end of the support. However, this type of bimodality seems empirically implausible, and Engelberg et al. propose to exclude it when fitting the distribution. See Appendix A.1 for details.) Second, the approach entails a discontinuity when moving from a histogram with two bins (approximated via a triangular distribution) to one with three bins (approximated via a generalized Beta distribution). Finally, practical implementation requires judgmental choices pertaining, e.g., to parameter limits imposed in numerical optimization or to the handling of certain 'undefined' cases that are not covered by EMW's proposal (because they did not or rarely occur in their SPF data) but that inevitably occur in large data sets like the SCE. Such implementation choices may reasonably be made differently by different authors. Full reproducibility hence requires careful documentation of all choices.

For the SPF data, the drawbacks of the EMW method arguably play a minor role since both the share of sparse histograms and the share of 'undefined' cases are small. This observation explains the widespread and successful use of the EMW method for the SPF and similar data sets. By contrast, given the properties of the SCE discussed above, the EMW method seems less well-adapted to large-scale consumer surveys.

3.2. Mass-at-midpoint method

The mass-at-midpoint (MAM) method (see Glas, 2020, and the references therein) assumes that the subjective distribution is discrete, with a point mass at $\{m_k\}_{k:p_k>0}$, where m_k denotes the midpoint of bin $k = 1, \ldots, K$. Hence the method assumes point mass at the subset of bins that receive nonzero probability. Under this assumption, the subjective mean and standard deviation can easily be computed. An advantage of this method is that it can be applied irrespective of the number of bins used. In particular, it avoids the discontinuity inherent in the EMW method. A disadvantage of the MAM method arises whenever the participants use one of the outer bins (i.e., whenever $p_1 > 0$ or $p_{10} > 0$). In this case, the subjective mean and standard deviation depend on the endpoints of the outer bins, which are not specified by the survey design and for which assumptions seem hard to justify. This disadvantage is especially relevant for the SCE, where about one-third of the participants use at least one outer bin. We provide evidence on this aspect in Appendix A.4.

4. A new approach to quantifying uncertainty in survey histograms

4.1. General idea: Quantifying uncertainty via entropy

We treat each survey response as a vector of probabilities $\underline{p} := (p_1, \quad p_2, \quad \ldots, \quad p_K)'$, where p_k denotes the subjective probability that the outcome is within the interval r_k that defines the range of bin k. 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 Fig. 1.

Our proposed measure of uncertainty is based on the concept of entropy. Informally, if the entropy of a 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, $\underline{p}$ corresponds to high uncertainty. Vice versa, under a low-entropy distribution $\underline{p}$, large forecast errors are unlikely, and hence low entropy corresponds to low uncertainty.

More formally, entropy relates to strictly proper scoring rules (Gneiting & Raftery, 2007). In economics, scoring rules are commonly used for eliciting beliefs in experiments (Schotter & Trevino, 2014) and for evaluating probabilistic forecasts (e.g. Boero, Smith, & Wallis, 2011). In a discrete setup, scoring rules are functions of the form $S(\underline{p}, k^*)$ 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 choice, 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 $\underline{p}$ (conditional on their information set); see Gneiting and Katzfuss (2014, Section 3.1.1) for a formal definition. The function

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

is called the entropy function associated with the scoring rule S (e.g. Gneiting & Raftery, 2007, Section 2.2). We propose to use this function to measure the subjective uncertainty in a probabilistic survey forecast $\underline{p}$.

4.2. Expected ranked probability score (ERPS)

As our preferred choice of scoring rule S, we consider the ranked probability score (RPS; Epstein, 1969):

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

where $P_k = \sum_{j=1}^{k} p_j$ is the cumulative probability of the first k bins. 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 of the underlying variable than $k^*=2$.[4] Thus, the RPS rewards forecasters who put much probability mass into bins equal to or close to the realizing bin k^*. For example, if a forecaster places unit probability mass on the third bin, then $k^*=2$ yields a lower (i.e., better) RPS than $k^*=1$. Boero et al. (2011) persuasively argue that this feature of the RPS is well in line with survey histograms, and propose to use it for evaluating the histograms' predictive accuracy.

The entropy function of the RPS is given by

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

The latter equation, our proposed uncertainty measure, is trivial to compute from the histogram probabilities.

Since it attaches only an ordinal but not a numerical interpretation to the bins, the ERPS at (1) does not depend on the bins' outcome ranges or the (unknown) distribution of probability mass within each bin. The ordinal interpretation renders parametric assumptions obsolete and explains the simplicity and robustness of the ERPS. For example, the ERPS easily accommodates sparse or multimodal histograms. A drawback of the ordinal interpretation is that the ERPS is not comparable across different bin definitions, such as design A involving ten bins of length one and design B involving five bins of length two covering the same interval. This concern may be relevant if the bin definitions must be adapted over time to account for changes in the distribution of the predictand. Such redefinitions occurred several times for the SPF since it was launched in 1968 (see Federal Reserve Bank of Philadelphia, 2022). However, the concern is less relevant for the SCE, whose probability ranges have remained unchanged since its start in 2013 and have also been adopted by many other consumer surveys such as the Bundesbank Online Panel (Deutsche Bundesbank, 2022).

[4] In our empirical analysis based on the SCE's bin definitions, the first bin, $k=1$, ranges from $(-\infty,-12]$, the second bin, $k=2$, ranges from $(-12,-8]$, and similarly for the other bins. The last bin, $k=10$, represents outcomes in the range $[12,\infty)$.

4.3. Comparison to other entropy-based measures

Here we relate the ERPS to entropy functions for two other popular scoring rules. The logarithmic score (LS; Good, 1952) and Brier score (BS; Brier, 1950) are given by

$$\mathrm{LS}(\underline{p}, k^*) = -\log p_{k^*}$$

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

where $\mathbb{I}_{k=k^*}$ is an indicator function that equals one if $k=k^*$, and equals zero otherwise. Their respective entropy functions are given by

$$\mathrm{ELS}(\underline{p}) = -\sum_{k=1}^{K} p_k \log p_k.$$

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

The ELS was famously developed by Shannon (1948) and is typically called 'Shannon Entropy'. In economics, it plays a key role in the theory of rational inattention (Sims, 2003). Rich and Tracy (2010) use the ELS to measure uncertainty in the SPF histograms. The EBS is much less widely used, with the interesting exception of López-Menéndez and Pérez-Suárez (2019), who quantify uncertainty in (aggregate) tendency surveys.

The BS and LS are designed for multinomial random variables; the outcome categories $k^* \in \{1,\ldots,K\}$ are considered interchangeable. Hence the EBS and ELS are invariant to permutations of the histogram 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 $\underline{p}_a$.

ELS and EBS are both maximized by the vector

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

where τ is a $K \times 1$ vector of ones (see López-Menéndez and Pérez-Suárez 2019, Shannon 1948). Hence flat probabilities represent maximal uncertainty, as seems natural in a multinomial setup. By contrast, we show in Appendix A.2 that the maximal ERPS is attained for the vector

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

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 $\mathrm{RPS}(\underline{p}^*, k)$, since $\underline{p}^*$ places no probability mass on the neighboring bins.

While the RPS accounts for ordering the outcome categories $k^* \in \{1, \ldots, K\}$, it does not reflect information about the width of the corresponding histogram bins. This information requires a numerical, rather than just ordinal, interpretation of the outcome categories. As mentioned, the numerical interpretation is challenging in the present context. Nevertheless, relating the (E)RPS to entropy-based uncertainty measures for numerical outcomes is interesting. We provide such a comparison in Appendix A.3.

5. Simulation studies

This section compares the ERPS to the EMW and MAM methods of quantifying survey uncertainty.

5.1. Survey histograms as noisy realizations

Our first simulation design views survey histograms as a noisy realization of an underlying true continuous distribution. Survey noise could arise, for example, from participants' limited attention when answering the survey. In the following, we analyze which histogram-based uncertainty measure is most closely aligned with the uncertainty of the true distribution. We implement this idea via the following design:

- Draw an independent sample of size n from a random variable X with continuous distribution F
- Set the 'survey' probability for the jth bin equal to

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \in \text{bin}_j),$$

  where $\mathbf{1}(A)$ is the indicator function of the event A, x_i is the ith realization of the simulated sample, and bin_j is the jth SCE interval as defined in Section 2, with $j \in \{1, 2, \ldots, 10\}$.
- Denote the corresponding true probability for bin j by

$$p_j = \mathbb{P}(X \in \text{bin}_j) = \int_{\text{bin}_j} \mathrm{d}F(x).$$

We next compute various uncertainty measures based on the simulated survey probabilities $\{\hat{p}_j\}_{j=1}^{10}$ and, possibly, the corresponding bin limits. We then compare these measures to the underlying ground truth measure of uncertainty. Specifically, for the EMW-SD and MAM-SD methods, we compare the estimated standard deviation $\hat{\sigma}$ to σ, the true standard deviation implied by F.[5] Similarly, we compare the interquartile range estimated by the EMW method (EMW-IQR) to the true interquartile range implied by F, and we compare the estimated ERPS (based on the $\hat{p}_j$s) to the true ERPS (based on the p_js).

The degree of noise in the histograms is governed by n. While a small sample size n may entail large deviations between $\hat{p}_j$ and p_j (and, possibly, empty bins $\hat{p}_j = 0$ for some j), each $\hat{p}_j$ converges in probability to p_j as $n \to \infty$. It remains to choose a true distribution F for simulating the data. We consider two variants: First, a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ and σ are the quantified mean and standard deviation associated with a randomly selected SCE histogram drawn from June 2013 to April 2020 waves and requiring that the histogram uses at least two bins. Second, the EMW method's quantified (triangular or generalized Beta) distribution for a randomly selected SCE histogram, again requiring at least two bins. While the first variant is somewhat simpler, the second variant intentionally favors the EMW method because it simulates data that, by construction, are closely in line with its functional form assumptions.

Table 2
Spearman rank correlation between estimated and true uncertainty, across 1000 Monte Carlo simulations.

	Normal		Quantified (triangular or gen. Beta)	
	$n = 20$	$n = 50$	$n = 20$	$n = 50$
EMW-SD	0.97	0.99	0.97	0.99
MAM-SD	0.95	0.97	0.95	0.96
EMW-IQR	0.97	0.98	0.97	0.99
ERPS	0.97	0.98	0.97	0.99

Table 2 summarizes the results, across 1 000 Monte Carlo simulations and for two sample sizes $n \in \{20, 50\}$. Reassuringly, there is generally a high agreement between the estimated and true uncertainty values, with all rank correlations in Table 2 exceeding 0.94. As expected, all methods perform better in the setup with less noise ($n = 50$). While the methods perform very similarly in the Gaussian case, the second variant indicates a modest advantage of the EMW and ERPS methods compared to MAM-SD. Given that the second variant favors EMW uncertainty measures by construction, the good performance of the ERPS confirms its robustness.

5.2. Sparse histograms

As we have argued, a key advantage of the ERPS over the EMW method is that the former requires no case distinction when moving from a sparse histogram (using two bins) to a histogram using three bins. We demonstrate the quantitative relevance of this point in a simulation study based on June 2013 to April 2020 waves of the SCE. For comparability, we focus on participants who use two adjacent bins, none of which is an outer bin in the SCE's histogram design shown in Section 2. We further require that the histogram probabilities sum to one and exceed one percent, which is the magnitude of the perturbation we consider. These selection criteria leave us with 15961 two-bin histograms. For each histogram, 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 each of 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

[5] In both simulation studies, we implement the MAM method by assuming that the left and right end of the histogram's support is given by -16 and 16, respectively. We provide further evidence of this implementation choice in Section 6 below.

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. We choose a perturbation size of one percentage point since it is the smallest size that seems empirically plausible.

For each setup (no perturbation, left perturbation, and right perturbation), we again consider the ERPS, as well as the standard deviation (EMW-SD) and interquartile range (EMW-IQR) of the distribution obtained via the EMW method, and the standard deviation obtained via the MAM method (MAM-SD). Given the small perturbation size, we contend that an uncertainty metric should be robust to the perturbation.[6] To measure the similarity between the perturbed and baseline histograms, we consider the rank correlation between the uncertainty measures and their mean absolute deviation (MAD). Table 3 summarizes the results, indicating that the perturbation significantly impacts the two EMW measures. The rank correlation between the original and perturbed measures can be as low as 0.38, which is remarkable given the small magnitude of the change. Similarly, the mean absolute deviations in the first two rows of Table 3 are considerable, given the mean values of the uncertainty measures reported in the first column. The results further indicate that the impact of the right perturbation is larger than the impact of the 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 in Section 2, the left neighbor of these bins is at (0, 2], whereas the right neighbor is at (8, 12]. Hence, the left perturbation expands the support of the histogram by two units, whereas the right perturbation expands the support by four units. This asymmetry matters here since the (Engelberg et al., 2009) algorithm supports the histogram if only interior bins are used.

MAM-SD is robust to both left and right perturbation, attaining rank correlations close to one and reducing mean absolute deviations by about 50% compared to the EMW method. For ERPS, 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 perturbation size (with $\delta = 0.01$ in our simulation study). Then Eq. (1) yields that

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

i.e., both perturbations lead to an additive increase in ERPS by $\delta\ (1 - \delta)$. Hence, the perturbation affects all histograms in the same way, leading to correlations of one and mean absolute deviations of about 0.01, demonstrating the robustness of the ERPS.

The present simulation experiment aims to quantify the impact of the EMW method's known discontinuity when moving from two to three bins. The two-bin case is empirically relevant for individual-level SCE histograms but less common in other contexts (such as average histograms across many survey participants). When the baseline histogram uses three or more bins, we expect the EMW method to be reasonably robust to the types of small perturbations considered above, as the switch from a triangular to a generalized Beta distribution for quantification does not occur in these situations.

[6] For larger perturbation sizes, it is no longer clear whether the uncertainty measure should be robust to the perturbation or not.

Table 3
First column: Mean of uncertainty measure (without perturbation). Second to fifth column: Rank correlation of uncertainty in perturbed and baseline histograms and mean absolute deviation (MAD) between uncertainty in perturbed and baseline histograms.

	Mean	Left perturbation		Right perturbation	
		Correlation	MAD	Correlation	MAD
EMW-SD	0.75	0.66	0.10	0.38	0.19
EMW-IQR	1.08	0.70	0.13	0.54	0.24
MAM-SD	1.00	0.99	0.05	0.97	0.09
ERPS	0.19	1.00	0.01	1.00	0.01

6. Empirical comparisons

We next compare the ERPS to the EMW and MAM methods based on empirical survey data. For EMW, we focus on the EMW-SD variant; the results based on EMW-IQR are qualitatively identical and are hence omitted for brevity. For MAM, we initially assume that the lower and upper support limits are given by -16 and 16, such that the open bins have the same length as their neighboring closed bins. We then study the impact of this parameter at the end of the section and in Appendix A.4.

Overall, the three uncertainty measures display strong positive associations. For a given survey date and variable, the rank correlation between any two uncertainty measures is at least 0.87 and often as high as 0.95. We next analyze whether respondents who express high uncertainty about inflation also express high uncertainty about house prices and their personal earnings. To this end, we consider the rank correlation coefficient of uncertainty across variables. We consider six pairs of variables and 83 monthly survey waves (June 2013 to April 2020). Fig. 2 illustrates house prices and inflation, indicating that ERPS and MAM generally yield a higher rank correlation than EMW-SD. Similar patterns also hold more broadly: Across all variable pairs and survey dates, the ERPS attains the highest rank correlation in about 64% of all cases. The corresponding shares for MAM and EMW are 34% and 2%. There is hence clear evidence that the ERPS is more consistent across variables than the EMW method. In the absence of a 'ground truth' measure of uncertainty, we cannot tell whether this feature of the ERPS is desirable. However, these findings indicate an interesting and robust difference between both measures.

We further compare the persistence of uncertainty as measured by EMW, ERPS, and MAM. We measure persistence by the rank correlation of uncertainty in two subsequent SCE waves for the subset of participants present in both waves. A small rank correlation may indicate a genuine shift in relative uncertainty from one month to the next (e.g., Anne is more uncertain than Bob in January, whereas Bob is more uncertain than Anne in February). Alternatively, a small rank correlation may reflect noise in the uncertainty measure.

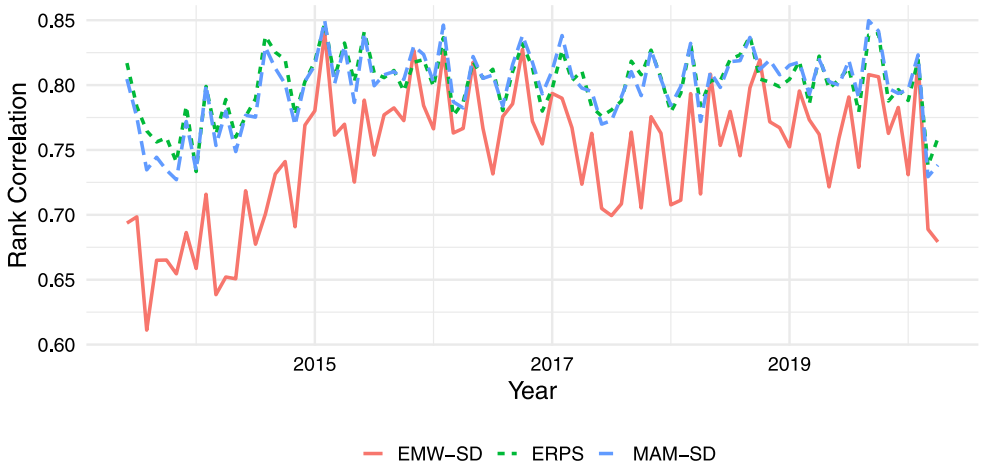

Fig. 2. Rank correlation of subjective uncertainty regarding house prices and inflation (one year ahead). The lines correspond to different measures of uncertainty.

Fig. 3. Rank correlation of subjective uncertainty over two subsequent survey months, based on participants present in both months.

Fig. 3 presents results on the persistence of uncertainty. The observed correlation of all three uncertainty measures is similar for personal earnings. Genuine shifts in relative uncertainty seem particularly plausible for this variable since it is individual-specific and prone to idiosyncratic information updates (such as Anne signing a new labor contract in February). MAM-SD is most persistent overall for house prices and inflation, followed by ERPS and EMW-SD. Similar to the findings across variables, this indicates that MAM-SD and ERPS are less sensitive to small changes in the raw probabilities $\underline{p}$ than EMW-SD.

Until now, all of our (simulation and empirical) results for MAM have assumed lower and upper support limits of -16 and 16. This choice is necessarily judgmental as the histograms yield no information on support limits. One

could argue that support limits of ±16 are too narrow and consider wider limits instead. However, any particular choice of wider limits is arbitrary as well. In Appendix A.4, we provide empirical results on this topic for inflation expectations in the SCE. As shown there, the largest standard deviations in the sample are especially sensitive to the choice of support limits and become very large for wide choices (such as ±38 used in the 'wide' scenario of Appendix A.4). On the other hand, these wide choices are not easily refuted as implausible, for example, when considering the presence of extreme point expectations in consumer surveys.[7]

7. Discussion

This paper introduces the ERPS, a new measure of uncertainty in probabilistic survey expectations. The ERPS is based on an ordinal interpretation of the survey outcome categories, which prevents parametric assumptions and explains its simplicity and robustness. The Engelberg et al. (2009, EMW) method, the current standard for quantifying uncertainty in economic surveys, uses a numerical interpretation of outcome categories instead. The numerical interpretation is more demanding and requires the researcher to make parametric assumptions about unknown aspects of the histogram. In return, it provides a full picture of subjective uncertainty.

We think that a user's choice between the ERPS and the EMW method should depend on the signal-to-noise ratio in the subjective probability data. If this ratio is high, then the EMW method – which is more sensitive to small changes in the probabilities – seems more appropriate. Examples of this situation include average histograms across time or across socio-demographic groups (which may be based on hundreds of individual responses) and perhaps probability assessments by individual expert forecasters. By contrast, the ERPS seems preferable in the context of individual-level probabilities by consumers, such as the ones covered by the SCE. This data type is an innovative source for monitoring and studying the general public's inflation expectations. In particular, microdata from the SCE and similar surveys allow us to analyze the heterogeneity in economic expectations across socio-demographic subgroups of society. Such analyses are highly relevant to study the general public's response to economic policy measures (see, e.g. D'Acunto, Malmendier, & Weber, 2023). Finally, for aggregate measures of uncertainty (obtained, e.g., by computing an individual-level measure of uncertainty and then averaging this measure across survey participants), we typically expect the difference between the EWM method and the ERPS to be limited. Due to averaging, the sensitivity of the EMW method will often be of minor importance in such situations.

Our simulation and empirical results also cover the mass-at-midpoint (MAM) method, which can estimate the standard deviation corresponding to an individual-level histogram. While MAM seems more attractive than the EMW method for two-bin histograms (see Section 5.2), its use of judgmental support limits makes it less attractive for histograms involving outer bins (see Appendix A.4). Compared to the ERPS, MAM enables more detailed information on a forecast histogram (in particular, estimates of its mean and standard deviation) at the cost of making assumptions and implementation choices that are hard to justify rigorously. This trade-off is similar to the trade-off that arises when comparing the ERPS and the EMW method.

Finally, while we have focused on measuring subjective uncertainty by itself, an interesting question is whether subjective uncertainty lines up with measures of realized uncertainty based on expectation errors. This comparison is of economic relevance since over- or underestimating objective uncertainty has possibly severe implications for decision-making (see, e.g. Ben-David, Graham, & Harvey, 2013). In Appendix A.5, we demonstrate that the ERPS can also be used in this context.

[7] Consider, for example, one year ahead inflation expectations in the SCE (variable code Q8v2part2). Across the June 2013 – April 2020 survey waves, 1.3% of the point expectations are −25% or lower, and 5% of the point expectations are +25% or higher.

Declaration of competing interest

Fabian Krüger reports financial support was provided by German Research Foundation. Lora Pavlova reports a relationship with Deutsche Bundesbank that includes: employment.

Acknowledgments

Financial support from the German Research Foundation (DFG) via grant KR 5214/1-1 is gratefully acknowledged. We thank three anonymous reviewers, seminar and conference participants at Heidelberg University, HKMEtrics, Humboldt-Universität Berlin, IWH (Halle), Joint Conference on Household Expectations (Bundesbank - Banque de France), 1st Bergamo Workshop on Statistics and Econometrics, KIT, as well as Konstantin Görgen, Axel Lindner, Malte Knüppel, Simas Kucinskas, Sebastian Rüth and Michael Weber for helpful comments.

Appendix

The appendix provides details on the EMW method (Appendix A.1), proves a claim on the ERPS (Appendix A.2), relates the ERPS to the CRPS (Appendix A.3), investigates the choice of bin limits for the MAM method (Appendix A.4), and sketches a comparison of the ERPS to its realized counterpart (Appendix A.5).

A.1. Details on the EMW method

Here we provide details on our implementation of the Engelberg et al. (2009, EMW) method for quantifying forecast histograms.

Case A: Forecaster uses one or two bins

Following Engelberg et al. (2009, EMW), 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 $c = (a + b)/2$.

If a forecaster uses only one bin, we use a triangular distribution with support equal to that of the bin used. This approach, which is recommended in EMW's Section 4.1.1, differs from the one implemented in the SCE, which assumes a uniform distribution over the support of the bin (Armantier et al., 2017, Footnote 28).

In the case of a forecaster using two adjacent bins, Becker, Duersch, Eife, and Glas (2021) note that the original procedure by EMW may yield counterintuitive triangular fits when applied to survey probability intervals of varying widths (like the SCE). Suppose the two bins with nonzero probability are given by $[L, M)$ and $[M, R)$, and denote the corresponding probabilities by p_L and p_R. Similar to Becker et al., we set $a = L$ if the left interval features weakly higher density, i.e., if $p_L/(M-L) \geq p_R/(R-M)$. Otherwise, we set $b = R$. Unlike Becker et al., we then choose the other endpoint of the isosceles triangle by numerically optimizing the squared difference between the empirical and fitted CDFs. This approach is motivated by the fitting criterion used in the case of three or more bins, described below. In most empirical two-bin cases, a squared difference of zero is attainable, and the numerical solutions coincide with the formulas proposed by Becker et al.. However, exceptions to this situation exist, including the example of a participant placing 30% and 70% probability on the (2,4] and (4,8] bins, respectively.

The preceding description does not cover two scenarios:

- 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 EMW method does not prescribe a solution for the former scenario. In the latter scenario, any solution would seem to hinge on an arbitrary choice of support limit. In our analysis, we drop observations from either of the two scenarios to not distort our findings on the EMW method.

Case B: Forecaster uses three or more bins

If the forecaster uses three or more bins, EMW propose to fit a generalized Beta distribution given by

$$F_{\text{gBeta}}(x; a, b, l, r) = \begin{cases} 0 & x \leq l, \\ \frac{1}{B(a,b)} \int_l^x \frac{(u-l)^{a-1}(r-u)^{b-1}}{(r-l)^{a+b-1}} du & l < x \leq r, \\ 1 & x > r, \end{cases} \quad (2)$$

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

$$\Gamma(a) = \int_0^\infty u^{a-1} \exp(-u)\, du.$$

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. EMW impose the constraint that $a > 1$ and $b > 1$ to obtain a unimodal shape, which seems plausible in the present context.

To fit the distribution at (2) to a vector of histogram probabilities $\underline{p}$, EMW 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

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

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.

R code for implementing the described quantification method is available via the first author's website. We drop a small number of individual survey responses (12, out of 367 728) for which our quantification method leads to excessive numerical challenges, primarily due to a large probability mass in one of the outer histogram bins.

A.2. Maximal ERPS

Here we prove a claim made in Section 4.3 of the paper.

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 Lagrangian and solving the resulting quadratic problem then shows that the maximand is given by

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

note that the second-order condition for a maximum is satisfied since CC' is strictly positive definite.

A.3. Relating the ERPS to probability distributions for numerical outcomes

Summary

The ERPS attaches an ordinal interpretation to the histogram bins and depends only on the vector $\underline{p}$ of bin probabilities. Here we relate this perspective to the unknown probability distributions for numerical outcomes that may underlie a given vector $\underline{p}$. In particular, consider two different probability distributions F_1, F_2, such that both F_1 and F_2 match $\underline{p}$, i.e., they both assign probability p_k to the interval defining bin $k = 1, \ldots, K$. F_1 and F_2 attain the same ERPS. By contrast, uncertainty measures for numerical outcomes will typically assign different levels of uncertainty to F_1 and F_2. In that sense, the ERPS summarizes the uncertainty of all probability distributions F that match $\underline{p}$. Below we discuss this conceptual aspect in more detail, providing explicit results for one particular subclass of distributions that match $\underline{p}$ and for the simplified case of a histogram with bins of equal width. The requirement of equal bin widths is natural and essential. Since the ERPS does not use information on bin length, there is no meaningful way to study the ERPS in a setup where bin length is a relevant parameter.

Details

Consider a discrete random variable X with support $v_1, v_2, \ldots, v_K$, where $v_j \in \mathbb{R}$ for all $j = 1, \ldots, K$, $v_a < v_b$ for $a < b$, $\mathbb{P}(X = v_j) = p_j$, and $\sum_{j=1}^{K} p_j = 1$. We think of X (and its modified versions below) as a draw from the probability distribution that underlies a given survey histogram. The cumulative distribution function (CDF) of X is given by

$$F(x) = \mathbb{P}(X \leq x) = \sum_{j:\ v_j \leq x} p_j.$$

Hence $F(x)$ is a piecewise constant function that satisfies $F(x) = 0$ for $x < v_1$ and $F(x) = 1$ for $x \geq v_K$. Since X is supported on the real line, we can measure its underlying uncertainty using the continuous ranked probability score (CRPS; Matheson & Winkler, 1976), a strictly proper scoring rule. Observe that using the CRPS requires a numerical interpretation of the support of X, in contrast to the ordinal interpretation underlying the ERPS. The expected CRPS (ECRPS), or CRPS entropy, for X is given by

$$\begin{aligned}
\mathrm{ECRPS}(F) &= \int_{-\infty}^{\infty} \mathrm{CRPS}(F, x)\, dF(x) \qquad (3)\\
&= \int_{-\infty}^{\infty} F(x)(1 - F(x))\, dx\\
&= \int_{v_1}^{v_K} F(x)(1 - F(x))\, dx\\
&= \sum_{j=1}^{K-1} (v_{j+1} - v_j) F(v_j)(1 - F(v_j))\\
&= \sum_{j=1}^{K-1} (v_{j+1} - v_j) P_j (1 - P_j), \qquad (4)
\end{aligned}$$

where $P_j = \sum_{l=1}^{j} p_l$ is the cumulative probability of the first j categories. The second equality follows the known properties of the CRPS. See, e.g., Gneiting and Raftery (2007, Section 4.2). The expression at (4) is identical to the expression for the ERPS at (1) if $v_{j+1} - v_j = 1$ for all j, i.e., if all support points of X are exactly one unit apart. We will focus on this case in the following, and we investigate the properties of the CRPS entropy in this setup.

For concreteness, suppose the support of X is given by $v_1, v_2, \ldots, v_K = 0, 1, \ldots, K-1$. Furthermore, for a given integer $n \in \mathbb{N}$ and $s = 1, 2, \ldots, n$, define the shifted random variables $X_s^n = X + s/(n+1)$, and $\mathbb{P}(X_s^n = j + s/(n+1)) = \mathbb{P}(X = j)$ for $j = 0, 1, \ldots, K-1$. Fig. 4 illustrates this construction for $n = 3$ (top panel) and $n = 20$ (bottom panel).

For given n and $\underline{p}$, we next consider the following family of mixture distributions:

$$\mathcal{F}_n(\underline{p}) = \left\{ Z : \mathbb{P}(Z \leq z) = \sum_{s=1}^{n} \omega_s^n F_s^n(z),\ 0 \leq \omega_s^n \leq 1, \sum_{s=1}^{n} \omega_s^n = 1. \right\}.$$

That is, $\mathcal{F}_n(\underline{p})$ collects all distributions that can be constructed as a finite mixture of the random variables $X_1^n, X_2^n, \ldots, X_n^n$, where $F_s^n(z) = F(z - s/(n+1))$ is the CDF of the sth mixture component, and F is the CDF of the discrete random variable X with support points $0, 1, 2, \ldots, K-1$ and associated probabilities $\underline{p}$. An interesting special case arises for $n \to \infty$ and $\omega_s^n = 1/n$ for all $s = 1, 2, \ldots, n$, yielding a piecewise uniform distribution between $[0, 1], (1, 2], \ldots, (K-1, K]$, as alluded to in the bottom panel of Fig. 4. Such a setup is sometimes assumed for quantifying survey histograms (see, e.g. Glas, 2020). Another practically relevant special case arises when n is odd, and $\omega_{(n+1)/2}^n = 1$, i.e., the 'central' mixture component receives a weight of one. In this case, we obtain the distribution assumed by the mass-at-midpoint method, with support at $0.5, 1.5, \ldots, K-0.5$ and associated probabilities $\underline{p}$.

Consider a forecast histogram with bins $[0, 1], (1, 2], \ldots, (K-1, K]$. Then for a given choice of n, each random variable in the family $\mathcal{F}_n(\underline{p})$ yields the same histogram, in the sense that $\mathbb{P}(Z \in \mathrm{bin}_j) = p_j$ for each $Z \in \mathcal{F}_n(\underline{p})$, where bin_j denotes the interval that defines the jth histogram bin. All members of $\mathcal{F}_n(\underline{p})$ yield the same ERPS. Furthermore, Eq. (4) implies that all shifted random variables $X_s^n, s = 1, \ldots, n$, yield the same expected CRPS. By contrast, the expected CRPS typically differs across members that are non-trivial mixtures of the X_s^n's, i.e., members that place strictly positive weight ω_s^n on at least two components s. The following result describes how the ERPS summarizes uncertainty across all members of $\mathcal{F}_n(\underline{p})$.

Proposition 1. *Consider the CDF $\sum_{s=1}^{n} \omega_s^n F_s^n$ of a member of the family $\mathcal{F}_n(\underline{p})$ described above, and let ECRPS $\left(\sum_{s=1}^{n} \omega_s^n F_s^n\right)$ denote the expected CRPS of this CDF. Then*

$$\min_{\omega_1^n, \omega_2^n, \ldots, \omega_n^n \in \Delta_n} \mathrm{ECRPS}\left(\sum_{s=1}^{n} \omega_s^n F_s^n\right) = \sum_{j=1}^{K-1} P_j(1 - P_j),$$

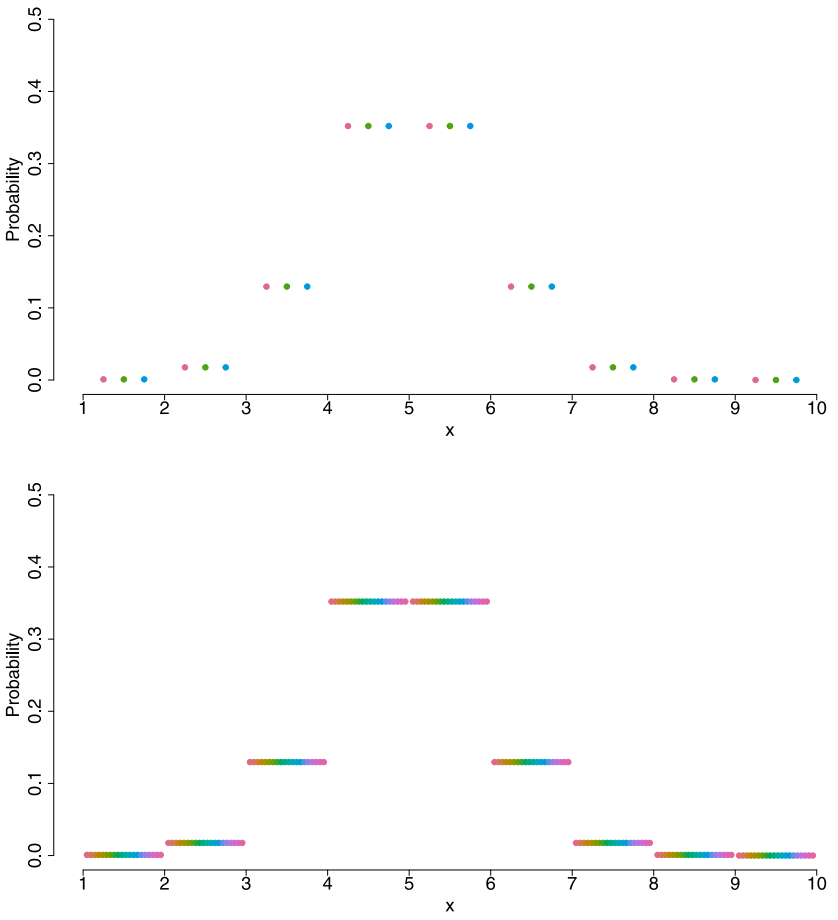

Fig. 4. Illustration of the shifted random variables X_s^n, for $n = 3$ (top panel) and $n = 20$ (bottom panel). Each color corresponds to one value $s \in \{1, 2, \ldots, n\}$.

where Δ_n is the set of all nonnegative weights that sum to one. Furthermore, the minimum is attained by setting $(\omega_1^n, \ldots, \omega_n^n)$ equal to a unit vector of length n (with $\omega_s^n = 1$ for exactly one value $s = s^*$, and $\omega_s^n = 0$ for all other values of s.)

Proof. The result follows directly from the fact that ECRPS is a strictly concave function of its (CDF-valued) argument, which in turn results from the CRPS being a strictly proper scoring rule (see Gneiting & Raftery, 2007, end of Section 2.1). Note that Gneiting and Raftery define strictly proper scoring rules in positive orientation. Since we use them in negative orientation (such that smaller scores are better), their 'strictly convex' must be replaced by 'strictly concave' in our setting. Strict concavity of the ECRPS function means that

$$\begin{aligned}\mathrm{ECRPS}\Big(\sum_{s=1}^{n} \omega_s^n F_s^n\Big) &\geq \sum_{s=1}^{n} \omega_s^n \underbrace{\mathrm{ECRPS}(F_s^n)}_{=\sum_{j=1}^{K-1} P_j(1-P_j)} \\ &= \sum_{j=1}^{K-1} P_j(1-P_j),\end{aligned}$$

with equality if and only if $\omega_1^n, \ldots, \omega_n^n$ is a unit vector. □

In words, the proposition states that $\mathrm{ERPS}(\underline{p}) = \sum_{j=1}^{K-1} P_j(1-P_j)$ is the minimal CRPS entropy in the family of distributions $\mathcal{F}_n(\underline{p})$, all of which are consistent with the same survey histogram $\underline{p}$. The analysis also implies that in the given setup of a histogram with bins $[0, 1], (1, 2], \ldots, (9, 10]$, each shifted random variable X_s^n is compatible with $\underline{p}$, and its ECRPS coincides with the ERPS of the histogram. Hence, we can identify several numerical random variables replicating the ERPS' uncertainty assessment.

The specific bin width of one considered here is not essential: Choosing equal-sized bins of another length c would render the ECRPS of each X_s^n equal to c times the ERPS of $\underline{p}$ (see Eq. (4)). By contrast, using bins of different lengths would make the link between ECRPS and ERPS less clear, which seems natural: Given that the ERPS is based on an ordinal interpretation of the bins, it cannot usefully reflect information on differences in bin length.

The family $\mathcal{F}_n(\underline{p})$ of distributions we consider ensures that each member is compatible with $\underline{p}$. Furthermore, it is easily possible to characterize the individual members' uncertainty (here, their ECRPS). This makes this family a suitable choice for studying the link between the ERPS and the uncertainty of distributions for numerical outcomes. At the same time, the result that the ERPS is

Table 4
Summary statistics for quantified standard deviations for the EMW and mass-at-midpoint (MAM) methods. Based on SCE histograms for inflation (one year ahead), between June 2013 and April 2020.

Std. dev. of all histograms ($n = 103\,734$)						
Variant	5% quant.	25% quant.	Median	Mean	75% quant.	95% quant.
EMW	0.41	0.82	1.52	2.61	3.14	8.66
MAM-Narrow	0.00	1.02	2.04	2.71	3.63	8.27
MAM-Medium	0.00	1.02	2.06	2.90	3.98	8.97
MAM-Wide	0.00	1.02	2.10	3.81	5.69	12.42
Std. dev. of histograms using at least one outer bin ($n = 38\,737$)						
Variant	5% quant.	25% quant.	Median	Mean	75% quant.	95% quant.
EMW	1.37	2.48	3.89	4.99	6.82	11.77
MAM-Narrow	2.30	3.25	4.15	4.96	6.45	9.42
MAM-Medium	2.55	3.60	4.61	5.45	6.93	10.34
MAM-Wide	3.48	5.18	6.95	7.89	9.69	15.01

the minimal expected CRPS does not hold across broader classes of distributions $\mathcal{G}(\underline{p})$ that are compatible with $\underline{p}$. We demonstrate this via the following example:

Example. Consider the bins $[0, 1], (1, 2], \ldots, (9, 10]$, and let $\underline{p} = (0.1, \ 0.1, \ \ldots, \ 0.1)'$. The discrete distribution G which places probability 0.1 on the ten points $g_1, g_2, \ldots, g_{10} = 0.99, 1.99, 2.99, 3.99, 4.99, 5.01, 6.01, 7.01, 8.01, 9.01$ is compatible with $\underline{p}$, and attains an expected CRPS of 1.405 (see Eq. (4)). This is strictly less than the ERPS of $\underline{p}$, given by 1.65.

A.4. Choice of support limits for mass-at-midpoint (MAM) method

Here we analyze the sensitivity of the MAM method to the choice of limit for the two outer bins, focusing on inflation expectations for brevity. Recall that the SCE's leftmost bin has an upper limit of -12, whereas the SCE's rightmost bin has a lower limit of $+12$. We consider the following three variants for closing the SCE's outer bins:

- **Narrow:** Leftmost bin equals $[-16, -12]$, rightmost bin equals $(12, 16]$. This implies a bin width of four, shared by the widest interior bins.
- **Medium:** Leftmost bin equals $[-20, -12]$, rightmost bin equals $(12, 20]$, i.e. doubling the bin width of the narrow variant.
- **Wide:** Leftmost bin equals $[-38, -12]$, rightmost bin equals $(12, 38]$. This choice corresponds to the maximal limit of 38 (or minimal limit of -38) that we impose in our implementation of the EMW method, following a proposal by Armantier et al. (2017). As noted in Footnote 7, a wide choice of bin limits could also be motivated by the empirical occurrence of extreme *point* expectations of inflation.

The narrow and wide choices of bin limits are at the lower and upper end of what we consider plausible. However, this assessment is necessarily subjective as no rigorous justification exists for one particular choice.

We implement the MAM method for these three choices of outer bins and consider the standard deviation obtained via the EMW method as a benchmark. Table 4 summarizes the empirical results, focusing on one year ahead of inflation expectations (SCE, variable code Q9). For histograms that use only interior bins, the choice of outer bin limit is irrelevant by construction. Hence the differences in standard deviations are driven entirely by histograms that use at least one outer bin (about 37% of all histograms). The bottom panel of Table 4 thus presents summary statistics for this subsample.

For example, in the bottom panel of Table 4, the average standard deviation is 4.96 for the narrow choice, and 5.45 for the medium choice, corresponding to a relative increase of about 10%. The wide choice of outer bins generates substantially larger standard deviations, with an average of 7.89 (about 46% larger than the medium choice). Compared to EMW, all MAM variants yield higher mean and median values. However, the right tail of standard deviations tends to be higher for EMW than for the narrow and medium variants of MAM.

The practical relevance (or irrelevance) of these differences in standard deviations seems specific to the application considered. However, given the lack of a rigorous justification for the choice of bin limit for the MAM method, checking the robustness of empirical results to this parameter will often be necessary and burdensome.

A.5. Comparing subjective and objective uncertainty

Here we provide a statistical justification for using the (expected and realized) RPS to compare subjective and objective measures of uncertainty and assess whether consumers' uncertainty assessment is realistic. The latter aspect is economically relevant in that misperceptions of uncertainty lead to economically suboptimal decisions in a wide range of situations (see, e.g. Ben-David et al., 2013, and the references therein). Comparisons of expected and realized loss that are conceptually similar to the ones sketched here have been proposed by Clements (2014), Galvao and Mitchell (2019), and Wei, Balabdaoui, and Held (2017).

We consider a so-called prediction space setup (Gneiting & Ranjan, 2013) that models the joint distribution of expectations and realizations. We treat the K histogram probabilities $\mathbf{p}$ 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 of interest, Ω, consists of forecast-observation pairs $(\mathbf{p}, \mathbf{k}^*)$. 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).[8] As in Ehm, Gneiting, Jordan, and Krüger (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 the expected and realized RPS 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 2. *Under Assumption 1, it holds that* $\mathbb{E}(RPS(\underline{\mathbf{p}}, k^*)) = \mathbb{E}(ERPS(\underline{\mathbf{p}}))$.

Proof. We have that

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

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. As noted by Gneiting and Resin (2022), the assumption is equivalent to $\underline{\mathbf{p}}$ being auto-calibrated (Tsyplakov, 2013), a notion of unbiasedness studied in the forecast evaluation literature. Under Assumption 1, Proposition 2 states that the RPS and ERPS of $\underline{\mathbf{p}}$ coincide in expectation. As a simple example (loosely following Gneiting, Balabdaoui, & Raftery, 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 $\mathbf{p}_1^{\mathrm{A}} = \Phi(-X), \mathbf{p}_2^{\mathrm{A}} = 1 - \Phi(-X) = \Phi(X)$. For forecaster A, Assumption 1 is satisfied with $\mathcal{F} = \sigma(X)$, the sigma algebra generated by X. In line with Proposition 2, it can be shown that the expected RPS and expected ERPS of forecaster A equal 1/6. In the notation of Proposition 2, it holds that $\mathbb{E}(\mathrm{RPS}(\underline{\mathbf{p}}^{\mathrm{A}}, k^*)) = \mathbb{E}(\mathrm{ERPS}(\underline{\mathbf{p}}^{\mathrm{A}})) = 1/6$. For a second forecaster B with $\mathbf{p}_1^{\mathrm{B}} = \mathbf{p}_2^{\mathrm{B}} = 0.5$, Assumption 1 is satisfied with $\mathcal{F} = \emptyset$, the empty information set. The expected ERPS and expected RPS of forecaster B equal 1/4, confirming the intuition that B's forecast is less informative than A's forecast.

[8] 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.

References

Armantier, O., Koşar, G., Pomerantz, R., Skandalis, D., Smith, K., Topa, G., et al. (2021). How economic crises affect inflation beliefs: Evidence from the Covid-19 pandemic. *Journal of Economic Behaviour and Organization*, *189*, 443–469.

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

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*, 1593–1636.

Becker, C., Duersch, P., Eife, T., & Glas, A. (2021). Extending the procedure of Engelberg et al. (2009) to surveys with varying interval-widths. Working Paper, available at https://www.awi.uni-heidelberg.de/md/awi/forschung/deseminar/dp707_beckerduerscheifeglas2021_note.pdf. (Last Accessed 21 July 2022).

Ben-David, I., Fermand, E., Kuhnen, C. M., & Li, G. (2019). Expectations uncertainty and household economic behavior. Working Paper, available at http://public.kenan-flagler.unc.edu/faculty/kuhnenc/RESEARCH/bfkl_v7.pdf. (Last Accessed 23 June 2022).

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

Binder, C. C. (2017). Measuring uncertainty based on rounding: New method and application to inflation expectations. *Journal of Monetary Economics*, *90*, 1–12.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, *77*, 623–685.

Boero, G., Smith, J., & Wallis, K. F. (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.

Carriero, A., Clark, T. E., & Marcellino, M. (2018). Measuring uncertainty and its impact on the economy. *The Review of Economics and Statistics*, *100*, 799–815.

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

Clements, M. P., Rich, R. W., & Tracy, J. S. (2023). Surveys of professionals. In R. Bachmann, G. Topa, & W. van der Klaauw (Eds.), *Handbook of Economic Expectations* (pp. 71–106). Academic Press, http://dx.doi.org/10.1016/B978-0-12-822927-9.00009-4.

Coibion, O., Gorodnichenko, Y., & Kumar, S. (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.

D'Acunto, F., Malmendier, U., & Weber, M. (2023). What do the data tell us about inflation expectations? In R. Bachmann, G. Topa, & W. van der Klaauw (Eds.), *Handbook of Economic Expectations* (pp. 133–161). Academic Press, http://dx.doi.org/10.1016/B978-0-12-822927-9.00012-4.

Deutsche Bundesbank (2022). Survey on consumer expectations. https://www.bundesbank.de/en/bundesbank/research/survey-on-consumer-expectations/survey-on-consumer-expectations-794568. (Last Accessed 1 March 2023).

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

ECB (2019). Inflation expectations and the conduct of Monetary policy. Speech by Benoît Cœuré, Member of the Executive Board of the ECB, at an event organised by the SAFE Policy Center, Frankfurt am Main, 11 July 2019. Available at https://www.ecb.europa.eu/press/key/date/2019/html/ecb.sp190711~6dcaf97c01.en.html. (Last Accessed 23 June 2022).

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (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., Manski, C. F., & Williams, J. (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.

Federal Reserve Bank of New York (2020). Survey of consumer expectations. https://www.newyorkfed.org/microeconomics/sce. (Last Accessed 9 March 2020).

Federal Reserve Bank of Philadelphia (2022). Survey of professional forecasters. https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/. (Last Accessed 23 July 2022).

Galvao, A. B., & Mitchell, J. (2019). Measuring data uncertainty: An application using the Bank of England's 'fan charts' for historical GDP growth. Working Paper, available at https://warwick.ac.uk/fac/soc/wbs/subjects/finance/mpf/working-papers/galvao_mitchell_may_19_escoe.pdf. (Last Accessed 19 July 2022).

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

Glas, A. (2020). Five dimensions of the uncertainty–disagreement linkage. *International Journal of Forecasting*, *36*, 607–627.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*, 243–268.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Gneiting, T., & Resin, J. (2022). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. Working paper, available at https://arxiv.org/abs/2108.03210v3. (Last Accessed 2 March 2023).

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *14*, 107–114.

Gosselin, M.-A., & Khan, M. (2015). A survey of consumer expectations for Canada. *Bank of Canada Review*, *2015*(Autumn), 14–23.

Grishchenko, O., Mouabbi, S., & Renne, J.-P. (2019). Measuring inflation anchoring and uncertainty: A US and Euro area comparison. *Journal of Money, Credit and Banking*, *51*, 1053–1096.

López-Menéndez, A. J., & Pérez-Suárez, R. (2019). Acknowledging uncertainty in economic forecasting. Some insight from confidence and industrial trend surveys. *Entropy*, *21*, 413.

Mackowiak, B., & Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *American Economic Review*, *99*, 769–803.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*, 1329–1376.

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

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*, 1087–1096.

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

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

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

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

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. *Working paper*, http://dx.doi.org/10.2139/ssrn.2236605, (Last Accessed 23 June 2022).

Wei, W., Balabdaoui, F., & Held, L. (2017). Calibration tests for multivariate Gaussian forecasts. *Journal of Multivariate Analysis*, *154*, 216–233.