European Journal of Operational Research 304 (2023) 661–675

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

# Testing the effectiveness of debiasing techniques to reduce overprecision in the elicitation of subjective continuous probability distributions

Valentina Ferretti[a,b], Gilberto Montibeller[c,d,*], Detlof von Winterfeldt[d]

[a] London School of Economics, UK
[b] Politecnico di Milano, Italy
[c] Loughborough University, UK
[d] University of Southern California, USA

ARTICLE INFO

*Article history:*
Received 18 November 2020
Accepted 7 April 2022
Available online 14 April 2022

*Keywords:*
Behavioral OR
Overconfidence
Overprecision
Judgment calibration
Expert judgment
Debiasing

ABSTRACT

Formal expert elicitation is a widely used method for quantifying uncertain variables in decision and risk analysis. When estimating uncertain variables, experts and laypeople exhibit overprecision, meaning that the ranges of their estimates are too narrow. Overprecision, a form of overconfidence, is pervasive and hard to correct, thus posing a challenge to expert elicitation. Following the increasing interest toward improving judgments in Behavioral Operational Research (OR), and the limited evidence about the effectiveness of debiasing tools, the aim of our research is to test the effectiveness of commonly employed practices for debiasing overprecision. We conducted two experiments, testing a set of debiasing techniques when eliciting points of a cumulative distribution functions for general knowledge questions. The debiasing procedures included hypothetical bets, counterfactual argumentation, and automatic stretching to increase the ranges of subjects' initial estimates. We find that two debiasing strategies that require further reasoning after initial estimates (hypothetical bets and counterfactuals) were not very effective for reducing overprecision, while the use of multipliers that increase the initial range of distributions, coupled with a re-elicitation of the distribution with the new range, provided more positive results. We provide some recommendations for expert elicitation in OR practice, based on our findings, and suggest avenues for further research into debiasing overprecision.



## 1. Introduction

Formal expert elicitation is a common method used in several Operational Research (OR) methods, such as decision analysis and risk analysis, whenever variables or events are uncertain, data are sparse, and models and expertise conflict (Aloysius et al., 2006; Dias et al., 2018; Keeney & von Winterfeldt, 1991; Morgan, 2014; Ortiz et al., 1990; Werner et al., 2017).

It is well known that experts and laypeople have biases when making probability judgments (Kirshner & Shao, 2019; Montibeller & von Winterfeldt, 2015), which can lead to low-quality OR models, as these judgments are critical inputs for the decision analysis. Among the many biases, overconfidence is the most common and important one, occurring both in informal and formal elicitations of uncertainties. For binary uncertain quantities, overconfidence is the tendency to assign unreasonably high probabilities to the event that one believes to be true. For continuous uncertain quantities, overconfidence is the tendency to provide ranges of estimates that are far too narrow. This phenomenon is also known as *overprecision*, or overconfidence in interval estimation (Soll & Klayman, 2004).

Overprecision is a persistent and poorly understood form of overconfidence (Moore et al., 2015b), which can lead to serious and detrimental consequences in decision and risk analysis. If the ranges of the estimated variable are too narrow, analysts are likely to miss the true value, i.e., it falls outside of their estimated range. For instance, when planning mitigation measures against sea level rise and estimating the range of this variable, overprecision can cause two potential problems. If the true value turns out to be above the estimated range, sea level rise would be underestimated, thus leading to unpreparedness and insufficient mitigation. On the other hand, if the true value turns out to be lower than the esti-

* Corresponding author at: Management Science and Operations Group, School of Business and Economics, Loughborough University, UK.
*E-mail address:* g.montibeller@lboro.ac.uk (G. Montibeller).

https://doi.org/10.1016/j.ejor.2022.04.008
0377-2217/

mated range, sea level rise would be overestimated, thus leading to inefficient use of resources due to overpreparedness, with costly and unnecessary risk mitigation.

Research has repeatedly found that the true value of an uncertain quantity falls outside of people's estimated ranges (i.e., outside the lower and upper bounds of an uncertain quantity) more often than should be expected (Alpert & Raiffa, 1982; Barberis & Thaler, 2003; Bolger & Harvey, 1995; Clemen, 2001; Fischhoff et al., 1978; Lichtenstein et al., 1982; Soll & Klayman, 2004; Wallsten et al., 1983). Expert predictions of future events tend to be closer to the true value, but they are also narrower than the predictions provided by laypeople (e.g., Graf-Vlachy, 2017; Haran et al., 2010; McKenzie et al., 2008; Onkal et al., 2003).

Within this context, most research has focused on overconfidence in estimating the probability of binary outcomes (e.g., Nguyen, 2018) and on overprecision in eliciting confidence intervals over a range (e.g., Schall et al., 2016). Our research focuses on a similar task as the latter one, on overprecision in eliciting continuous probability distributions, as these distributions are often required in decision and risk analysis (Keeney & von Winterfeldt, 1991; McNamee & Celona, 2008; Spetzler et al., 1975; Wallsten et al., 2016) and in policy analysis practices (Morgan, 2014).

While there is now extensive knowledge about the bias and its causes and consequences, we know less about how to effectively reduce overprecision in the elicitation of continuous probability distributions. Four main directions for debiasing overprecision have been suggested so far: (i) varying the elicitation presentation format (e.g., Abbas et al., 2008; Seaver et al., 1978); (ii) encouraging the consideration of more information (e.g., Haran et al., 2010); (iii) warning against the bias (e.g., Schall et al., 2016); and (iv) providing training and/or immediate feedback (e.g., Mannes & Moore, 2013; von Winterfeldt & Edwards, 1986). However, very limited research has been conducted to test and compare the effectiveness of different debiasing strategies to reduce overprecision, with the exception of Abbas et al. (2008) and Seaver et al. (1978).

Following the increasing interest toward improving judgments in Behavioral Operational Research (Franco et al., 2021), the objective of the research presented in this paper is to experimentally test and compare widely employed practices in decision and risk analysis for debiasing overprecision in the elicitation of continuous probability distributions.

The most commonly used debiasing procedures consist of using hypothetical bets, counterfactuals, and some form of stretching of the ranges of the elicitation variables after an expert defines what they believe are the absolute minimum and the absolute maximum bounds for a variable (Ferretti et al., 2016; Montibeller & von Winterfeldt, 2015; Morgan, 2014). To test the effectiveness of these commonly used methods for debiasing, we conducted two experiments. We started by testing and comparing the hypothetical bet and counterfactual methods but found low effectiveness in reducing overprecision in terms of the indicators employed in the comparison. We subsequently tested the automatic stretching procedure and found this to be a more promising debiasing approach.

This research makes three contributions. This is the first study, to our knowledge, that has attempted to experimentally test and systematically compare the effectiveness of commonly employed methods for debiasing overprecision in the elicitation of points of continuous probability distributions. The second contribution consists of a comparative study of debiasing strategies, which moves beyond traditional "think harder" strategies and includes adjustments of ranges using multipliers. Our results indicate that both the use of counterfactuals and hypothetical bets were not effective in reducing this bias. Linked to these results, our third contribution is to provide practical guidelines on how the use of multipliers may be implemented in decision and risk analysis practice, with the use of automatic stretching, the selection of adequate automatic multiplier, and the adoption of the fixed value elicitation protocol (Abbas et al., 2008).

The remainder of the paper is organized as follows. Section 2 reviews the literature on debiasing overprecision. Section 3 presents the research design and the evolution of the study across the two experiments. Section 4 provides a discussion of the findings of the experiments and is followed by conclusions and directions for further research.

## 2. Debiasing overprecision: literature review

Since Alpert & Raiffa (1982) discovered the overprecision bias, several researchers have found that this bias occurs strongly and consistently across many contexts and people, including different professions, ages, genders, levels of expertise, cultures, and elicitation formats (e.g., Jonsson & Allwood, 2003; Moore et al., 2015b, 2015a).

After more than 50 years of research on identifying and describing biases in judgment and decision-making, the more recent literature has highlighted the need to develop strategies and techniques to reduce these biases (e.g., Jain et al., 2013; Lahtinen et al., 2020; Milkman et al., 2009; Montibeller & von Winterfeldt, 2015; Morewedge et al., 2015; Schall et al., 2016). This is evidenced by the increasing focus on improving judgments in behavioral economics (e.g., Thaler & Sunstein, 2008), behavioral operations management (e.g., Ren & Croson, 2013), behavioral policy (e.g., Galizzi, 2014), behavioral decision research (e.g., Milkman et al., 2009), and behavioral operational research (e.g., Franco & Hamalainen, 2016; Franco et al., 2021).

As mentioned in the introduction, efforts for debiasing overprecision in individual estimates have taken several paths and can be grouped into the following four categories: (i) varying the elicitation/presentation format; (ii) encouraging the consideration of more information; (iii) using warnings against the bias; and (iv) providing training and/or feedback. We revise briefly the main developments within each category next.

Within the first category, several elicitation protocols have been suggested to reduce overprecision in expert elicitation. Several studies (e.g., Juslin et al., 1999; Seaver et al., 1978; Teigen & Jorgensen, 2005) found that judging probabilities with fixed intervals produces less overprecision than estimating intervals for fixed probabilities. Winman et al. (2004) proposed a method for adaptive interval assessment of eliciting judgments, starting with a confidence level for a given interval and interactively revising this during the elicitation task. They found that the results of this method displayed less overprecision than intervals that are elicited directly, given a confidence level. Soll & Klayman (2004) tested interval estimates and used three elicitation formats, showing that asking for fractiles separately led to better calibration of estimates than asking for confidence intervals. Bedford & Cooke (2001) suggest a formula for adjusting the original upper and lower bound from the expert's stated confidence (see also Burgman (2016) for a more detailed discussion on eliciting intervals from experts and on quality measures for estimates, such as calibration).

The elicitation of continuous distribution has been studied by Seaver et al. (1978), who tested five procedures for eliciting subjective probability distributions over continuous variables. They showed that asking for probabilities for given intervals of the uncertain quantity performed better in reducing overprecision than asking for fractiles. In addition, asking for odds performed slightly better than asking for probabilities in their experiment. Within the practice of probability assessment, Abbas et al. (2008) compared different methods using direct quantile assessments – fixed probability (FP), fixed variable (FV), and a mixture of the two – and found the fixed variable method to be superior when measured

in terms of monotonicity and accuracy. However, Budescu & Du (2007) have shown that this better performance of FV for probability judgements is not consistently true for all levels of confidence.

When examining a forecast of time to complete a task, Jain et al. (2013) found that decomposing the interval estimation task into its time elements can substantially widen the intervals and thereby potentially improve calibration. Welsh & Begg (2018) showed that the more-or-less elicitation (MOLE) approach reduced overconfidence ranges when participants were asked to successively choose between computer-generated values rather than making their interval estimates directly. Once a participant chose the value which they believed was closer to the true value of the parameter under analysis, the MOLE approach then proposed two new values from the revised range and repeated the process for several iterations.

Camilleri & Newell (2019) explored how overprecision is affected by information processing. They showed that asking participants to generate 90% confidence intervals and presenting uncertainty information sequentially leads to underprecision, while a summary format of presenting information leads to overprecision.

The second category of debiasing tools encompasses strategies that encourage the consideration of more information, contrary evidence, and possible alternatives ("think harder" strategies). Within this category, Koriat et al. (1980) asked participants to make estimates in a forced choice format (two-alternative questions from which to choose an answer). Subsequently, they asked subjects to list arguments contradicting their selection. When asked to provide counterarguments, participants showed lower overconfidence, measured by comparing the assessed probabilities with the observed relative frequencies of being correct ("hit rates"). For interval estimates, Soll & Klayman (2004) asked their participants to specify a lower and upper bound of their initial fractiles. However, the efficacy of counterfactuals with confidence intervals and continuous distributions remains untested (Moore et al., 2015b).

More recently, Haran et al. (2010) proposed the subjective probability interval estimates (SPIES) method, which forces subjects to consider a pre-specified range of possible outcomes of a target variable. This range is then divided into intervals, and the participants are asked to estimate the probability that the true value falls within each interval. They found that SPIES led to significantly lower surprise rates compared to other elicitation methods (90% confidence intervals and fractiles). In addition, Walters et al. (2017) introduced another debiasing technique, "consider the unknowns" and tested it with two-alternative forced-choice questions. In this technique, participants first generate a list of unknowns before stating their uncertainties. They showed that the protocol reduced overprecision more than a simple "consider the alternative" technique (Koriat et al., 1980).

Herzog & Hertwig (2009) proposed a special version of the "consider the opposite" strategy for point estimates and called it dialectical bootstrapping. This strategy has shown to create better quantitative judgments by averaging the respondent's first estimate with a second, dialectical estimate. This approach leads to a gain in accuracy if the second estimate by the same judge is based on non-redundant knowledge and assumptions, thus leading to a different error compared to the first estimate.

The third category of overprecision debiasing tools, providing warnings against the occurrence of the bias, has usually been considered one of the least effective measures to debias interval overconfidence (e.g., Plous, 1995). However, more recently, Schall et al. (2016) showed that a significantly higher reduction of overprecision in interval estimates can be achieved by combining a dynamic process of warning content and stimulus change to make the warnings more salient.

Finally, research in the fourth category has shown that immediate feedback can improve the quality of subject's probability assessments (e.g., Mannes & Moore, 2013, for point estimates; Block & Harper, 1991, for both point and interval estimates). For instance, Murphy & Winkler (1977) demonstrated that weather forecasters, who are given daily feedback on their forecasts, have excellent calibration results for point probability estimates. There is also behavioral evidence that training (see, for example, Hora, 2007, for probability estimates) may improve the calibration of subjects (e.g., Chang et al., 2016, for probability estimates with binary forecast questions; Keren, 1987, for point probability estimates; Alpert & Raiffa, 1982, for probability intervals), although results lack generalizability (e.g., Lichtenstein & Fischhoff, 1980, for both probability and interval estimates). Table 1 summarizes the type of elicitation tasks tested with the debiasing approaches discussed in this section.

In conclusion, despite a plethora of elicitation protocols suggested in the literature, there is limited systematic comparisons of the effectiveness of different debiasing strategies against overprecision for cumulative distribution functions, except for Abbas et al. (2008) and Seaver et al. (1978). The comparison that we thus propose in this paper is an attempt to address this gap.

## 3. Overview of two experiments: methods and research questions

This section presents an overview of two behavioral experiments designed to test some of the commonly employed practices in decision and risk analysis to debias overprecision. We focus here on elicitation protocols, as other approaches, reviewed in the previous section, may be less relevant in practice. For instance, providing immediate feedback about the accuracy of estimates is not feasible in most real-world decision and risk analysis applications, which are often characterized by long-term horizons and one-off decisions (Larrick, 2004). In the same spirit, we do not consider methods or protocols that require pre-specified ranges, e.g., SPIES by Haran et al. (2010), because it is often not possible to identify ranges a priori for variables with high levels of uncertainty, which are typical in situations that require expert judgment elicitation (e.g., Dias et al., 2018; Hora, 2007; Keeney & von Winterfeldt, 1991). In addition, providing a pre-specified range might anchor the responses on these extreme values (Morgan, 2014) or be interpreted by participants as information (i.e., defining the "correct" range of uncertainty). Finally, we note that probability training and cautioning experts against biases are usually part of a formal elicitation process (e.g., Keeney & von Winterfeldt, 1991) and, therefore, we assume these preliminary activities as common practice.

### 3.1. Methods

In the first experiment, we tested two widely employed debiasing tools for stretching the bounds of the distribution (the $x_0$ and the $x_{100}$ fractiles): the use of counterfactual (CF) questions (Morgan, 2014) and hypothetical betting (HB) questions (e.g., Fox & Tversky, 1998; Kennedy, 1986; Montibeller & von Winterfeldt, 2015; Rapoport, 1964; Winkler, 1971). Given the low effectiveness of these two debiasing tools, which we observed in the first experiment, we introduced in the second experiment a new treatment which automatically stretches (AS) the initial bounds provided by participants ($x_0$ and $x_{100}$) and subsequently re-elicits the 10th, 50th, and 90th fractiles using these new ranges. The stretching is "automatic", since it provides the participants with new extremes by multiplying their highest estimate by 2 and dividing their lowest estimate by 2. (We acknowledge that other multipliers could have been considered; however, we also note that this was the first controlled test of this debiasing approach, as far as we know, and starting from a factor of two seemed a reasonable

**Table 1**
Type of Tasks Tested with the Main Debiasing Approaches.

| Categories of Debiasing | Type of task | | |
|---|---|---|---|
| | Point estimates | Interval estimates | Probability estimates |
| Varying the elicitation/presentation format | | Bedford and Cooke (2001), Camilleri & Newell (2019), Jain et al. (2013), Juslin et al. (1999), Seaver et al. (1978), Soll & Klayman (2004), Teigen & Jorgensen (2005), Welsh & Begg (2018), Winman et al. (2004) | Budescu & Du (2007), Seaver et al. (1978), Abbas et al. (2008) |
| Encouraging the consideration of more information | Herzog & Hertwig (2009), Koriat et al. (1980), Walters et al. (2017) | Haran et al. (2010), Soll & Klayman (2004) | |
| Using warnings against the bias | | Schall et al. (2016), Plous (1995) | |
| Providing training and/or feedback | Mannes & Moore (2013), Block & Harper (1991) | Alpert & Raiffa (1982), Block & Harper (1991), Lichtenstein & Fischhoff (1980) | Hora (2007), Keren (1987), Lichtenstein & Fischhoff (1980), Murphy & Winkler, 1977 |

assumption. We conduct a sensitivity analysis on this parameter in Section 4.)

While the automatic stretching is a computational change of the range of the estimates presented to the participants, a perfectly calibrated participants should not change their fractiles, just because the presented range has changed. By re-eliciting the three internal fractiles (10th, 50th, and 90th) with the stretched extremes, we can determine how the change in the extremes affects the assessed cumulative probability distribution and whether it reduces overprecision. Some may argue that automatic stretching is not really a debiasing method, but rather a computational intervention without a behavioral implication. However, in real world elicitations, ranges of uncertain variables are not always given, but absolute minima and maxima often exist. Additionally, there is, in principle, no difference in creating ranges artificially or asking the participants for it. In the second experiment we re-elicited the cumulative distribution after changing the ranges, thus adding the behavioral component of an effect of the ranges.

To elicit fractiles of the distributions ($x_{10}$, $x_{50}$ and $x_{90}$) within these three methods we tested and compared the fixed value (FV) and fixed probability (FP) elicitation protocols, which are the most common methods for eliciting probability distributions of continuous variables in decision and risk analysis. In the FP approach, the decision analyst selects a set of cumulative probabilities ($p_i$, $i=1$, …, $n$) and subjects are asked to report variable values ($x_i$, $i=1$, …, $n$) such that the $Pr\ (X \leq x_i) = p_i$, where $n$ is the number of data points. In the FV approach, instead, the decision analyst selects a set of variable values ($x_i$, $i=1$, …, $n$) and asks the subjects to provide their cumulative probabilities ($p_i$, $i=1$, …, $n$) such that $Pr\ (X \leq x_i) = p_i$ (Abbas et al., 2008). Many practitioners limit the number of elicitation points to three or five to reduce the elicitation burden. For instance, in Cooke's classical model, experts are asked to provide their 5%, 50%, and 95% quantiles. Other protocols, such as the one suggested by Burgman (2016), ask for the lowest and highest plausible estimates before eliciting other quantiles (for details see Dias et al., 2018; Morgan, 2014).

In summary, the independent variables in this series of experiments are the debiasing tools used to stretch the bounds (CF, HB, and AS) as well as the elicitation protocols (FV and FP) for a three-by-two design. We used general knowledge questions, as typically employed in many previous behavioral experimental studies (e.g., Koriat et al., 1980; Langnickel & Zeisberger, 2016; Moore et al., 2015a; Ren & Croson, 2013; Seaver et al., 1978). We have selected difficult questions, as it has been shown that those questions tend to produce overprecision, whereas easy questions may produce underprecision (e.g., Moore & Healy, 2008; Moore et al., 2015b). The advantage of this type of general knowledge question design is that, contrary to prediction tasks, the true answer is available to provide an immediate reward for accuracy.

To minimize the influence of "anchoring and adjustment" when eliciting probability distributions, the decision analyst usually does not begin with central tendency questions but, instead, probes the extreme lower and upper values (Connolly & Dean, 1997; Morgan, 2014). We thus set up the elicitation tasks by asking first the extremes, then the median, followed by the 10% and 90% fractiles. To replicate as much as possible real-world elicitation processes, we allowed each subject to see their initial distributions as a graph and revise them as much as they wished before proceeding to the next elicitation step, as opposed to more artificial settings in which revisions are not allowed.

As experts and novices exhibit similar levels of overprecision, the participants in our experiment were master's degree students that had taken a course in statistics during their studies. All participants in this study watched a short training video at the beginning of the experiment to make sure that they clearly understood the elicitation task and the way their performance was measured (i.e., scoring rule).

We incentivized accuracy in two ways. We paid a fixed fee for participation in the experiment (£10 per hour) and rewarded the best performer with a lump sum (£100). For the latter, we used the Matheson & Winkler (1976) scoring rule, a widely accepted rule for measuring the accuracy of continuous distributions (e.g., Seaver et al., 1978):

$$S = \int_0^{x_t} [F(x)]^2 \, dx + \int_{x_t}^{\infty} [1 - F(x)]^2 \, dx, \tag{1}$$

where $x_t$ is the true value (standardized as 100) and $F(x)$ is the elicited cumulative probability distribution[1] of variable $x$. The lower $S$ is, the more accurate the judgments. The duration of the experiment was 60 minutes, encompassing 10 general knowledge questions presented in a random order, and there was no incentive for a faster/slower response time. The average S over the 10 questions was used to measure the accuracy of each subject and reward the best performer.

### 3.2. Research questions, design and dependent variables

The research question in the two experiments is whether the experimental debiasing methods can reduce overprecision. In experiment 1 we used two debiasing methods (CF and HB) and two elicitation methods (FV and FP) in a two-by-two design. The design

[1] We assumed a piecewise linear function between each pair of elicited data points to obtain a continuous CDF. This linearization takes into account the lack of information between each of those two points and assumes a uniform distribution based on the principle of maximum entropy (Jaynes, 1957).

**Table 2**
General knowledge questions for Experiment 1.

| Ref. Number | Variables | True Value |
|---|---|---|
| 1 | Height of Turin Tower "La mole Antonelliana" [m] | 167.5 |
| 2 | Distance between Turin and Milan central train stations [km] | 147 |
| 3 | "Piazza Vittorio Veneto" inclination from its highest to its lowest point [cm] | 719 |
| 4 | Length of Turin's Arcades [km] | 18 |
| 5 | Area of Italy [$km^2$] | 301,340 |
| 6 | Height of the Turin skyscraper by Renzo Piano [m] | 167.5 |
| 7 | Mean global sea level rise in 2100 projected by the 5th IPCC report for the high emission scenario [cm] | 75 |
| 8 | Mean global average increase in the surface temperature in 2100 projected by the 5th IPCC report for the high emission scenario [°C] | 4 |
| 9 | Number of days in which Turin has gone beyond the legislative limit of air pollution in 2013 | 106 |
| 10 | % of reused waste from the construction phase (from January 2010 to August 2014) of the Renzo Piano Skyscraper in Turin | 93 |

was similar in experiment 2, but we replaced the hypothetical bet method (HB) with the automatic stretching method (AS).

In both experiments we used the following dependent variables (we denote estimated variable values from the initial elicitation as $x_i$ and from the revised elicitation as $x'_i$):

DV1: number of judgments revised after the subject has been exposed to different debiasing tools;
DV2: proportion of surprises among the responses, measured by 10% surprises (if the true value is outside the 90% range of the distribution) and by 20% surprises (if the true value is outside the 80% range of the distribution);
DV3: width of the variable's overall range ($x_0$ to $x_{100}$);
DV4: width of the 80% inner range ($x_{10}$ to $x_{90}$) under the fixed value or fixed probability condition.

In experiment 2 we used two additional dependent variables:

DV5: marginal improvement of the 80% inner range overprecision (measured as initial $x_{90} - x_{10}$ and revised $x'_{90} - x'_{10}$);
DV6: the proportion of inner revised estimates that are moving toward the right direction after auto-stretching (lower for the 10% fractile and higher for the 90% fractile in relation to their original estimates).

### 3.3. Experiment 1

The aim of this first experiment was to test and compare the effectiveness of two debiasing protocols (CF and HB) and two elicitation methods (FV and FP), introduced above, on the bounds of the participants' subjective probability distributions.

#### 3.3.1. Methods - experiment 1

One hundred and ten subjects participated in this experiment in a large computer laboratory of the Politecnico of Torino (Italy). All participants had been exposed to probability encoding in a class lecture. Participants were asked to provide estimates for the 10 general knowledge variables (listed in Table 2).

Each participant had been randomly allocated to one of the four experimental conditions: CF-FV, CF-FP, HB-FV, HB-FP. The elicitation task for each variable consisted of the following two phases: (i) elicitation of the initial estimates for $x_0$, $x_{100}$, $x_{50}$, $x_{10}$, $x_{90}$ (in this order); and (ii) revision of the estimates for $x'_0$, $x'_{100}$, $x'_{50}$ (in this order).

Participants were working on a pre-programmed interactive Excel file which allowed them not only to see, while responding, the cumulative probability distributions in linear piecewise form plotted on the side of the screen, but also to revise their estimations before proceeding to each subsequent phase. The elicitation protocol followed two phases. These two phases were repeated for all 10 variables, which were randomly presented to each subject.

Phase 1 in the FP-CF and FP-HB conditions required the participants to provide the following values for the variable under analysis: (a) the lowest number such that they are absolutely sure that the true answer would not be below it ($x_0$); (b) the highest number such that they are absolutely sure that the true answer would not be above it ($x_{100}$); (c) the best guess so that the chances of the true answer falling below or above is 50/50 ($x_{50}$); (d) a low end such that there is a 10% chance that the true answer is between this low end point and their lowest value ($x_{10}$); and a high end such that there is a 10% chance that the true value answer is between this high end point and their highest value ($x_{90}$). This elicitation protocol is illustrated on the left-hand side of Fig. 1.

In the FV-CF and FV-HB conditions, participants were first asked to provide the following values: (a) the lowest number such that they are absolutely sure that the true answer would not be below it ($x_0$) and (b) the highest number such that they are absolutely sure that the true answer would not be above it ($x_{100}$). Subsequently, we provided (by automatic calculation in the Excel spreadsheet) 50% of the interval value ($x_{50}$), then 10% ($x_{10}$) and 90% ($x_{90}$) of the interval value, and subjects were asked to provide the cumulative probabilities for each of these three values. The probability estimates were then elicited from each of these values (but subjects were not informed that it was 10%, 50%, and 90% of the range, to avoid anchoring their probability estimate on those proportions). Therefore, in the FP protocol participants provided the median, while in the FV protocol the median was interpolated from the participants' probability judgments, assuming a piece-wise linear cumulative distribution. This elicitation protocol is illustrated on the right-hand side of Fig. 1, with the calculated median highlighted in the same graph.

For example, consider the general knowledge question number 7 employed in the experiment, *the mean sea level rise in 2100* (see Table 2) and the CDF elicited on the lefthand side of Fig. 2. In this case, the subject identified $p(X \leq 40$ centimeters$) = 0$, $p(X \leq 70$ centimeters$) = 100\%$, followed by $p(X \leq 60$ centimeters$) = 50\%$, $p\ (X \leq 44$ centimeters$) = 10\%$, and $p(X \leq 68$ centimeters$) = 90\%$. The thresholds in the same graph indicate the limits for two measures of surprise, 10% surprises (if the true value were below 42 centimeters or above 69 centimeters) and 20% surprises (if the true value were below 44 centimeters or above 68 centimeters). The true value is $x_t = 75$ centimeters, thus the initial distribution is both a 10% surprise and a 20% surprise.

During Phase 2 of the elicitation protocol, the counterfactuals and hypothetical betting debiasing techniques were used to reduce overprecision on the initial estimates. In the FP-CF and FV-CF conditions, participants were asked if they could think of plausible explanations under which the true answer was (i) lower than their initial estimate for $x_0$ and (ii) higher than their initial estimate for $x_{100}$. If they stated they could think of an explanation, they were requested to revise their initial estimates downward for the former estimate (i.e., $x'_0 < x_0$), and upward for the latter estimate (i.e., $x'_{100} > x_{100}$). Otherwise, no revision was required.

In the FP-HB and FV-HB conditions, participants were posed hypothetical bets for both the lower bound and upper bound of the range. For the former, they would need to pay £100 if the true

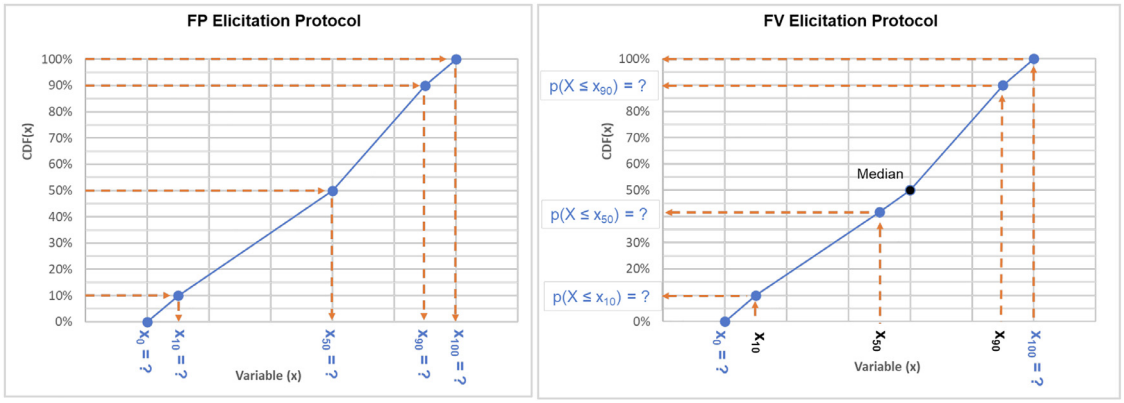



**Fig. 1.** Fixed-Probability (FP) and Fixed-Value (FV) Elicitation Protocols Employed in the Experiments.

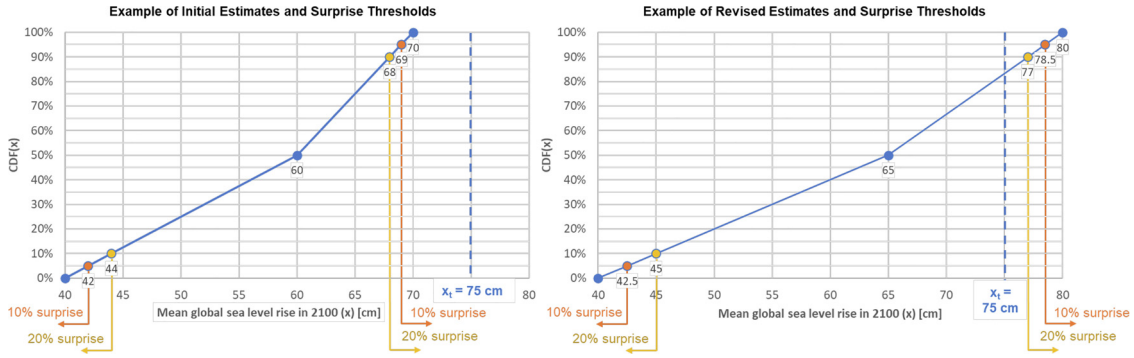



**Fig. 2.** Example of Initial and Revised Estimates and Surprise Thresholds.

value was below their initial estimate for the lowest value or receive £1 otherwise. For the latter, they would have to pay £100 if the true value was above their initial estimate for the highest value or receive £1 otherwise. If they rejected the bet, they were asked to revise their initial lower bound estimate downward (i.e., $x'_0 < x_0$) or the upper bound estimate upward (i.e., $x'_{100} > x_{100}$), respectively.

Hypothetical betting was also used for the revision of the median, as this type of protocol is widely employed to elicit fractiles of continuous distributions (Dias et al., 2018; Morgan & Henrion, 1990). Subjects were asked if they wished that the true value would fall below or above the median. They were then asked to adjust their median estimate ($x'_{50}$) until they would be indifferent to betting on its lower or upper side.[2]

For instance, in the same example above illustrated in Fig. 2, let us assume that the subject has decided to revise the upper bound, with $x'_{100} = 80$ centimeters, either by considering a counterfactual (CF treatment) or from the hypothetical betting (HB treatment), and has revised the median to $x'_{50} = 65$ centimeters. This revised distribution is shown on the righthand side of the same figure and is now neither a 10% surprise nor a 20% surprise, as the true value (75 centimeters) is below both surprise thresholds (78.5 centimeters and 77 centimeters, respectively).

### 3.3.2. Results - experiment 1

All 110 participants completed the probability estimates for all 10 questions. We considered only estimates that produced monotonic cumulative distributions and focused on responses that were more informed about the specific question, in the sense that respondents had an approximate idea about the location of the true value. Specifically, we considered an elicited distribution to be informed if the median fell in a range between 1/5th and 5 times the true value. In fractile terms, we defined an informed response as $x_{50} \leq 0.2x_t$ or $x_{50} \geq 5x_t$ (where $x_t$ is the true value). In contrast, an uninformed distribution was one where the median fall outside of the 1/5th to 5 time range around the true value. While dividing and multiplying the true value by five creates a wide range, our main intent was to avoid considering responses that were based on pure guesswork or misunderstanding of the metric of the uncertain variable.

We had 1028 monotonic distributions (93.5% of the total number of estimates), of which 817 (79.5%) were classified as informed. Most questions had responses with a rate of informed responses over 80%, but three variables had a lower rate, as detailed in Table 3. Question number 3 in particular proved difficult for many participants. This was probably due to the fact that the response was to be given in cm, while the true answer was several meters, leading to a severe response bias. The number of monotonic and informed responses per treatment are shown in Table 4.

As mentioned previously, to allow comparisons across the variables, we set up the $x_t$ as 100 and linearly normalized the other data points between 0 (the actual zero value) and 100 ($x_t$). We explored the effectiveness of the debiasing tools and elicitation methods by examining four dependent variables considering the informed responses, as detailed next.[3]

*DV1: number of judgments revised after the subject was exposed to a debiasing tool*

[2] We tested the effectiveness of this protocol in moving the median estimate toward the true value but found inconclusive results. The proportion of revisions toward the true value were 56.5% in Exp. 1 but only 48.5% in Exp. 2.

[3] The statistical tests that we employed in the data analysis are non-parametric version of t-tests or ANOVAs. We established that the data are not normally distributed, primarily due to skewness and/or kurtosis of the distributions. Thus, we reverted to the most common and powerful non-parametric tests.

**Table 3**
Monotonic distributions and informed responses for Exp. 1 (110 responses per variable).

| Reference Variable | Monotonic Distributions | % Monotonic Distributions | Informed Responses | % Informed Responses |
|---|---|---|---|---|
| 1 | 106 | 96.4% | 101 | 95.3% |
| 2 | 97 | 88.2% | 94 | 96.9% |
| 3 | 102 | 92.7% | 21 | 20.6% |
| 4 | 102 | 92.7% | 90 | 88.2% |
| 5 | 102 | 92.7% | 70 | 68.6% |
| 6 | 105 | 95.5% | 100 | 95.2% |
| 7 | 103 | 93.6% | 64 | 62.1% |
| 8 | 105 | 95.5% | 89 | 84.8% |
| 9 | 105 | 95.5% | 96 | 91.4% |
| 10 | 101 | 91.8% | 92 | 91.1% |
| Totals | 1028 | 93.5% | 817 | 79.5% |

**Table 4**
Monotonic and informed responses per treatment for Exp. 1.

| | Fixed-Value (FV) Elicitation | | | Fixed Probability (FP) Elicitation | | |
|---|---|---|---|---|---|---|
| | Monotonic | Informed | % Informed | Monotonic | Informed | % Informed |
| Counterfactual (CF) | 249 | 194 | 77.9% | 275 | 221 | 80.4% |
| Hypothetical Betting (HB) | 237 | 182 | 76.8% | 267 | 220 | 82.4% |

**Table 5**
Counts and percentage of revisions of both bounds of the probability distribution for counterfactuals vs. hypothetical betting for Exp. 1.

| | Total estimates | Revised $x_0$ | Revised $x_{100}$ |
|---|---|---|---|
| Counterfactuals (CF) | 415 | 58 (14.0%) | 80 (19.3%) |
| Hypothetical bets (HB) | 402 | 37 (9.2%) | 72 (17.9%) |
| Two sample test for equality of proportions with continuity correction | | $p = 0.04$ | $p = 0.68$ |
| Total (CF and HB) | 817 | 95 (11.6%) | 152 (18.6%) |

**Table 6**
Counts and percentage of 10% surprises for CF and HB (for FP and FV, and across all variables) for Exp. 1.

| | Counterfactuals (CF) | | | Hypothetical Betting (HB) | | |
|---|---|---|---|---|---|---|
| | Surprises | Non-surprises | Total | Surprises | Non-surprises | Total |
| Initial estimates | 200 (48.2%) | 215 (51.2%) | 415 | 213 (53.0%) | 189 (47.0%) | 402 |
| Revised estimates | 191 (46.0%) | 224 (54.0%) | 415 | 198 (49.3%) | 204 (50.7%) | 402 |

Regarding the first dependent variable, Table 5 shows that there were more revisions for the upper bound ($x_{100}$) than for the lower bound ($x_0$) across debiasing treatments, 18.6% vs 11.6% respectively, and this difference is statistically significant (X-squared = 14.96, df = 1, $p$-value < 0. 01). Moreover, CF has led to slightly more revisions for both bounds compared to HB (the $p$ value is statistically significant for the lower bound, but not for the upper bound).

If we consider the number of subjects that made these judgments, there are 31 subjects that revised the lower bound under the CF condition (i.e., 28.2% of the total) and 25 subjects that revised the same estimate under the HB condition (i.e., 22.7%). Similarly, 35 subjects revised the upper bound under the CF condition (i.e., 31.8%) and 34 under the HB condition (i.e., 30.9%). Not only do these figures confirm that counterfactuals lead to slightly more people revising their judgment, they also indicate that the response to the treatment is judgment-based and not subject-based (large number of subjects making few revisions vs. a small number of people revising most of the variables).

*DV2: proportion of 10% surprises*

Moving to the second dependent variable, we consider as surprises those instances in which the true value falls outside the 90% range of the participant's estimates, hence the denomination of 10% surprises. Table 6 summarizes the number and percentage of 10% surprises before and after employing each debiasing technique. There were many surprises, about 50% for both debiasing techniques vs. the 10% expected result with perfect calibration. We note that the proportion of surprises in revised estimates is slightly lower, although not statistically significant, for counterfactuals compared to hypothetical betting (McNemar's Chi-squared test with continuity correction leading a Chi-squared= 0.73 and $p = 0.39$). Overall, there was no significant effect of the treatments (i.e., CF or HB) in reducing surprises.

*DV3: width of the variable's overall range*

When considering the third dependent variable, the width of the variable's range (initial $x_{100} - x_0$ vs. revised $x'_{100} - x'_0$), we obtain the results shown in Fig. 3 (for CF, the median of the initial width is $\tilde{x}_{CF} = 69.89$ with an inter-quartile range of $IQR_{CF} = 92.08$, and for the revised width the median is $\tilde{x}'_{CF} = 75.00$ with an inter-quartile range of $IQR'_{CF} = 94.62$; for HB, the median of the initial width is $\tilde{x}_{HB} = 55.56$ with $IQR_{HB} = 78.34$, and for the revised width, $\tilde{x}'_{HB} = 59.14$ with $IQR'_{HB} = 82.81$). These differences are significant for both CF and HB (Wilcoxon signed rank test with continuity correction,[4] for CF: $V = 1849$, $z = -6.72$, $p < 0.01$; for HB: $V = 849.5$, $z = -6.86$, $p < 0.01$).

Fig. 4 shows the comparison between treatments of the second dependent variable (DV2), measured as the change in percentage of 10% surprises, where values are calculated as the ratio of the post-treatment surprise proportion to the pre-treatment surprise proportion (i.e., lower ratios are preferable), and the third depen-

[4] The Wilcoxon tests were selected whenever the data were non-normally distributed, as verified with the Shapiro-Wilk normality test.

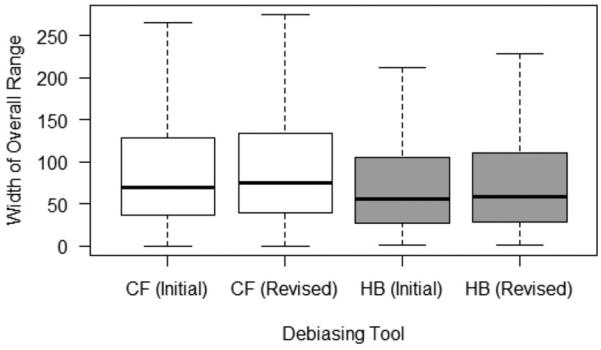



**Fig. 3.** Boxplot of initial ($x_{100}$ – $x_0$) vs. revised ($x'_{100}$ – $x'_0$) width of the overall ranges, for counterfactuals vs. hypothetical betting in Exp. 1.

**Table 7**
Results of the statistical test for range change and for the percentage change in 10% surprises – Exp. 1.

| | Overall Range Change | Change in Surprises |
|---|---|---|
| Elicitation (FP vs FV) | $H = 0.41$, $p = 0.52$ | $H = 0.07$, $p = 0.78$ |
| Treatment (CF vs HB) | $H = 0.27$, $p = 0.61$ | $H = 1.17$, $p = 0.28$ |
| Elicitation*Treatment | $H = 0.43$, $p = 0.52$ | $H = 1.57$, $p = 0.21$ |

dent variable (DV3), where the overall range change is calculated as the ratio of the post-treatment range to the pre-treatment range (i.e., higher ratios are preferable).

No statistically significant result has been found in effectiveness between the tested overprecision debiasing tools (CF vs. HB) and elicitation protocols (FP vs. FV). Specifically, Fig. 4a shows that there is an interaction effect between treatment type and elicitation protocol on the percentage change in 10% surprises, with the combination HB-FV performing slightly better than the others. However, there seems to be no interaction between the elicitation protocol and the treatment on the bounds for the width of the range, as shown in Fig. 4b. We tested the significance of these results using the Scheirer-Ray-Hare test[5] as shown in Table 7.

*DV4: width of the 80% inner range*

Let us consider the fourth dependent variable, the 80% inner range overprecision, measured as initial [$x_{90}$ – $x_{10}$] when using fixed value vs. fixed probability elicitation protocols. The FV protocol led to significantly wider initial inner ranges [$x_{90}$ – $x_{10}$] when compared to the FP protocol (see Fig. 5, $\tilde{x}_{FP} = 33.67$, $IQR_{FP} = 54.67$; $\tilde{x}_{FV} = 48.44$, $IQR_{FV} = 66.84$; Wilcoxon rank sum test with continuity correction, $W = 67{,}258$, $z = -4.51$, $p < 0.01$).

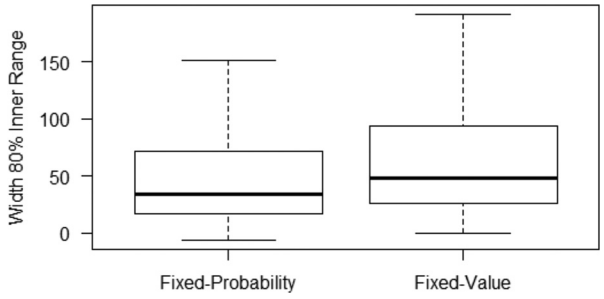



**Fig. 5.** Boxplot of the width of 80% inner range of initial estimates ($x_{90}$ – $x_{10}$) for FP vs. FV in Exp. 1.

### 3.4. Experiment 2

As presented above, neither of the two methods to stretch the ranges tested in Experiment 1 (CF and HB) was very effective in reducing the number of surprises. However, they were both effective in increasing the initial ranges, albeit not sufficiently. We decided to further explore CF efficacy while eliminating HB as the informal feedback from the participants highlighted that HB questions were more difficult to understand when compared to CF questions.

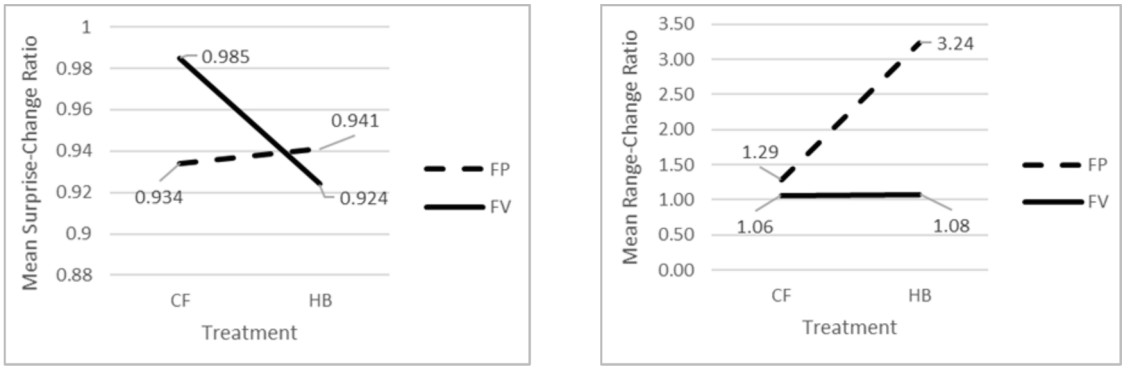
We thus introduced automatic stretching (AS) as a new debiasing method for the bounds for Experiment 2. The aim of this second experiment was to test and compare the effectiveness of CF vs. AS on the tails of the participants' subjective probability distributions and to explore whether there is any behavioral effect of AS on overprecision. We investigate the 80% inner ranges provided in the revised estimates ($x'_{90}$ – $x'_{10}$) as a main indicator of debiasing the extremes.

#### 3.4.1. Methods – experiment 2

One hundred and one students participated in this experiment in the Behavioral Lab of the London School of Economics and Political Science. All participants had been exposed to probability encoding in their studies. Participants were asked to provide estimates for the 10 general knowledge variables (listed in Table 8). Each participant was randomly allocated to one of the following four experimental conditions: FP-CF, FP-AS, FV-CF, or FV-AS.

As in Experiment 1, participants were working on a pre-programmed interactive Excel file which allowed them not only to see, while responding, the cumulative probability distributions in linear piecewise form plotted on the side of the screen, but

[5] The Scheirer-Ray-Hare test is a factorial extension of the Kruskal-Wallis test, which is itself a non-parametric version of a one-way ANOVA. The test assumes equal distributions across all groups and is less powerful than a traditional ANOVA.




a. 10% Surprise Change (lower ratios are preferable) b. Overall Range Change (higher ratios are preferable)

**Fig. 4.** Interaction between treatments and elicitation protocol for Exp. 1.

**Table 8**
General knowledge questions for Exp. 2.

| Ref. Number | Variables | True Value |
|---|---|---|
| 1 | Malfunction rate of an iPhone over a two-year period after purchase [%] | 7.5 |
| 2 | UK present unemployment rate (seasonally adjusted) [%] | 5% |
| 3 | Number of immigrants into the UK in the year ending March 2015 [thousand persons] | 636 |
| 4 | Average number of sunny days (clouds cover less than 30% of the sky) per year in Los Angeles [days] | 186 |
| 5* | Area of Italy [$km^2$] | 301,340 |
| 6 | $CO_2$ per capita in the USA in 2015 [metric tons per person] | 15.9 |
| 7* | Mean global sea level rise in 2100 projected by the 5th IPCC report for the high emission scenario [cm] | 75 |
| 8* | Mean global average increase in the surface temperature in 2100 projected by the 5th IPCC report for the high-emission scenario [ °C] | 4 |
| 9 | Mean loss of sea ice extent in the Artic from 1981 to 2010 [million $km^2$] | 15.5 |
| 10 | Number of sheets of paper that a typical tree produces [sheets] | 8333 |

* = identical general knowledge as in Experiment 1.

**Table 9**
Monotonic distributions and informed responses for Exp. 2 (101 responses per variable).

| Reference Variable | Monotonic Distributions | % Monotonic Distributions | Informed Responses | % Informed Responses |
|---|---|---|---|---|
| 1 | 100 | 99.0% | 57 | 57.0% |
| 2 | 97 | 96.0% | 80 | 82.5% |
| 3 | 100 | 99.0% | 53 | 53.0% |
| 4 | 99 | 98.0% | 98 | 99.0% |
| 5 | 97 | 96.0% | 56 | 57.7% |
| 6 | 97 | 96.0% | 48 | 49.5% |
| 7 | 100 | 99.0% | 47 | 47.0% |
| 8 | 97 | 96.0% | 66 | 68.0% |
| 9 | 101 | 100.0% | 49 | 48.5% |
| 10 | 97 | 96.0% | 33 | 34.0% |
| Totals | 985 | 97.5% | 587 | 59.6% |

also to revise their estimations before proceeding to each subsequent phase. The elicitation task for each variable consisted of two phases, as described next.

In Phase 1, there was the elicitation of the initial estimates for $x_0$, $x_{100}$, $x_{50}$, $x_{10}$, $x_{90}$ (in this order), following the same elicitation protocol for Experiment 1 described previously. The elicitation Phase 2 asked for the revision of the estimates for $x'_0$, $x'_{100}$, $x'_{50}$ (in this order). We replicated FP-CF and FV-CF as in Experiment 1, but now also elicited the 10th ($x'_{10}$) and 90th ($x'_{90}$) revised percentiles, either following FP or FV for this estimate. The auto stretching (AS) treatment consisted of eliciting the initial lower and upper bound ($x_0$, $x_{100}$) then dividing the former by two ($x_0/2$) and multiplying the latter by two ($2\ x_{100}$) and informed the subject that: "You might have missed the true value in your original range. We have thus automatically stretched the original range dividing by ½ the lower bound and multiplying by 2 times the upper bound". Starting with the new extremes $x'_0 = x_0/2$ and $x'_{100} = 2\ x_{100}$, we then re-elicited the 50th ($x'_{50}$), 10th ($x'_{10}$), and 90th ($x'_{90}$) percentiles. The elicitation protocols for AS were either FP or FV.

For example, if auto stretching were employed for the initial distribution shown on the lefthand side of Fig. 2, the lower bound would be automatically adjusted to 20 centimeters and the upper bound to 140 centimeters. The subject would be asked then to re-estimate the 50th ($x'_{50}$), 10th ($x'_{10}$), and 90th ($x'_{90}$) percentiles given these wider bounds.

### 3.4.2. Results - experiment 2

All 101 participants completed the probability estimates for all 10 questions. Again, we considered only estimates that produced monotonic cumulative distributions, and focused on responses that were informed the specific question (i.e., distributions with $x_{50} \leq 1/5x_t$ or $x_{50} \geq 5x_t$). We thus had 985 monotonic distributions (97.5% of the total number of estimates), of which 587 (59.6%) were classified as informed, as detailed in Table 9.

This proportion of informed responses is a significantly lower rate of informed responses than in Experiment 1 (two-sample test for equality of proportions with continuity correction, x-squared=60.99, $p < 0.01$). On closer analysis, the lower informed response rate was largely due to the environmental questions variables (numbered between 7 and 10 in Exp. 1 and between 6 and 10 in Exp. 2), with 80.2% in Experiment 1 vs. 48.1% in Experiment 2 (two-sample test for equality of proportions with continuity correction, x-squared=117.31, $p < 0.01$), while the non-environmental question variables (numbered between 1 and 6 in Exp. 1 and between 1 and 5 in Exp. 2) had similar response rates in both experiments (68.4% and 68.1%, respectively), which were not significantly different. The number of monotonic and informed responses per treatment is shown in Table 10.

If we consider the number of subjects that made these judgments, there are 48 subjects that revised the lower bound under the CF condition (i.e., 47.5% of the total) and 44 subjects that revised the upper bound (i.e., 43.5%). Out of 587 informed distributions, there were 118 revisions of the lower bound (20.1%) and 126 revisions of the high bound (21.5%). We provide below the results of the experiment looking at the six dependent variables described in Section 3.2 considering the distributions that were classified as informed responses.

*DV1: number of judgments revised after the subject was exposed to a debiasing tool*

We analyzed for this first dependent variable only the number the number of revised judgments for CF, as the bound revisions ($x_0$ and $x_{100}$) for AS are revised by default and not decided by the subject (however, we have analyzed the revisions of the 10th and the 90th fractiles under the AS condition). There were 293 informed responses for the CF treatment, with 118 revised $x_0$ (40.27% of total) and 126 revised $x_{100}$ (43.00%). The difference in these proportions were not statistically significant.

*DV2: proportion of 10% surprises*

Table 11 summarizes the number and percentage of 10% surprises before and after employing each debiasing technique (CF and AS, respectively), where initial estimates represent the no-treatment condition. We observed a significant effect of both treatments on the reduction of surprises for both CF (McNemar's Chi-squared test with continuity correction leading a Chi-

**Table 10**
Monotonic and informed responses per treatment for Exp. 2.

| | Fixed-Value (FV) Elicitation | | | Fixed Probability (FP) Elicitation | | |
|---|---|---|---|---|---|---|
| | Monotonic | Informed | % Informed | Monotonic | Informed | % Informed |
| Counterfactual (CF) | 246 | 137 | 55.7% | 257 | 156 | 60.7% |
| Auto Stretch (AS) | 266 | 156 | 58.6% | 216 | 138 | 63.9% |

**Table 11**
Counts and percentage of 10% surprises with CF and AS (for FP and FV, and across all variables) for Exp. 2.

| | Counterfactuals (CF) | | | Automatic Stretch (AS) | | |
|---|---|---|---|---|---|---|
| | Surprises | Non-surprises | Total | Surprises | Non-surprises | Total |
| Initial estimates | 121 (41.3%) | 172 (58.7.0%) | 293 | 106 (36.1%) | 188 (63.9%) | 294 |
| Revised estimates | 102 (34.8%) | 191 (65.2%) | 293 | 53 (18.0%) | 241 (82.0%) | 294 |

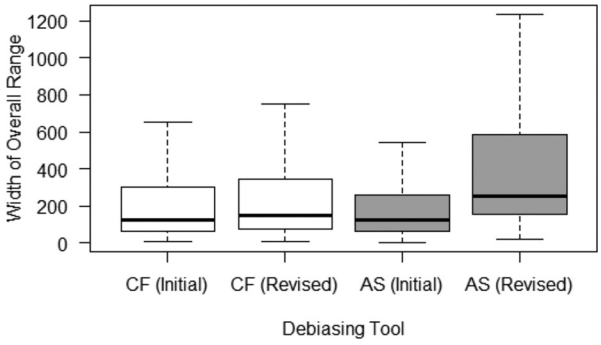



**Fig. 6.** Boxplot of initial ($x_{100} - x_0$) vs. revised ($x'_{100} - x'_0$) width of the overall ranges, for counterfactuals vs. automatic stretching for Exp. 2.

**Table 12**
Results of the statistical test for range change and for the change in percentage of 10% surprises for Exp. 2.

| | Overall range change | Change in% Surprises |
|---|---|---|
| Elicitation (FP vs FV) | $H = 1.38$, $p = 0.24$ | $H = 3.73$, $p = 0.05$ |
| Treatment (CF vs AS) | $H = 65.69$, $p = 0.00$ | $H = 15.84$, $p = 0.00$ |
| Elicitation*Treatment | $H = 0.56$, $p = 0.45$ | $H = 2.24$, $p = 0.13$ |

squared=17.38 and $p < 0.01$) and for AS (Chi-squared=73.70 and $p < 0.01$). Considering the comparison between CF and AS, the same table shows that the proportion of surprises in revised estimates is much lower for AS compared to CF and this difference is statistically significant (McNemar's Chi-squared test with continuity correction leading a Chi-squared= 76.92 and $p < 0.01$).

*DV3: width of the variable's overall range*

When considering the width of the variable's range (initial $x_{100} - x_0$ vs. revised $x'_{100} - x'_0$) we obtain the results shown in Fig. 6 (for CF, initial $\tilde{x}_{CF} = 127.66$ with $IQR_{CF} = 236.17$ and for the revised $\tilde{x}'_{CF} = 148.94$ with $IQR'_{CF} = 269.39$; for AS, initial $\tilde{x}_{AS} = 127.66$ with $IQR_{AS} = 195.48$ and for the revised $\tilde{x}'_{AS} = 254.09$ with $IQR'_{AS} = 431.59$). The difference between the revised and initial ranges for both treatments is significant (Wilcoxon signed rank test with continuity correction, $V = 1435.5$, $z = -9.28$, $p < 0.01$ for CF, and $V = 2547$, $z = -20.39$, $p < 0.01$ for AS).

Fig. 7 shows the comparison between treatments (CF and AS) and elicitation protocols (FP and FV) on the change in percentage of 10% surprises ratio (DV 2) and on the overall range change ratio (DV 3). The figure shows that AS performs better than CF on all dimensions of the analysis (lower 10% surprise ratio and higher overall change ratio). Table 12 confirms the statistical significance of tail treatments on each of the two dimensions. This analysis shows an interaction for the surprise change (Fig. 7a) but no interaction effect between the elicitation protocol and the treatment on the tails for the overall width of the range (Fig. 7b). We tested the significance of these results using the Scheirer-Ray-Hare test as shown in the same table.

*DV4: width of the 80% inner range*

Considering the 80% inner range overprecision, measured as ($x_{90} - x_{10}$), once again the use of the FV protocol led to wider initial inner ranges compared to the FP protocol (see Fig. 8) with similar spread ($\tilde{x}_{FP} = 93.71$, $IQR_{FP} = 152.50$; $\tilde{x}_{FV} = 100.36$, $IQR_{FV} = 143.89$). However, this difference in medians is not statistically significant (Wilcoxon rank sum test with continuity correction, $W = 40{,}512$, $p = 0.21$).

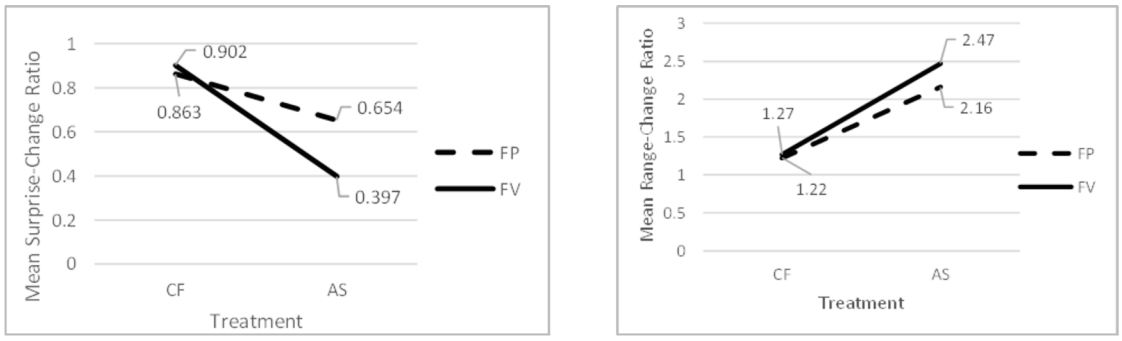
In Experiment 2 we were also interested in understanding the effect of treatments on estimates for the 10% and 90% fractiles, which were elicited from the subjects for both for CF and AS. They are analyzed next.

*DV5: inner range overprecision marginal improvement*

We considered the 80% inner range overprecision marginal improvement (measured as initial $x_{90} - x_{10}$ and revised $x'_{90} - x'_{10}$) when using FV vs. FP. Starting from the CF treatment (boxplot shown in Fig. 9), we noticed that the revised ranges are wider than the initial ones for both FP (initial $\tilde{x}_{FP} = 91.95$, $IQR_{FP} = 156.29$, revised $\tilde{x}'_{FP} = 96.78$, $IQR'_{FP} = 164.79$; Wilcoxon signed rank test with continuity correction, $V = 1083.5$, $z = -2.83$, $p < 0.01$) and for FV (initial $\tilde{x}_{FV} = 106.66$, $IQR_{FV} = 186.86$, revised $\tilde{x}'_{FV} = 119.36$, $IQR'_{FV} = 208.26$; Wilcoxon signed rank test with continuity correction, $V = 1235$, $z = -2.83$, $p < 0.01$). Considering the AS treatment (boxplot shown in Fig. 10), again the revised ranges are wider than the original ones for both FP (initial $\tilde{x}_{FP} = 96.26$, $IQR_{FP} = 149.61$, revised $\tilde{x}'_{FP} = 138.35$, $IQR'_{FP} = 191.60$; Wilcoxon signed rank test with continuity correction $V = 573$, $z = -1.25$, $p < 0.01$) and FV (initial $\tilde{x}_{FV} = 94.58$, $IQR_{FV} = 113.30$, revised $\tilde{x}'_{FV} = 201.63$, $IQR'_{FV} = 292.29$; $V = 0$, $z = -1.25$, $p < 0.01$).

We also analyzed the change in percentage of 20% surprises (i.e., $x_t < x_{10}$ or $x_t > x_{90}$), for the CF and AS treatments under each elicitation protocol (FP or FV), where values are calculated as the ratio of the post-treatment surprise proportion to the pre-treatment surprise proportion, as shown in Fig. 11a. We also considered the ratio of the post-treatment 80% range to the pre-treatment range, i.e., $(x'_{90} - x'_{10})/(x_{90} - x_{10})$, as shown in Fig. 11b.

We tested the significance of these results using the Scheirer-Ray-Hare test, as shown in Table 13. Statistically significant results in effectiveness (measured as the width of the variable's 80% inner range and proportion of 20% surprises) have been found for both the elicitation protocols (with FV performing better than FP) and the tested overprecision debiasing tools (with AS performing better than CF). We observe no interaction between the elicitation protocol and the treatment on the tails for the 20% surprise (Fig. 11a) and for the width of the 80% inner range (Fig. 11b).




a. 10% Surprise Change (lower ratios are preferable) b. Overall Range Change (higher ratios are preferable)

**Fig. 7.** Interaction between treatments and elicitation protocol for Exp. 2.

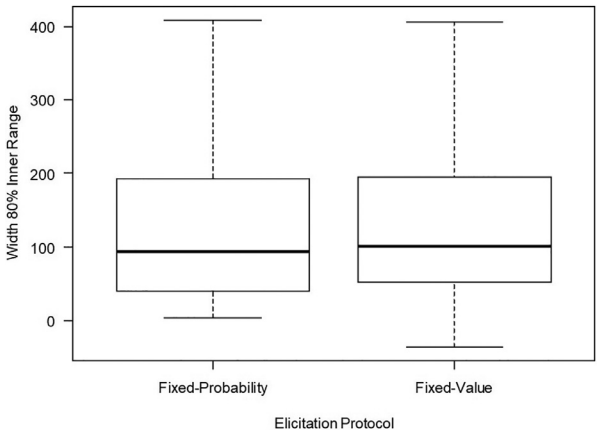



**Fig. 8.** Boxplot of the width of 80% inner range of initial estimates (x90 – x10) for FP vs. FV in Exp. 2.

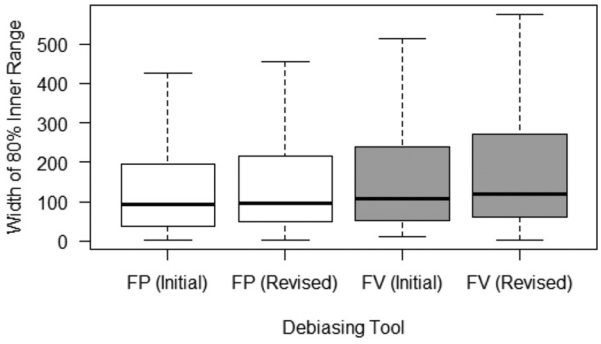



**Fig. 9.** Boxplot of the width of the initial ($x_{90} - x_{10}$) and the revised ($x'_{90} - x'_{10}$) inner range estimated under the FP and FV conditions within the CF treatment in Exp. 2.

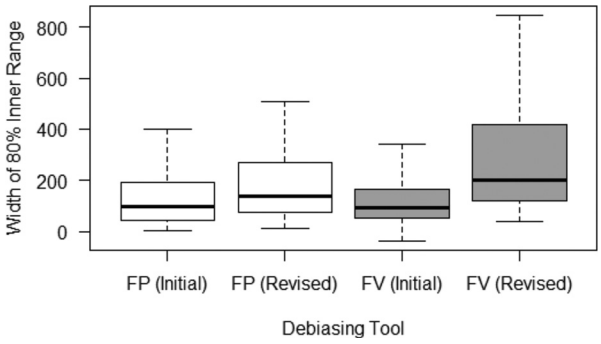



**Fig. 10.** Boxplot of the width of the initial ($x_{90} - x_{10}$) and the revised ($x'_{90} - x'_{10}$) inner range estimated under the FP and FV conditions within the AS treatment in Exp. 2.

**Table 13**
Results of the statistical test for 80% inner range change and for the change in percentage of 20% surprises – Exp. 2.

| | 80% Inner range change | Change in 20% Surprises for inner range |
|---|---|---|
| Elicitation (FP vs FV) | $H = 5.91$, $p = 0.01$ | $H = 9.32$, $p < 0.01$ |
| Treatment (CF vs AS) | $H = 37.42$, $p = 0.00$ | $H = 9.91$, $p < 0.01$ |
| Elicitation*Treatment | $H = 1.94$, $p = 0.16$ | $H = 1.84$, $p = 0.17$ |

*DV6: the proportion of revised judgments that are moving toward the right direction after auto-stretching*

We now consider the sixth dependent variable, i.e., the proportion of revised judgments that are moving toward the right direction after auto-stretching (proportion of $x'_{10}$ becoming lower than $x_{10}$, or $x'_{90}$ becoming higher than $x_{90}$). We tested the significance of the proportions with a two-sample test for equality of proportions with continuity correction (see Table 14). Considering the instances in which $x'_{10}$ became smaller than $x_{10}$ after revision, this table shows that 36.9% of judgments improved under CF and 59.2% under AS. The proportion of no revision for the inner lower bound (i.e., $x'_{10} = x_{10}$) is 34.1% for CF and 6.5% for AS. For the inner upper bound ($x'_{90}$), a much higher proportion of judgments under AS (85.7%) was revised upward in comparison with CF (49.1%). Again, the proportion of no revisions for the inner upper bound ($x'_{90} = x_{90}$) is very small for AS (1.7%) when compared with CF (33.4%). These results are statistically significant.

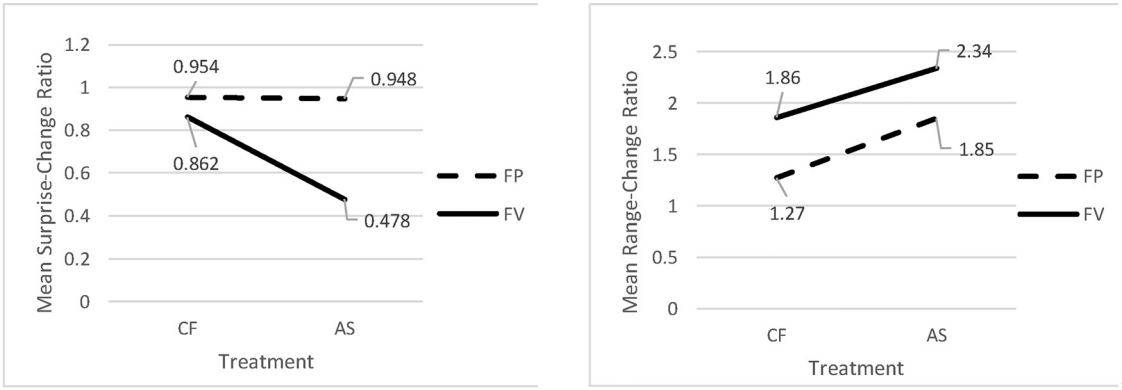
## 4. Summary and discussion

This section compares relevant findings across the two experiments, considering the research question, design and six dependent variables (refer to Section 3.2). We start with the results from observing the number of revisions (DV1). Both CF and HB had low revision rates in Exp. 1, as shown in Table 5. Considering the proportion of surprises (DV2), both CF and HB had a larger proportion of 10% surprises in the revised judgments in Exp. 1 (Table 6), but there was a statistically significant reduction in the number of surprises for both CF and AS in Exp. 2 (Table 11). Noticeably, the new treatment (AS) introduced in Experiment 2 was by far the most successful treatment in reducing surprises.

Considering the width of the overall ranges (DV3) for both experiments (Figs. 3 and 6), every tail treatment has resulted in a significant increase in overall ranges, with ranges under CF becoming wider in Experiment 2 and AS, by design, generating much wider ranges. The analysis of the 80% inner range under the fixed value and fixed probability conditions (DV4) shows that both experiments resulted in FV leading to significantly wider ranges than FP in Exp. 1 (Fig. 5) but not in Exp. 2 (Fig. 8).

**Table 14**
Revision of inner lower and upper bound for Exp. 2.

| | Inner lower bound ($x_{10}$) | | | Inner upper bound ($x_{90}$) | | |
|---|---|---|---|---|---|---|
| | $x'_{10} < x_{10}$ | $x'_{10} > x_{10}$ | $x'_{10} = x_{10}$ | $x'_{90} < x_{90}$ | $x'_{90} > x_{90}$ | $x'_{90} = x_{90}$ |
| CF (Total = 293) | 108 (36.9%) | 85 (29.0%) | 100(34.1%) | 51 (17.4%) | 144 (49.1%) | 98 (33.4%) |
| AS (Total = 294) | 174 (59.2%) | 101 (34.4%) | 19 (6.5%) | 37 (12.6%) | 252 (85.7%) | 5 (1.7%) |
| 2-sample test for equality of proportions | $X^2 = 28.41$, $p < 0.01$ | $X^2 = 1.70$, $p = 0.19$ | $X^2 = 67.80$, $p < 0.01$ | $X^2 = 2.31$, $p = 0.13$ | $X^2 = 87.74$, $p < 0.01$ | $X^2 = 100.04$, df = 1, $p < 0.01$ |




a. 20% Surprise Change (lower ratios are preferable)

b. 80% Range Change (higher ratios are preferable)

**Fig. 11.** Interaction between treatments and elicitation protocol for 80% inner range – Exp. 2.

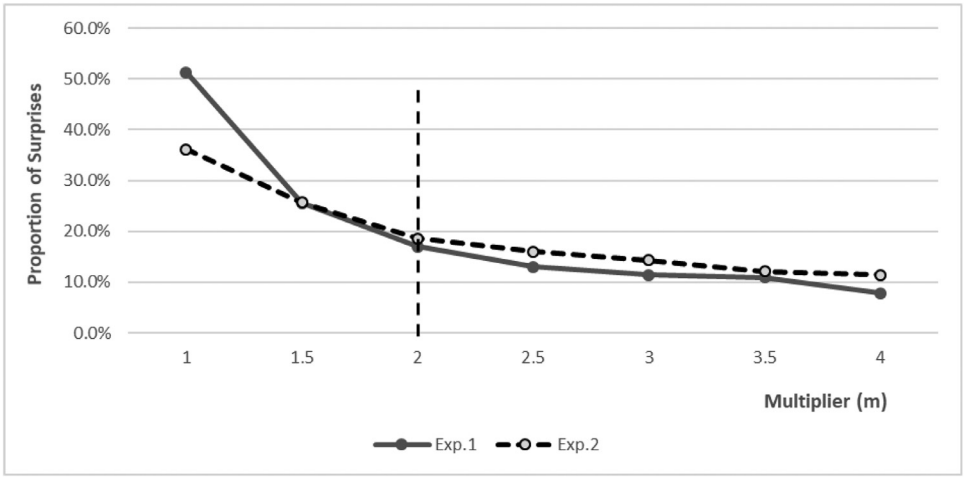



**Fig. 12.** Proportion of 10% surprises as a function of the automatic stretch multipliers for both upper and lower bounds (the vertical dashed line indicates the multiplier used in Exp. 2).

Experiment 2 also enabled us to consider the calibration of the elicited probabilities of the ranges of the variables. We analyzed the influence of the elicitation protocol and the debiasing tool on the 80% inner ranges (DV5). We found that, under both CF and AS treatments, FV led to wider 80% ranges and lower 20% surprise ratios, as shown in Figure 11, thus confirming the preliminary findings from Abbas et al. (2008). At the same time, AS proved more effective than CF in reducing 20% surprises and in widening the 80% inner ranges (DV5), a behavioral effect, stimulating participants to revise their initial inner estimates toward a wider range by a significantly larger extent than those under CF, as shown in Table 14 (DV6).

To summarize, these two experiments led to three key findings. First, our research confirms that overprecision is a pervasive bias (e.g., Moore et al., 2015b), as subjects often did not revise their estimates even if encouraged to do so. Second, traditional "think harder" debiasing strategies (e.g., CF and HB) were not very effective, as demonstrated by both the low number of revisions and the large proportion of surprises for those participants that indeed revised. Third, we found that automatically stretching the initial range provided by the subject was the most effective treatment to reduce overprecision. Furthermore, the FV elicitation protocol generated wider revised 80% inner ranges in the second experiment, in comparison with the FP protocol. Therefore, our recommendation to reduce overprecision in continuous probability distribution elicitation tasks is to stretch automatically the tails, then re-elicit inner fractiles with the FV protocol. Alternatively, plausible ranges can be decided upon by the analyst, making sure that they are wide enough to accommodate different opinions, but small enough to avoid unreasonable or implausible estimates.

One possible reason why auto stretching works better may be that subjects anchor (Tversky & Kahneman, 1974) revised estimates on their initial estimates and insufficiently adjust from there (Moore et al., 2015b). The key decision when employing this treatment is thus which multiplier to use for the initial extreme estimates. We explored the answer to this question by analyzing the effects of different multipliers on the number of surprises using the data from our two experiments. For this analysis, we considered all the monotonic estimates and stretched the initial upper and lower bounds using the multiplier m, with $x'_0 = (1/m)\ x_0$ and $x'_{100} = m\ x_{100}$. The revised median was defined as the mid-point between extremes, i.e.: $x'_{50} = x'_0 + (x'_{100} - x'_0)/2$. The intermediate points $x'_{10}$ and $x'_{90}$ were interpolated linearly between the lower bound and the median (with $p'_{10} = 10\%$) and between the median and the upper bound (with $p'_{90} = 90\%$), respectively.

The graph in Fig. 12 shows the results of using different multipliers on the proportion of 10% surprises. There is a significant marginal reduction in the number of surprises in both experiments when a multiplier on the original bounds is employed. However, the multiplier which we employed in the second experiment ($m = 2$) would only reduce to 17.0% the proportion of surprises in Experiment 1 and to 18.6% in Experiment 2. Only a multiplier of 3.5 would bring them near the 10% target (10.9% of surprises for Experiment 1 and 12.1% for Experiment 2).

## 5. Conclusions

Biases in expert elicitation can lead to low-quality decision analysis or risk analysis models (Montibeller & von Winterfeldt, 2015; Morgan, 2014). Overprecision is a bias that is particularly hard to correct. The research described in this paper has explored debiasing tools that are commonly used to reduce overprecision of expert probability judgments in decision and risk analysis. Our first contribution was a systematic comparison of the effectiveness of these tools, using controlled behavioral experiments to evaluate their effectiveness. We found that traditional "think harder" strategies (e.g., use of counterfactuals or hypothetical bets on the bounds) were not very effective in making people revise their initial judgments. Moreover, in the relatively few instances in which subjects indeed decide to revise, these strategies were not effective in sufficiently expanding the range. Our second contribution consisted of the automatic stretching of the bounds, which proved to be, by far, the most effective debiasing approach in reducing surprises and in improving the 10% and 90% quartile estimates. While we are not the first ones to suggest automatic stretching as a debiasing tool against overprecision (e.g., Moore et al., 2015a; Walters et al., 2017), we are the first, as far as we know, to test it empirically and compare it to other debiasing tools. Our final contribution was to provide preliminary guidelines on how this automatic adjustment may be implemented in decision and risk analysis and a new evidence-based protocol for the elicitation of subjective continuous probability distributions: start by eliciting the extremes of the distribution, then use a suitable multiplier for each extreme and elicit the internal fractiles using the fixed-value elicitation protocol.

In terms of implication for OR, our study indicates some pointers. First, it confirms the challenges associated with the debiasing of judgments, which are often required by experts and decision makers in OR models. Second, the study highlights the importance of systematic and evidence-based assessment of debiasing strategies to guide best practices in OR modeling. Third, it indicates that more automatic debiasing tools may be more effective in the elicitation of judgments for OR modeling than tools that rely solely on the willingness of experts and decision makers to revise their initial estimations. Fourth, and perhaps the most important lesson for OR is that end points do matter, in experiments and in practice. This is especially true for highly uncertain distributions where experts may disagree about both the central tendency and the absolute minima and maxima. Asking experts to provide minima and maxima and then stretching the endpoints is only one of several possible approaches to defining the starting points of an elicitation. Our research, combined with findings from Seaver et al. (1978) and Abbas et al. (2008), shows the merit of this procedure. The downside of this approach is that it may lead to over-stretching for some easier variables and under-stretching for some very difficult variables. Another possibility is for the analyst to pre-specify the endpoints, as suggested by Welsh and Begg (2018), using plausibility and logic to cover the widest possible range without violating natural laws or logical boundary conditions. For example, many uncertain variables are naturally bounded below by zero and plausibility arguments may lead to a logical upper bound. The percent scale is bounded both below and above. The downside of this latter procedure is that it may exaggerate the range and lead to distributions that are too wide. A compromise between expert elicited endpoints and the analyst provided ones is to ask multiple experts for endpoints and then use the minimum of the $x_0$'s and the maximum of the $x_{100}$'s as the endpoints for all experts to use. This is a testable procedure that should be explored in future research.

We also would like to stress some limitations of our study. The first one is that our conclusions come from laboratory experiments with a student population. Indeed, debiasing studies with experts are rare because it is often challenging to obtain samples of experts to perform the required experiments (Graf-Vlachy, 2017). The second limitation is that we had a relatively limited number of informed distributions, particularly in Experiment 2. A possible explanation for this issue is that the questions might have been too hard. Indeed, we have neither pre-assessed the participants' degree of knowledge about each variable, nor their level of confidence on their estimates. Lastly, we employed general knowledge questions; they entail a different task compared to prediction questions about future events, which would have better replicated the elicitation task in settings where predictions are needed.

We can thus identify some areas for further research. To address the above limitations, future experiments could explore the effects of: (i) involving experts in real decision and risk analysis tasks, instead of students making estimates of known quantities, who are guided by a decision analyst/facilitator that can encourage subjects to think harder and further revise their estimates; (ii) employing prediction-type questions more aligned with subjects' knowledge; and (iii) improving the amount of pre-task training on probability elicitation and evaluating this knowledge before the experiment started.

To further explore behavioral implications for overprecision debiasing effectiveness, future research could also investigate the impact of individual differences/personality variables (e.g., age, gender, numeracy, cognitive style, affect) on subjects' estimates. Finally, more research on automatic rules and multipliers would be welcome, as we had encouraging results from these comparative experiments.

Concluding, this research has confirmed that overprecision is indeed a pervasive bias. However, and encouragingly, successful debiasing tools can – and should – be designed to counter this bias and improve the quality of judgments in decision and risk analysis.

## Acknowledgments

The authors are grateful to the three anonymous referees whose feedback helped them to improve the manuscript. They would like to also thank Richard John for his help with the data analysis and Sule Guney for her help with the first experiment. We are also

grateful to the Center for Risk and Economic Analysis of Threats and Emergencies, University of Southern California, for the support with this research project.

## References


Abbas, A. E., Budescu, D. V., Yu, H. T., & Haggerty, R. (2008). A comparison of two probability encoding methods: Fixed probability versus fixed variable values. *Decision Analysis, 5*(4), 190–202.

Aloysius, J. A., Davis, F. D., Wilson, D. D., Taylor, A. R., & Kottemann, J. E. (2006). User acceptance of multi-criteria decision support systems: The impact of preference elicitation techniques. *European Journal of Operational Research, 169*(1), 273–285.

Alpert, M., Raiffa, H., Kahneman, D., Slovic, P., & Tversky, A. (1982). A progress report on the training of probability assessors. *Judgement under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Barberis, N., Thaler, R., Constantinides, G. M., Harris, M., & Stulz, R. (2003). A survey of behavioural science. *Handbook of the economics of finance* (pp. 1053–1128). Amsterdam: Elsevier.

Bedford, T., & Cooke, R. (2001). *Probabilistic risk analysis: Foundations and methods*. United Kingdom: Cambridge University Press.

Block, A. R., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*, 188–207.

Bolger, F., & Harvey, N. (1995). Judging the probability that the next point in an observed time-series will be below, or above, a given value. *Journal of Forecasting, 14*(7), 597–607.

Budescu, D. V., & Du, N. (2007). The coherence and consistency of investors' probability judgments. *Management Science, 54*(11), 1731–1744.

Burgman, M. A. (2016). *Trusting judgements: How to get the best out of experts*. United Kingdom: Cambridge University Press.

Camilleri, A. R., & Newell, B. R. (2019). Better calibration when predicting from experience (rather than description). *Organizational Behavior and Human Decision Processes, 150*, 62–82.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making, 11*(5), 509–526.

Clemen, B. (2001). Assessing 10.50.90s: A surprise. *Decision Analysis, 20*, 2.

Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science, 43*(7), 1029–1045.

Dias, L. C., Morton, A., & Quigley, J. (2018). Elicitation. *The science and art of structuring judgments*. ChamSwitzerland: Springer.

Ferretti, V., Guney, S., Montibeller, G., & von Winterfeldt, D. (2016). Testing best practices to reduce the overconfidence bias in Multi-criteria Decision Analysis. In *Proceedings of the Hawaii International Conference on Systems Sciences (HICSS)* (pp. 1547–1555).

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4*(2), 330–344.

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*(7), 879–895.

Franco, L. A., Hamalainen, R. P., Kunc, M., et al., (2016). Engaging with behavioural OR: On methods, actors, and praxis. *Behavioural operational research: Theory, methodology and practice*. Palgrave.

Franco, L. A., Hamalainen, R. P., Rouwette, E. A. J. A., & Leppanen, I. (2021). Taking stock of behavioural OR: A review of behavioural studies with an intervention focus. *European Journal of Operational Research, 293*(2), 401–418.

Galizzi, M. (2014). What is really behavioral in behavioral health policy? And does it work? *Applied economic perspective, 36*(1), 25–60.

Graf-Vlachy, L. (2017). Like student like manager? Using student subjects in managerial debiasing research. *Review of Managerial Science*, 1–30.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgement. *Judgement and Decision Making, 5*(7), 467–476.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231–237.

Hora, S. C., Edwards, W., Miles, R. F., & von Winterfeldt, D. (2007). Eliciting probabilities from experts. *Advances in decision analysis* (pp. 129–153). Cambridge: Cambridge University Press.

Jain, K., Mukherjee, K., Bearden, J. N., & Gaba, A. (2013). Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Science, 59*(9), 1970–1987.

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical review, 106*, 620–630.

Jonsson, A. G., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgements over time, content domain, and gender. *Personality individual differences, 34*(4), 559–574.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of experimental psychology: Learning, memory and cognition, 25*, 1038–1052.

Keeney, R. L., & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management, 38*(3), 191–201.

Kennedy, P. (1986). The Bayesian approach to research in economic education. *The Journal of Economic Education, 17*(1), 9–24.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*(1), 98–114.

Kirshner, S. N., & Shao, L. (2019). The overconfident and optimistic price-setting newsvendor. *European Journal of Operational Research, 277*(1), 166–173.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118.

Lahtinen, T. J., Hämäläinen, R. P., & Jenytin, C. (2020). On preference elicitation processes which mitigate the accumulation of biases in multi-criteria decision analysis. *European Journal of Operational Research, 282*(1), 201–210.

Langnickel, F., & Zeisberger, S. (2016). Do we measure overconfidence? A closer look at the intervalproduction task. *Journal of Economic Behavior & Organization, 128*, 121–133.

Larrick, R. P., Koehler, D. J., & Harvey, N. (2004). Debiasing. *Blackwell handbook of judgment and decision making*: 1. Blackwell Publishing Ltd.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organization Behavior and Human Decision Processes, 26*(2), 149–171.

Lichtenstein, S., Fischhoff, B., Phillips, L. D., Kahneman, D., Slovic, P., & Tversky, A. (1982). Calibration of probabilities: The state of the art in 1980. *Judgment under uncertainty: Heuristics and biases* (pp. 306–333). Cambridge England: Cambridge University Press.

Mannes, A. E., & Moore, D. A. (2013). A behavioral demonstration of overconfidence in judgment. *APS Association for Psychological Science, 24*(7), 1190–1197.

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

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational behavior and human decision processes, 107*, 179–191.

Mcnamee, P., & Celona, J. (2008). *Decision analysis for the professional* (4th edition). SmartOrg, Inc electronic.

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science, 4*(4), 379–383.

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis, 35*(7), 1230–1251.

Moore, D. A., Carter, A. B., & Yang, H. H. J. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes, 131*, 110–120.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502–517.

Moore, D. A., Tenney, E. R., Haran, U., Gideon, K., & Wu, G. (2015). Overprecision in judgement. *The wiley blackwell handbook of judgement and decision making*. John Wiley & Sons.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 129–140. https://doi.org/10.1177/2372732215600886.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *PNAS, 111*(20), 7176–7184.

Morgan, M. G., & Henrion, G. (1990). *Uncertainty*. Cambridge University Press.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable forecasts of precipitation and temperature? *Bulletin of the American Netereological Society, 53*, 1449–1453.

Nguyen, K. D. (2018). *Evaluating aleatory uncertainty assessment*. University of Southern California PhD Thesis.

Onkal, D., Yates, J. F., Simga-Mugan, C., & Oztin, S. (2003). Professional vs. amateur judgement accuracy: The case of foreign exchange rates. *Organizational behavior and human decision processes, 91*(2), 169–185.

Ortiz, N. R., Wheeler, T. A., Breeding, R. J., Hora, S., & Keeney, R. (1990). Use of expert judgment in NUREG-1150. *Nuclear Engineering and Design*, 1990.

Plous, S. (1995). A comparison of strategies for reducing interval overconfidence in group judgements. *Journal of Applied Psychology, 82*(4), 443–454.

Rapoport, A. (1964). Sequential-decision making in a computer-controlled task. *Journal of Mathematical Psychology, 1*, 351–374.

Ren, Y., & Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management Science, 59*(11), 2502–2517.

Schall, D. L., Doll, D., & Mohnen, A. (2016). Caution! warnings as a useless countermeasure to reduce overconfidence? An experimental evaluation in light of enhanced and dynamic warning designs. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.1946.

Seaver, D. A., von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distribution on continuous variables. *Organisational Behavior and Human Performance, 21*, 379–391.

Soll, J., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory and Cognition,, 30*, 299–314.

Spetzler, C. S., & Stael von Holstein, C. A. S. (1975). Probability encoding in decision analysis. *Management Science, 22*(3), 340–358.

Teigen, K. H., & Jorgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology, 19*, 455–475.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness* (pp. 1–293).

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science (New York, N.Y.), 184*, 1124–1131.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: CUP.

Wallsten, T. S., Forsyth, B., & Budescu, D. V. (1983). Stability and coherence of health experts' upper and lower subjective probabilities about dose-response curves. *Organizational Behaviour and Human Decision Processes, 31*, 277–302.

Wallsten, T. S., Shlomi, Y., Nataf, C., & Tomlinson, T. (2016). Efficiently encoding and modeling subjective probability distributions for quantitative variables. *Decision, 3*(3), 169–189.
Walters, D. J., Fernbach, P. M., Fox, C. R., & Slomanc, S. A. (2017). Known unknowns: A critical determinant of confidence and calibration. *Management Science, 62*(12), 4298–4307.
Welsh, M. B., & Begg, S. H. (2018). More–or–less elicitation (MOLE): Reducing bias in range estimation and forecasting. *EURO Journal on Decision Processes, 6*, 171–212.
Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales-Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research, 258*(3), 801–819.
Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association, 66*(336), 675–685.
Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition., 30*(6), 1167–1175.