frontiers in Psychology

ORIGINAL RESEARCH
published: 23 April 2019
doi: 10.3389/fpsyg.2019.00862

Experimental Investigation on the Elicitation of Subjective Distributions

Carlos J. Barrera-Causil[1], *Juan Carlos Correa*[2] *and Fernando Marmolejo-Ramos*[3]*

[1] Davinci Research Group, Faculty of Applied and Exact Sciences, Metropolitan Technological Institute, Medellín, Colombia, [2] Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia, [3] School of Psychology, The University of Adelaide, Adelaide, SA, Australia

Elicitation methods aim to build participants' distributions about a parameter of interest. In most elicitation studies this parameter is rarely known in advance and hinders an objective comparison between elicitation methods. In two experiments, participants were first presented with a fixed random sequence of images and numbers and subsequently their subjective distributions of percentages of one of those numbers was elicited. Importantly, the true percentage was set in advance. The first experiment tested whether receiving instructions as to the elicitation method would assist in estimating a true value more accurately than receiving no instructions and whether accuracy was determined by the numerical skills of the participants. The second experiment sought to compare the elicitation method used in the first experiment with a variation of a graphical elicitation method. The results indicate that (i) receiving instructions as to the elicitation method does assist in producing estimates closer to a true percentage value, (ii) the level of numerical skills does not play a part in the accuracy of the estimation (Experiment 1), and (iii) although the average estimates of the betting and graphical method are not significantly different, the betting method leads to more precise estimations than the graphical method (Experiment 2). Both studies featured statistical procedures (functional data analysis and a novel clustering technique) not considered in past research on the elicitation of subjective distributions. The implications of these results are discussed in relation to a recent key study.

Keywords: cluster analysis, expert knowledge elicitation, functional data analysis, prior distribution, subjective probability

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Adriano Velasque Werhli,
Fundação Universidade Federal do Rio Grande, Brazil
Mohd Bakri Adam,
Putra Malaysia University, Malaysia

Correspondence:
Fernando Marmolejo-Ramos
fernando.marmolejoramos@adelaide.edu.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *09 September 2018*
Accepted: *01 April 2019*
Published: *23 April 2019*

Citation:
Barrera-Causil CJ, Correa JC and Marmolejo-Ramos F (2019) Experimental Investigation on the Elicitation of Subjective Distributions. Front. Psychol. 10:862. doi: 10.3389/fpsyg.2019.00862

1. INTRODUCTION

"The objective world is no more than a reflection of any person" (Tomás Carrasquilla, 1915)[1].

When people are asked to provide numeric estimates of capital accumulations after a series of annual changes they tend to underestimate the accumulated financial growth even when they are to assume they have enough funds to cushion potential losses (Gonzalez and Svenson, 2014). People's responses thus rely on their subjective experience with and understanding of financial fluctuations and wealth. In other words, information about an uncertain parameter (e.g., an issue of interest) is essential for people to make decisions. Since this information relies on subjective experience acquired over time, it is thus conceivable that a person has various estimates, or proportion

[1] This phrase appears in a short essay titled "Elogio de la viuda sabia" (In praise of the wise widow) published in a Colombian newspaper and for which, to our knowledge, there is no published English translation. The original sentence reads as: "El mundo objetivo no es más que un reflejo de cualquier sujeto."

of estimates, for a specific parameter. This is a key component in Bayesian statistics known as the prior distribution (Berger, 1985).

In some instances, the only possibility is to work with an informative prior distribution, for example, in cases where sample data is unavailable, or the event will occur just once in a life. One illustration of this situation is the determination of the probability that an asteroid destroys the earth. In this case the researcher faces the need of eliciting an informative prior distribution based on personal knowledge (Schlag et al., 2015).

The elicitation of priors consists of extracting information about a parameter of interest from the subjective experience of a person and expressing it as a probability distribution (see **Figure 1** and Anscombe and Aumann, 1963). So, if the elicitation process is applied to a group of persons, then the researcher will end up with several prior distributions. Indeed, several persons may have very different beliefs for the same parameter (Plous, 1993). However, different procedures are available to reduce several prior distributions to one. Winkler (1967, 1968, 1969) studied the problem of consensus in which persons produce several distributions that are combined into a single distribution to be used for posterior Bayesian analysis. For example, Albert et al. (2007) combined opinions from more than one person by using a hierarchical model that considers the bias and precision of the person as well as the consensus and diversity within the group. More recently, expert elicitation has been used in an educational context to foster teacher's self-reflection purposes (Lek and Van de Schoot, 2018).

Obtaining prior information is a very complex procedure that requires quantifying the knowledge of one or several participants in the area under study in order to build personal prior distributions (O'Hagan et al., 2006). Both the process of extracting information from the person's mind and the quantification of it are further affected by factors that increase the complexity of these procedures. Some of these factors are numerical skills and cognitive variables (Albert et al., 2007)[2]. For instance, attitudes have an effect in that they are context dependent (e.g., one's attitude differs when betting on a football game or picking a presidential candidate) (Plous, 1993). Research conducted by Hastorf and Cantril (1954) and Loy and Andrews (1981) are examples of this attitudinal changes. These individual characteristics thus suggest that the individual elicited prior distributions could represent different populations.

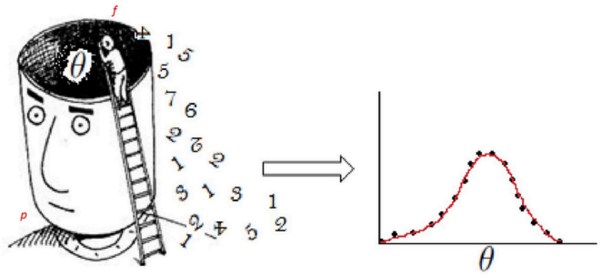

FIGURE 1 | Illustration of the elicitation of priors. **Left**: a person (p) has knowledge-based experience that influences his/her beliefs about an issue of interest (θ). The dark area surrounding θ represents latent cognitive factors that also affect θ itself and the elicitation process. The person who elicits information about θ, or "facilitator" (f), has the task of reaching p's beliefs. **Right**: as beliefs are largely qualitative, f also has to quantify them and render them into a probability distribution that captures what p knows about the issue (more technically, parameter) of interest θ. The distribution of θ can take any form in practice; for illustration purposes we showed a Gaussian distribution.

Due to individual differences (subjective experience), it cannot be guaranteed that different persons have the same grade of expertize or they have been exposed to the same events in their work. This is a default constraint that challenges the comparison of different elicitation techniques. An attempt to lessen this constraint was proposed by Wang et al. (2002) via an objective approach for evaluating an elicitation method that avoids the assumptions and pitfalls of existing approaches. However, their approach does not guarantee that people's knowledge is the same.

Traditionally, because elicitation methods have been compared in non-experimental situations (see Anscombe and Aumann, 2014), their results are not comparable. One reason for this is that people have different levels of knowledge and beliefs. Thus, if an elicitation method is applied to knowledgeable people (i.e., experts), it is very likely that their prior distributions will be good even if the elicitation method is deficient[3]. However, if the level of expertize of the persons is not controlled, it would be difficult to compare the elicitation methods. Also, this is impossible to achieve in real world situations.

One of the first comparisons of elicitation methods was proposed by Schweickert et al. (1987), where three techniques were used to extract the knowledge base from experts on lighting for industrial inspection tasks. Hudlicka (1996) compared three indirect knowledge elicitation techniques based on the number of attributes elicited, the ease with which these data were obtained, and the degree of post-analysis and interpretation required. In the same direction, Zhang (2007) compared three requirements elicitation techniques, but like in Schweickert and Hudlicka, this comparison did not control the level of the experts' knowledge[4].

In this paper, we examine the resulting personal prior distributions about a percentage when participants receive or do not receive instructions about the elicitation process. Importantly, it is ensured that participants receive the same amount of information about a parameter of interest and a computer application is designed to elicit prior distributions via an interactive questionnaire. This interactive elicitation process provides a distribution of estimates for the parameter of interest for each participant. Further, a cluster analysis is carried out with

[2] Some of those cognitive biases and variables are representativeness, availability, adjustment, and anchoring (Kahneman and Tversky, 1972; Tversky and Kahneman, 1974), range-frequency, and overconfidence (O'Hagan, 2019). These cognitive variables play a key role in expert knowledge elicitation and thus need further investigation.

[3] A prior distribution could be considered as good enough when it correctly reflects the expert's previous beliefs and do not suggest quite different plausible regions for the parameters of interest.

[4] Other techniques, such as interviewing, protocol analysis, multidimensional scaling, logistic regression (Wright and Ayton, 1987), and item response theory (Andrade and Gosling, 2018) have also been proposed.

the group who received elicitation instructions in order to detect if participants with different degrees of mathematical and/or statistical skills produce distributions of percentages that better capture the parameter of interest (Experiment 1). The elicitation method used in Experiment 1 is then compared with a variation of a graphical elicitation method (Experiment 2). Functional data analysis (FDA) techniques (see Wang et al., 2016) are used to characterize prior distributions of the participants and a novel method is used for clustering distributions (see Methods section for details) (Barrera and Correa, 2015).

TABLE 1 | Random allocation of participants in the two experimental groups.

Group	G1	G2	G3	Total
Instruction (I)	6	13	6	25
Non-instruction (NI)	8	13	13	34
Total	14	26	19	59

G1, mathematical and statistical skills; G2, mathematical skills only; and G3, non-numerical skills.

2. EXPERIMENT 1

2.1. Participants

Fifty-nine undergraduate students verbally consented to volunteer for the experiment (age$_{\text{range}}$ = 16–27). Of these participants, 14 had approved a course in basic mathematics and statistics at the university (mathematical and statistical skills group, G1; Mean$_{\text{age}}$ = 21.7, *SD* = 2.8, females = 7), 26 had approved basic mathematics at the university (mathematical skills group, G2; Mean$_{\text{age}}$ = 20.9, *SD* = 2.0, females = 11), and 19 had not completed either basic mathematics or statistics at the university (non-numerical skills group, G3; Mean$_{\text{age}}$ = 22.8, *SD* = 2.6, females = 11). The study was carried out according to the Declaration of Helsinki (World Medical Association, 2013) and approved by the local ethics committee at the Metropolitan Technological Institute in Medellín-Colombia (ethical application ref: FGN-006).

2.2. Materials

The experiment was implemented in Microsoft Visual C++ and ran in a room hosting 40 computers with 2GHz Intel(R) Core(TM) i5–4590T processors and 8GB of RAM. Data were analyzed using R (R Development Core Team, 2016) using the add-on packages `fda` (Ramsay et al., 2014) and `fda.usc` (Febrero-Bande and Oviedo de la Fuente, 2012) for FDA, and `cluster` (Maechler et al., 2016) and `clv` (Nieweglowski, 2013) for cluster analysis.

2.3. Procedure

Participants were pseudo-randomly[5] assigned into the elicitation instruction (I) and non-instruction (NI) groups (**Table 1**). Twenty-five participants formed the I group (Mean$_{\text{age}}$ = 21, *SD* = 1.9, females = 12) and 34 participants formed the NI group (Mean$_{\text{age}}$ = 22.2, *SD* = 2.8, females = 17).

Participants in the I and NI groups were informed they would see a random sequence of numbers and images and their task was to determine the percentage of times that the number one appeared (the actual value was 23% and each item was shown for 500 ms and with Interstimulus Interval; ISI = 0). In order to ensure both groups received the same input information, a fixed random order was used for the presentation of items (phase I). This part of the experiment lasted $\sim$ 1 min. The random sequence of items consisted of 26 items; 10 '1' numbers, 10 '2' numbers, and six images.

Subsequently, both groups of participants underwent the elicitation process (phase II) but only those in the I group received instructions as to what the goal of the elicitation process was. The betting elicitation method was used. This is an interactive method in which the computer application asks questions and provides feedback to participants in order to gauge a range of minimum and maximum estimates and probability values for each. Specifically, the participant is asked about the bets he/she would be willing to place for or against the occurrence of a certain event (E). Assuming that x_a is the amount of money that a person is willing to bet for a total of M dollars, and that the utility function is linear, Cooke (1991) showed that the the expected utility of the betting is given by $MkP(E)$ for some constant k, and that the expected utility of x_a is simply kx_a. Setting these two expectations equal it follows that $P(E) = M^{-1}x_a$. In this work, we assume that utility functions are linear.

2.4. Statistical Analyzes

The goal of the elicitation process is to gauge data that can be used to build personal distributions for a specific parameter $\theta \in \Theta$, where Θ is the parameter space of θ.

Thus, let $\mathcal{A}_i$ be fixed subintervals of Θ for the i-th participant ($i = 1, 2, \ldots, n$) such that $\mathcal{A}_i = [\theta_1^i, \theta_m^i]$, with θ_1^i and θ_m^i correspond to the minimum and maximum value that θ can take according to the belief of the i-th participant, respectively. Now, let $\mathcal{A}^* = [\theta_1^*, \theta_m^*] = \bigcup_{i=1}^n \mathcal{A}_i$, and let us consider the grid $\theta_1^* < \theta_2^* < \cdots < \theta_{m-1}^* < \theta_m^*$. Thus, for the values $\{\theta_j^i\}_{j=1}^m$ of θ, each participant provides points (weights) $\{y_j^i\}_{j=1}^m$ which are represented in a graph; these points correspond to the levels of certainty that he/she has about each value in the sequence $\{\theta_j^i\}_{j=1}^m$. For example, if $\theta = \theta_j^3$, then y_j^3 would be the level of credibility that the third participant has about that statement.

For n participants, the above set up will result in a graph with n sequences of discrete and non-negative points $\{\theta_j^i, y_j^i\}_{j=1}^m$ for $i = 1, 2, \ldots, n$. FDA enables to represent the elicited priors in a continuous form by using numeric functions for curve fitting, such as B-splines, and to obtain information about measures that vary on a continuum (e.g., density curves and functional data like time-series). FDA makes use of descriptive measures, such as the functional mean, the (median) deepest curve, the functional boxplot, and analytical measures such as functional clustering methods. These measures are extensions of classical

[5] Initially, participants were randomly assigned into two balanced groups (i.e., 30,30), but a failure in one of the computer rooms and the absence of one participant led to having two unbalanced groups.

statistics methods, such as the mean, median, boxplot and the k-means clustering method (see Ramsay and Silverman, 2005 for technical details).

The cluster analysis is carried out here using a novel hierarchical clustering method, which works as follows. After obtaining the values $\{\theta_j^i\}_{j=1}^m$ and the corresponding certainty levels $\{y_j^i\}_{j=1}^m$ specified by the i-th participant ($i = 1, 2, \ldots, n$), a B-spline is fitted to the $\{y_j^i\}_{j=1}^m$ of each participant. Doing so results in a grid of k points in the (0,1) interval, which corresponds to the range of possible values for the percentages of ones being displayed (in this study, k = 10,000). Further, a matrix of distances between these functions is obtained; this distance measure corresponds to the Hellinger's distance for the curves x^s, x^t and is given by:

$$d(x^s, x^t) = \sqrt{\sum_{j=1}^{k} \left(\sqrt{h_j^s} - \sqrt{h_j^t}\right)^2},$$

where $h_j^s = \frac{y_j^s}{\sum y_j^s}$, $h_j^t = \frac{y_j^t}{\sum y_j^t}$, and y_j^s and y_j^t are the heights of the curves x^s and x^t in the point j, respectively.

Subsequently, the function `hclust` of R is used to construct a hierarchical cluster that uses this Hellinger's metric in combination with the Ward's method (see Murtagh and Legendre, 2014). This novel clustering method is used in this paper as a recent simulation study indicates this proposed method performs better than both agglomerative hierarchical clustering approaches, which combine Eucledian metrics with the unweighted pair-group arithmetic average method, and the Ward's method (Barrera and Correa, 2015).

Location and scale estimations are reported via the Mean and the standard deviation (SD) and bias-corrected-and-accelerated (BCA) (Efron, 1987) confidence intervals (CI) via bootstrap are estimated for values of interest.

2.4.1. Hellinger Distance

We know that Euclidian distance is sensitive to the measurement units of the variables. Therefore, changes in scale affect changes in the distance between individuals. In this paper, we use prior distributions with different symmetries and kurtoses. Thus, changes in the heights of the curves, may represent problems in the Euclidean metric. In scenarios like this the Hellinger distance is more appropriate for density functions and adaptable to discrete distributions (Cuadras and Fortiana, 1993).

There are ways to measure distances between probability measures and these distances do not depend on the parametrizations. In probability and statistics, the Hellinger distance is used to quantify the similarity between two probability distributions without depending on the parametrizations (van der Vaart, 2000).

The Hellinger distance between two probability measures is the L_2-distance between the square roots of the corresponding densities in terms of the elementary probability theory. If we denote the densities as f and g, respectively, the squared Hellinger distance can be expressed as a standard calculus integral (van der Vaart, 2000)

$$\int \left(\sqrt{f(\theta)} - \sqrt{g(\theta)}\right)^2 d\theta.$$

For two discrete probability distributions $P = (p_1 \ldots p_m)$ and $Q = (q_1 \ldots q_m)$, their Hellinger distance is defined as

$$H(P, Q) = \sqrt{\sum_{i=1}^{m} (\sqrt{p_i} - \sqrt{q_i})^2}.$$

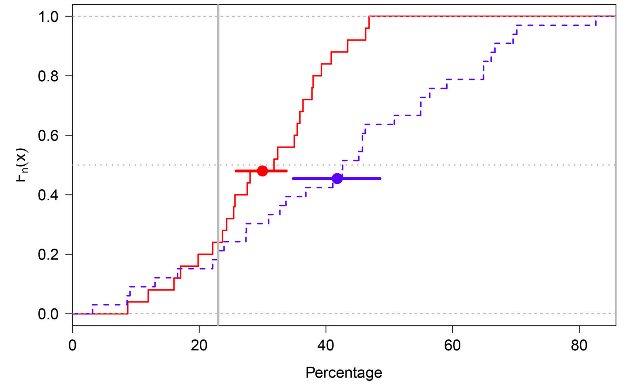

FIGURE 2 | Empirical cumulative distribution function plot of the I (red solid line) and NI (blue dotted line) groups. The groups' means are shown as solid dots. The error bars around the means represent 95%$_{\text{BCA}}$CIs. The gray dotted horizontal line cuts across the groups' median values and the gray solid vertical line represents the true percentage value (23%).

2.5. Results

A test of the difference between the average values of the I and the NI groups was carried out by calculating the median value in each participant's distribution of percentages, and then performing a Welch t-test comparing the means of the two resulting distributions. The parametric pairwise comparison was performed via Q–Q plots (Vélez and Correa, 2015; Loy et al., 2016) and the Shapiro-Wilk (SW) normality test (Marmolejo-Ramos and González-Burgos, 2013), indicating that data in the I and NI groups are normally distributed ($p_{SW} = 0.70$ in both groups). The pairwise comparison also indicated the average percentages of ones in the I and NI groups (I group: Mean = 29.98%, 95%$_{\text{BCA}}$CI = [25.79,33.83]; NI group: Mean=41.77%; 95%$_{\text{BCA}}$CI = [34.62,48.57]) were statistically significantly different ($t_{49.94} = -2.85, p = 0.006$; see **Figure 2**)[6].

These results thus suggest that participants in the NI group had more difficulties than participants in the I group in estimating percentages close to the true value (23%). In other words, explaining what the elicitation process was about (i.e., its

[6] These statistics and the ECDFs were estimated after one outlying observation in the NI group was removed. Such outlier (value = −4.66) was the median percentage of a participant's distribution who exhibited very low and illogical values and the B-spline smoothing simply exacerbated such result. A pairwise comparison between the I and NI groups remained significant even when such outlier was not excluded ($t_{50.27} = -2.44, p = 0.018$).

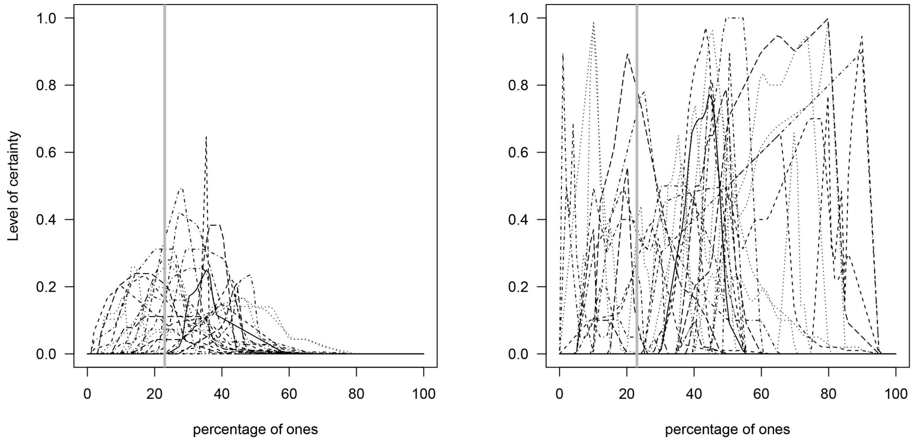

FIGURE 3 | Elicited prior distributions of the percentages given by each of the participants in the I **(Left)** and NI **(Right)** groups. The gray solid vertical line represents the true percentage value (23%).

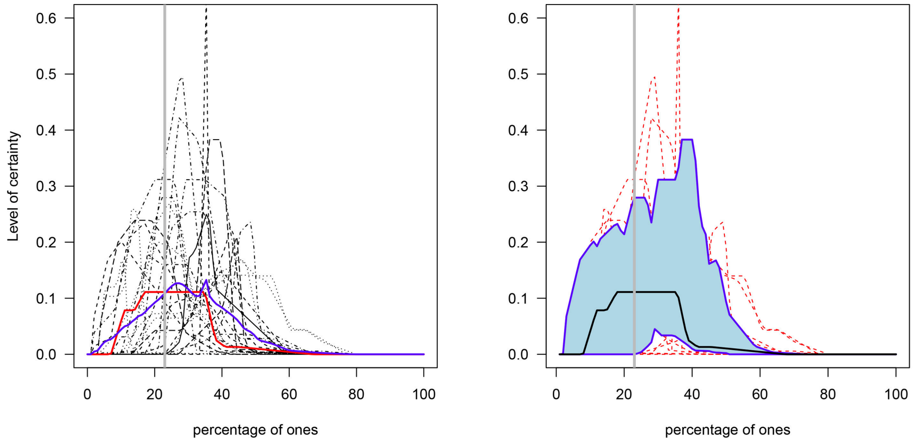

FIGURE 4 | Elicited prior distributions of the percentages in the I group. **Left**: the blue and red lines represent the functional mean and median curves, respectively. **Right**: the solid black line represents the median curve and the lower and upper blue solid lines represent the functional first and third quantiles, respectively (individual distributions are shown in the background in red dotted lines). The gray solid vertical line represents the true percentage value (23%).

goals and steps) assisted participants in the I group to produce estimates closer to the true value (see **Figure 3**). Indeed, a closer look at the distributions obtained in the I group indicates their median deepest curve has narrower spread around the true value than their mean curve (median curve = 25.3% and mean curve = 29.3%) (**Figure 4**).

A cluster analysis was performed on the I group data in order to investigate if members of the G1, G2, and G3 groups (**Table 1**) generated distributions for the percentage of ones that better capture the true value[7]. That is, the goal is to determine whether the three levels of numerical skills are reflected in clusters of skills such that those with the highest level exhibit distributions closer to the true value. The results indicate that around 50% of participants in each of the three groups were grouped in cluster 1, around 33% were grouped in cluster 2, and ∼ 17% were grouped in cluster 3 (see **Table 2**). As **Figures 5**, **6** show, cluster 2 grouped those participants whose distributions' highest levels of certainty were closer to the true value. In clusters 1 and 3 the true value occurred, respectively, on the lower and upper areas of the distributions' tails.

These results thus indicate that the level of numerical skills do not determinate the confirmation of clusters. That is, the clusters were conformed by a mixture of participants representing three levels of numerical skills and the cluster that better captured the true value was indeed no different in this regard. Although unknown cognitive factors (e.g., fatigue) and other demographics (e.g., gender) could have had an effect on the prior distributions obtained for each participant, it is also likely that the method used to build such distributions has had an effect. The elicitation method itself is therefore central to the construction of personal

[7] A permutation test of the equality of two density estimates (Bowman and Azzalini, 2014) indicated the distributions of functional means were different (the FDR-adjusted *p*-values of the three comparisons were close to zero).

prior distributions about a parameter of interest. This experiment showed that the betting (elicitation) method did help participants to build their prior distributions but it is open to question if another elicitation method could have led to a comparable outcome. Experiment 2 had thus the goal of comparing the betting method with a method that elicits knowledge via probability distribution plots.

TABLE 2 | Clusters of the three groups with varying mathematical and/or statistical skills.

Cluster	G1	G2	G3	Total
1	3	6	4	13
2	2	4	1	7
3	1	3	1	5
Total	6	13	6	25

G1, mathematical and statistical skills group; G2, mathematical skills group; and G3, non-numerical skills group.

3. EXPERIMENT 2

3.1. Participants

Thirty-three undergraduate students verbally consented to volunteer in the experiment (Mean$_{\text{age}}$ = 21.9, age$_{\text{range}}$ = 17–29, *SD* = 2.5, females = 16). None of the participants was involved in Experiment 1. The study was carried out according to the Declaration of Helsinki (World Medical Association, 2013) and approved by the local ethics committee at the Metropolitan Technological Institute in Medellín-Colombia (ethical application ref: FGN-006).

3.2. Materials

As in Experiment 1.

3.3. Procedure

Participants were randomly assigned into two groups: the betting (B) and graphical (G) elicitation groups. The betting elicitation method was the same used in Experiment 1, with the consideration that people were instructed before the elicitation session. The graphical elicitation method enables to represent the

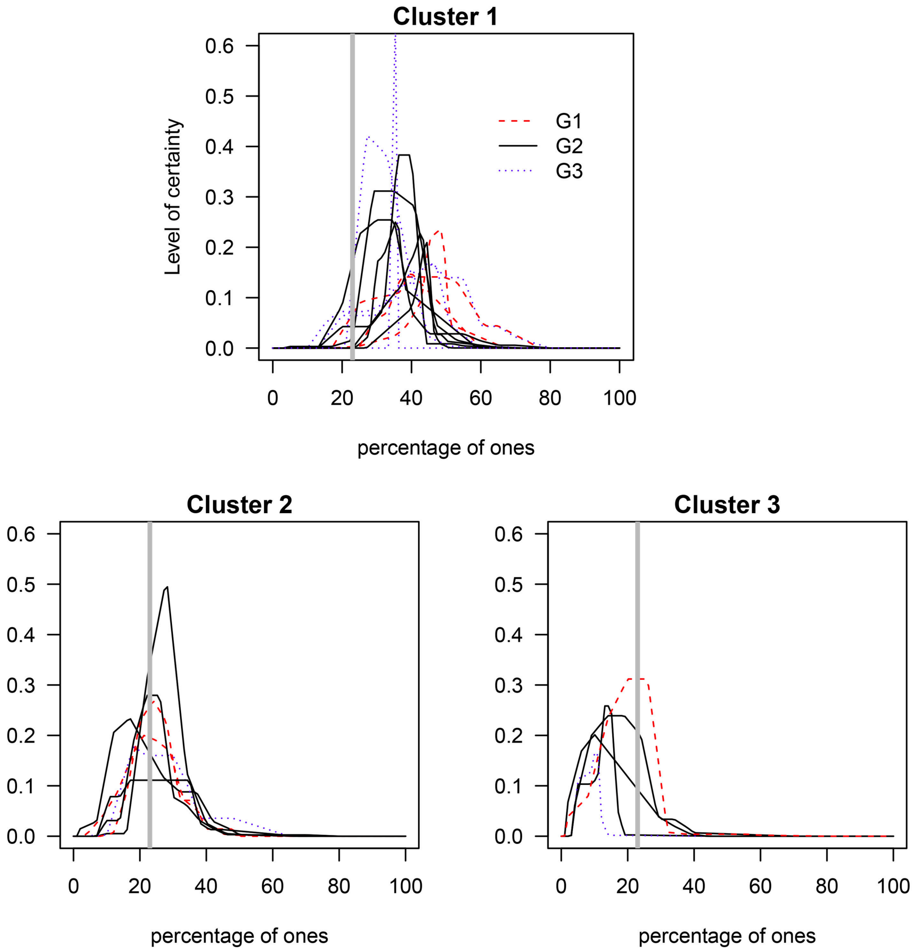

FIGURE 5 | Clusters of the elicited prior distributions of the three groups with varying mathematical and/or statistical skills. G1, mathematical and statistical skills group; G2, mathematical skills group; and G3, non-numerical skills group. The gray solid vertical line represents the true percentage value (23%).

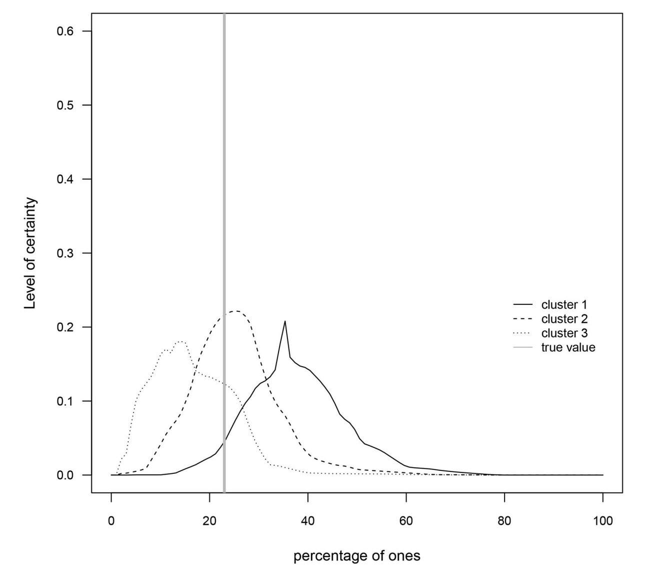

FIGURE 6 | Prior means for each cluster. The gray solid vertical line represents the true value of the percentage of ones (23%).

degree of knowledge about a parameter of interest via histograms, smooth curves (akin to probability density function plots), or points in the Cartesian plane (Chesley, 1975). The ultimate goal is therefore to approximate a probability distribution. In this method, participants are asked to pinpoint on a grid of possible values the level of certainty they have about a parameter. While the X axis represents the values the parameter of interest can obtain, the Y axis represents degrees in probability via adjectives or adverbs of frequency (see Mosteller and Youtz, 1990; Renooij and Witteman, 1990) (**Figure 7**).

Fifteen participants formed the B group ($\text{Mean}_{\text{age}} = 21.5$, $\text{age}_{\text{range}} = 17$–$24$, $SD = 2$, females $= 8$) and 18 participants formed the G group ($\text{Mean}_{\text{age}} = 22.3$, $\text{age}_{\text{range}} = 19$–$29$, $SD = 2.9$, females $= 8$). As in Experiment 1, participants in both groups were informed they would see a random sequence of numbers and images and their task was to determine the percentage of times that the number one appeared (the actual value was 77% and each item was shown for 500 ms with Interstimulus Interval ISI $= 0$). In order to ensure both groups received the same input information, a fixed random order was used for the presentation of items (phase I). This part of the experiment lasted $\sim$ 1 min. The random sequence of items was the same used in Experiment 1. Subsequently, both groups of participants underwent the elicitation process (phase II).

3.4. Statistical Analyzes

As in Experiment 1, FDA tools were used.

3.5. Results

The individual distributions for each elicitation method are shown in **Figure 8**. As in Experiment 1, the median value in each participant's distribution of percentages was estimated and the two resulting distributions were compared via a Welch t-test. This test suggested the groups' mean percentages (B group: M $= 73.19\%$; $95\%_{\text{BCA}}\text{CI} = [65.18,76.83]$; G group: M $= 71.04\%$; $95\%_{\text{BCA}}\text{CI} = [66.27,74.97]$) did not statistically differ ($t_{28.77} = 0.59, p = 0.55$; **Figure 9**)[8].

A visual analysis suggested that although the B group was more left-skewed than the G group (due to two very low median values: 40.4 vs. 60.6%; **Figure 9**), the B group had less variability than the G group ($\text{MAD}_{\text{B}} = 2.99$; $\text{MAD}_{\text{G}} = 7.48$). Indeed, when the two outlying values were removed from the data in the B group (the prior distribution of the participants were illogical respect to their values), this group exhibited average percentages that included the true value (M $= 76.68\%$; $95\%_{\text{BCA}}\text{CI} = [74.66,79.40]$).

4. DISCUSSION

The first study set out to investigate if receiving instructions as to the elicitation method would assist in estimating a true value more accurately than receiving no instructions and whether accuracy was determined by the numerical skills of the participants. The second study sought to compare the elicitation method used in Experiment 1 with a variation of a graphical elicitation method. As to the Experiment 1, the results suggest that receiving instructions as to the elicitation method does assist in producing estimates closer to a true percentage value and the level of numerical skills does not play a part in the accuracy of the estimation. In regard to Experiment 2, the data indicate that although the average estimates of the betting and graphical method are not significantly different, the betting method leads to more precise estimations than the graphical method. Methodologically speaking, both studies featured statistical procedures (FDA tools and a novel clustering technique) not considered in past research on the elicitation of subjective distributions. The implications of these results are discussed in relation to a recent key study.

Grigore et al. (2016) compared the histogram (graphical) and the hybrid elicitation methods in order to obtain subjective probability distributions as to the cost-effectiveness analysis of alternative treatments for prostate cancer. Their results showed that although participants gave more positive ratings to the graphical than to the hybrid method[9] as to the ease of use, the hybrid method was assessed as more accurate. If we entertain the idea that the hybrid method is somewhat akin to the betting method, the results of our Experiment 2 indicate that non-graphical methods seem to lead to estimates closer to the true value (see **Figure 9**).

According to the results of Grigore et al. (2016), the graphical method exhibited less variability around the location parameter than the hybrid method. These results differ from what our

[8] Because the Shapiro-Wilk normality test indicated that the data in the group B did not distribute normally ($p < 0.001$), groups were also compared using the Wilcox rank sum test with continuity correction. This test also showed no difference in the mean percentage of ones ($W = 167.5, p = 0.24$).

[9] In this method, participants are first prompted to give the highest and lowest possible estimates for the parameter of interest; then, intervals within the highest and lowest values are built and the participant is required to assign probabilities to each interval.

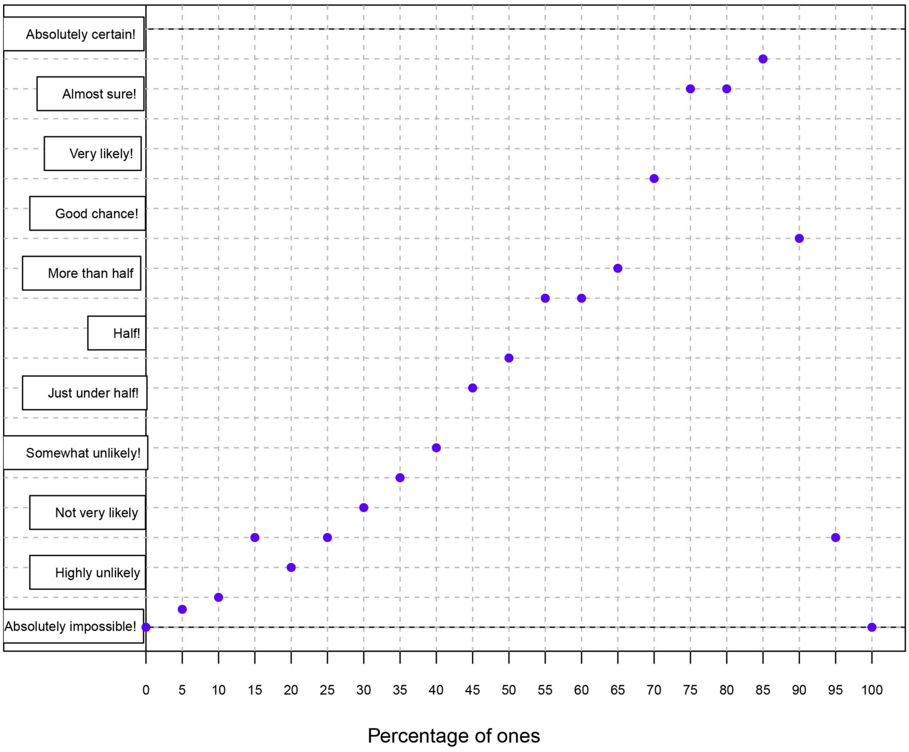

FIGURE 7 | Illustration of the graphical elicitation method. The participant sees a grid without dots and his/her task is to assign a degree of probability (*Y* axis) to each of the percentage values (*X* axis). The *Y* axis represents degrees in probability via 11 linguistic forms (from bottom to top: absolutely impossible, highly unlikely, not very likely, somewhat unlikely, just under half, half, more than half, good chance!, very likely!, almost sure!, and absolutely certain!) (Mosteller and Youtz, 1990).

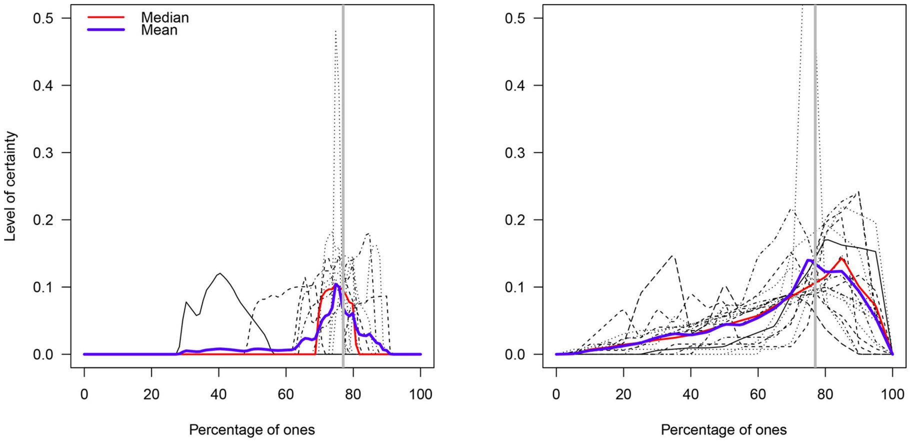

FIGURE 8 | Prior distributions of the percentage of ones in the B and G elicitation groups. The blue and red lines represent the functional mean and median curves, respectively. The gray solid vertical line represents the true percentage value (77%).

Experiment 2 showed in that the graphical method had more variance than the betting method. Interestingly, though, Grigore et al. (2016) found that the location parameters obtained via the graphical method were lower than those given by the hybrid method. Our Experiment 2 also showed that the graphical method lead to lower average estimations of the true parameter than those given by the betting method. Thus, although the graphical method seems easy to use, other methods (e.g., the betting and the hybrid methods) tend to shift participants distributions toward more precise estimates. Having said this, graphical methods need to be tested under different scenarios in order to assess their usability. For example, one could speculate that graphical methods could lead to more homogeneous distributions and accurate estimates than other elicitation

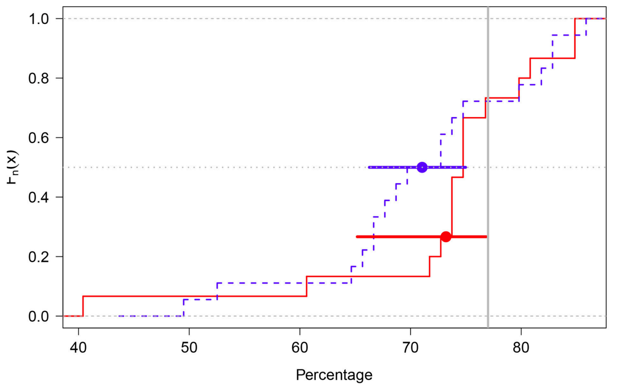

FIGURE 9 | Empirical cumulative distribution function plot of the B (red solid line) and G groups (blue dotted line). The groups' means are shown as solid dots. The error bars around the means represent 95% BCA confidence intervals. The gray dotted horizontal line cuts across the groups' median values and the gray solid vertical line represents the true percentage value (77%).

methods when the parameter of interest refers to a topic relevant to participants who quotidianly rely on graphical displays (e.g., graphic designers, architects, or researchers on statistical graphics). Indeed, research on the assessment of normality of data distributions indicates that graphical displays can be more powerful than traditional goodness of fit tests (see Loy et al., 2016). The key message therefore is that rich information can be extracted from a simple visual assessment of probability distributions. Thus, the elicitation of subjective probabilities via graphical displays demands further investigation.

In Experiment 1, we found that walking the participant through the elicitation method does help in building subjective distributions with low variance around an average estimate that is close to the true value compared to not doing so (see **Figure 2**). Our elicitation sessions (I group in Experiment 1, and B and G groups in Experiment 2) resembled that used by Grigore et al. (2016) (see section "elicitation sessions" in their article). However, a key extra step performed by these authors was to have the participants provide ratings as to the ease of completion of the elicitation method, their face validity, and comments (via open questions) as to the task itself. We did not include such extra questions but we believe it is something to be aware of for future elicitation experiments. In our Experiment 1, though, we assessed participants numerical skills since this was a variable of explicit interest in our study and, as the results indicated, it seems to have no effect on the precision of the true estimate. Nevertheless, we believe that extra information as to the participants (e.g., basic demographics and emotional and cognitive states) needs to be used for weighting their distributions. FDA tools can be used to build subjective distributions and the cluster method proposed in Experiment 1 can be used to re-group subjective distributions according to variables of interest. We believe using these statistical tools in the context of the elicitation of priors enables to build more accurate subjective distributions and perform proper distributional analyzes[10].

A point that we believe needs extra attention and is a central step in familiarizing the participant with the elicitation process is to explain to participants general concepts in probability. Recent brain imaging evidence suggests that while assessing prior probabilities (i.e., the degree of prior certainty) requires frontal brain activation, assessing likelihoods correlates with parietal activation (Kopp et al., 2016). It might be the case that the definitions given to participants as to what probability entails could reflect not only on their brain activations but also on their statistical behavior. In most research of elicitation, probability seems to be understood as a blend between frequency distributions and hypotheses (e.g., opinions) for measuring relative degrees of uncertainty (Monari, 2015). However, probability has also been defined as a pure mathematical concept and as propensity (natural tendency of a concrete thing to be in a certain state or to experience certain changes) (Bunge, 1981). These definitional issues need to be stated and clarified in elicitation studies.

ETHICS STATEMENT

All participants received course credit or participated voluntarily. The study's protocol was approved by the ethics committee of the Instituto Tecnologico Metropolitano. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

CB-C and JC conceived and design of the study. CB-C organized the database. CB-C and FM-R performed the statistical analysis. CB-C wrote the first draft of the manuscript. JC and FM-R wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

FM-R thanks Iryna Losyeva and Alexandra Marmolejo-Losyeva (la patulya) for their support. The authors thank Rosie Gronthos for proofreading the manuscript.

[10] The topic of distributional analysis is essential to research on the elicitation of distributions. Other tools that are designed to deal with distributions and that should be considered in future studies are Generalized Additive Models for Location Scale and Shape (GAMLSS) (Stasinopoulos and Rigby, 2007), Linear Quantile Mixed Models (LQMM) (Geraci and Bottai, 2014), and finite mixture distributions (McLachlan and Peel, 2000).

REFERENCES

Albert, I., Donnet, S., Guihenneuc, C., Low, S., Mengersen, K., and Rousseau, J. (2007). Combining expert opinion in prior elicitation. *Biostatistics.* 8, 1–33. doi: 10.1214/12-BA717

Andrade, J. A., and Gosling, J. P. (2018). Expert knowledge elicitation using item response theory. *J. Appl. Stat.* 45, 2981–2998. doi: 10.1080/02664763.2018.1450365

Anscombe, F. J., and Aumann, R. J. (1963). A definition of subjective probability. *Ann. Math. Stat.* 34, 199–205. doi: 10.1214/aoms/1177704255

Anscombe, F. J., and Aumann, R. J. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.* 6, 103–128. doi: 10.1146/annurev-economics-080213-040927

Barrera, C., and Correa, J. (2015). *Analysis of the Elicited Prior Distributions Using Tools of Functional Data Analysis*. PhD thesis, Universidad Nacional de Colombia Sede Medellín, Medellín, Colombia.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis. 2nd Edn.* New York, NY: Springer-Verlag.

Bowman, A. W., and Azzalini, A. (2014). *R Package sm: Nonparametric Smoothing Methods (version 2.2-5.4)*. University of Glasgow, UK and Università di Padova, Italia.

Bunge, M. (1981). Four concepts of probability. *Appl. Math. Model.* 5, 306–312. doi: 10.1016/S0307-904X(81)80051-0

Chesley, G. (1975). Elicitation of subjective probabilities: a review. *Account. Rev.* 50, 325–337.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York, NY: Oxford University Press.

Cuadras, C., and Fortiana, J. (1993). Aplicacion de las distancias en estadistica. *Qestiió.* 37, 39–74.

Efron, B. (1987). Better bootstrap confidence intervals [with discussion]. *J. Am. Stat. Assoc.* 82, 171–200. doi: 10.1080/01621459.1987.10478410

Febrero-Bande, M., and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda.usc. *J. Stat. Softw.* 51, 1–28. doi: 10.18637/jss.v051.i04

Geraci, M., and Bottai, M. (2014). Linear quantile mixed models. *Stat. Comput.* 24, 461–479. doi: 10.1007/s11222-013-9381-9

Gonzalez, N., and Svenson, O. (2014). Growth and decline of assets: on biased judgments of asset accumulation and investment decisions. *Polish Psychol. Bull.* 45, 29–35. doi: 10.2478/ppb-2014-0005

Grigore, B., Peters, J., Hyde, C., and Stein, K. (2016). A comparison of two methods for expert elicitation in health technology assessments. *BMC Med. Res. Methodol.* 16:3. doi: 10.1186/s12874-016-0186-3

Hastorf, A. H., and Cantril, H. (1954). They saw a game: a case study. *J. Abnorm. Soc. Psychol.* 49, 129–134. doi: 10.1037/h0057880

Hudlicka, E. (1996). "Requirements elicitation with indirect knowledge elicitation techniques: comparison of three methods," in *Proceedings of the Second International Conference on Requirements Engineering* (Colorado Springs, CO: IEEE), 4–11.

Kahneman, D., and Tversky, B. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kopp, B., Seer, C., Lange, F., Kluytmans, A., Kolossa, A., Fingscheidt, T., et al. (2016). P300 amplitude variations, prior probabilities, and likelihoods: a bayesian erp study. *Cogn. Affect. Behav. Neurosci.* 16, 911–928. doi: 10.3758/s13415-016-0442-3

Lek, K., and Van de Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Front. Educ.* 3:82. doi: 10.3389/feduc.2018.00082

Loy, A., Follett, L., and Hoffmann, H. (2016). Variations of q-q plots: the power of our eyes. *Am. Stat.* 2, 202–214. doi: 10.1080/00031305.2015.1077728

Loy, J. W., and Andrews, D. S. (1981). They also saw a game: a replication of a case study. *Replicat. Soc. Psychol.* 1, 45–49.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2016). *Cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.5—For new features, see the 'Changelog' file (in the package source).

Marmolejo-Ramos, F., and González-Burgos, J. (2013). A power comparison of various normality tests of univariate normality on ex-gaussian distributions. *Methodology.* 4, 137–149. doi: 10.1027/1614-2241/a000059

McLachlan, G., and Peel, D. (2000). *Finite Mixture Distributions*. New York, NY: John Wiley and Sons.

Monari, P. (2015). Considerations on probability: from games of chance to modern science. *Statistica.* 75, 345–360. doi: 10.6092/issn.1973-2201/6223

Mosteller, F., and Youtz, C. (1990). Quantifying probabilistic expressions. *Stat. Sci.* 5, 2–12. doi: 10.1214/ss/1177012242

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif.* 31, 274–295. doi: 10.1007/s00357-014-9161-z

Nieweglowski, L. (2013). *clv: Cluster Validation Techniques*. R package version 0.3-2.1.

O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *Am. Stat.*, 73(Sup1), 69–81. doi: 10.1080/00031305.2018.1518265

O'Hagan, A., Buck, E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. New York, NY: John Wiley and Sons.

Plous, S. (1993). *The Psychology of Judment and Decision Making*. New York, NY: McGraw-Hill, Inc.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramsay, J., and Silverman, B. (2005). *Functional Data Analysis*. Berlin: Springer.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.3.

Renooij, C., and Witteman, C. (1990). Talking Probabilities: Communicating Probabilistic Information With Words and Numbers. *Int. J. Approx. Reason.* 22, 169–194.

Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Exp. Econ.* 18, 457–490. doi: 10.1007/s10683-014-9416-x

Schweickert, R., Burton, A., Taylor, N., Corlett, E., Shadbolt, N., and Hedgecock, A. (1987). Comparing knowledge elicitation techniques: a case study. *Artif. Intell. Rev.* 1, 245–253.

Stasinopoulos, D. M., and Rigby, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in r. *J. Stat. Softw.* 23, 1–46. doi: 10.18637/jss.v023.i07

Tversky, B., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science.* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Vélez, J. I., and Correa, J. C. (2015). A modified Q-Q plot for large sample sizes. *Comun. Estad.* 8, 163–172. doi: 10.15332/s2027-3355.2015.0002.02

Wang, H., Dahs, D., and Druzdel, M. (2002). A method for evaluating elicitation schemes for probabilistic models. *IEEE Trans. Syst. Man Cyber. B* 32, 38–43. doi: 10.1109/3477.979958

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Review of functional data analysis. *Annu. Rev. Stat. Appl.* 3, 257–295. doi: 10.1146/annurev-statistics-041715-033624

Winkler, R. (1967). The quantification of judgment: some methodological suggestions. *J. Am. Stat. Assoc.* 62, 1105–1120.

Winkler, R. (1968). The consensus of subjective probability distributions. *Manage. Sci.* 15, B61–B75.

Winkler, R. (1969). Scoring rules and the evaluation of probability assessors. *J. Am. Stat. Assoc.* 64, 1073–1078.

World Medical Association (2013). Ethical principles for medical research involving human subjects. *J. Am. Med. Assoc.* 310, 2191–2194. doi: 10.1001/jama.2013.281053

Wright, G., and Ayton, P. (1987). Eliciting and modelling expert knowledge. *Decis. Support Syst.* 3, 13–26.

Zhang, Z. (2007). "Effective requirements development-a comparison of requirements elicitation techniques," in *INSPIRE 2007* (Tampere).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.