Delft University of Technology

Decision Making Under Uncertainty: Applying Structured Expert Judgment

SEJ2x

# Using Structured Expert Judgment to predict a university ranking in the Humanities

*Nicolas Robinson-Garcia*
Delft Institute of Applied Mathematics, TU Delft
April 27, 2020

**Abstract**

This report is an assignment for the MOOC **Decision Making Under Uncertainty: Applying Structured Expert Judgment**. The goal of the exercise it to familiarize with the application of Structured Expert Judgment, the design and implementation of an elicitation and the reporting of findings. Structured expert judgment uncertainty quantification is a methodology developed by R. Cooke [3] at TU Delft and used in situations where decision making takes place despite a lack of information. This course is part of the Training Plan of the MSCA project Unveiling the Ecosystem of Science: A Contextual Perspective on the Many Roles of Scientists funded by the LEaDing Fellows Programme under the Marie Sklodowska-Curie grant agreement No 707404.

The exercise aims at reporting a hypothetical ranking of Spanish universities in the field of Humanities. While Structured Expert Judgment is usually applied in fields in which there is a lack of data, in the particular example displayed here the starting point is different. We are analyzing a field in which there is a lack of consensus towards which data and how such data should be used to assess performance in these fields. The experts were directed to provide answers on the estimated rank a university should receive based on their knowledge and opinion. No indication as to which indicator should they used was given, hence the ranking should be considered as an overall ranking, comprising not only research performance, but also teaching and knowledge transfer. Having said this, as an exploratory exercise, the results should not be in any case considered as definite.

# Contents

# 1 Background of the study and motivation

The purpose of this study is to construct a universities' ranking of Spanish universities in the Humanities fields. University rankings are highly controversial and criticized in the scientific literature due to the lack of consensus on the criteria employed arguing for serious conceptual and methodological limitations [7]. Most university rankings rely heavily on research output and citation-based indicators, and even when they do not, results show that much of their actual position in ranks can be explained by bibliometric indicators [6]. In principle, this is a field in which lack of data is not an issue and hence should not qualify for the use of structured expert judgment uncertainty quantification. But this lack of consensus, partly reflected on the heterogeneity of missions, institutional statements and peculiarities of higher education institutions [2] has led to the use of experts as a means to quantify the prestige or reputation of institutions (e.g., Times Higher Education World University Rankings). This has been highly controversial, as in no way has the expertise of these individuals been proven or quantified [1].

The fields of the Humanities have been, to a large extent, ignored in this whole debate. They have little or no weight in rankings as they are considered *problematic* due to the incapability of bibliometric indicators to adapt to the specificities of the communication and dissemination mechanisms used in these fields [4, 5]. The term Humanities is usually used to give room to a wide range of very diverse fields which range from linguistics to fine arts or anthropology. In this sense, it seems adequate that experts' judgment should be used to provide a fair assessment on the positioning of universities in these fields, as they should be able to balance and combine these different aspects based on their own knowledge, adapting their judgment to the heterogeneity of disciplines, as well as the variety of outputs and communication mechanisms used in these fields.

# 2 Elicitation design

## 2.1 Description of experts

The elicitation exercise was designed as follows. 11 experts on research evaluation were selected and contacted. Two types of experts were selected, those specialized specifically on the evaluation of the Humanities fields, and those more familiarized with the methodological challenges and constraints of rankings, with a deep knowledge of the Spanish university system. Out of these 11 experts, four accepted to participate in the elicitation exercise. Next we briefly describe each of the experts who participated in the elicitation.

**Elena Castro Martinez.** Elena is Tenured Scientist at the Spanish Council for Scientific Research (CSIC). Among other research interests she has worked on the analysis of knowledge transfer mechanisms in the Arts & Humanities. Institutional profile

**Domingo Docampo Amoedo.** Domingo is Full professor on Signal Theory at the University of Vigo. He was Rector between 1998 and 2006. Since then he has specialized on the analysis of university rankings and conducted several reports for Spanish, French and Australian universities. Institutional profile

**Julia Olmos Peñuelas.** Julia is Associate Professor at the Department of Business Management at the University of Valencia. Her interests revolve around the production and transfer of knowledge. During her PhD she studied science-society interactions in the Social Sciences and Humanities. Institutional profile

**Daniel Torres-Salinas.** Daniel teaches at the Department of Information and Communication at the University of Granada. His is specialized on bibliometrics and is CEO of a spin-off specialized on conducting institutional reports on research performance. Personal website

After agreeing to participate, experts received information on the elicitation exercise (see Appendix A) in Spanish language in which a brief explanation of the exercise was provided but no questions were given at that stage. An individual Skype call was set per expert to conduct the elicitation. These calls have an average duration of 30 – 40 minutes. During the call, the process was explained once again and experts could ask further questions if needed. Prior to querying them, they were invited to explain out loud the reasoning the followed on the answers provided.

## 2.2 Elicitation questions

During the Skype call, experts were asked to respond to 24 questions: 10 calibration questions and 14 questions of interest. For each question they were asked to provide three estimates, the median, the 5th percentile and the 95th percentile. There was no order enforced and while some experts would first determine the 90% uncertainty threshold, others would provide their best estimate and construct the threshold based on it.

The calibration questions were designed to provide information on two peculiarities: 1) their expertise on the Spanish Higher Education system, and 2) their knowledge on the Arts & Humanities fields. The information

| University | Region | City |
|---|---|---|
| Univ Alcala | Com. de Madrid | Alcala de Henares |
| Univ Aut Barcelona | Catalunya | Barcelona |
| Univ Aut Madrid | Com. de Madrid | Madrid |
| Univ Barcelona | Catalunya | Barcelona |
| Univ Complutense | Com. de Madrid | Madrid |
| Univ Granada | Andalusia | Granada |
| Univ Jaume I | Com. Valenciana | Castellon |
| Univ Pais Vasco | Basque Country | Bilbao |
| Univ Pol Valencia | Com. Valenciana | Valencia |
| Univ Pont Comillas | Com. de Madrid | Madrid |
| Univ Salamanca | Castilla Leon | Salamanca |
| Univ Sevilla | Andalusia | Seville |
| Univ Vigo | Galicia | Vigo |
| Univ Zaragoza | Aragon | Zaragoza |

Table 1: List of universities ranked.

used to develop the calibration questions was retrieved from the National Institute of Statistics [1] as well as from reports conducted by the Spanish Ministry of Universities. The calibration questions can be consulted in Appendix A.

Regarding the questions of interest, experts were asked to rank 14 Spanish universities within a rank of 88 Spanish universities (which is the actual number of universities in Spain currently). The selection of ranked universities was made randomly, trying to showcase both large, medium and small universities. In all cases they should include at least one programme in the Humanities. These are shown in Table 1 along with their location.

## 3 Experts' performance

In Table 2 the performance of experts is reported. Four indicators are shown for each expert, their calibration score, their information score (considering only calibration questions), their total information score and their normalized weight. As observed, E1 and E2 are the two experts reporting better calibration scores. E4 shows the lowest calibration scores and hence is the one for whom the correct answer was most times out of his 90% confidence range. However, E4 shows the highest information score, that is, her answers where the most informative. This was followed by E1, with E3 being the least informative expert of the four.

| Expert ID | Calibration | Inf. Score (Cal) | Inf. Score | Norm. weight |
|---|---|---|---|---|
| **E1** | 0.075 | 0.770 | 0.825 | 0.543 |
| **E2** | 0.075 | 0.601 | 0.71 | 0.424 |
| **E3** | 0.006 | 0.386 | 0.590 | 0.002 |
| **E4** | 0.001 | 0.987 | 1.08 | 0.001 |

Table 2: Experts' performance indicators

Overall, all five experts reported more informative answers when responding to the questions of interest, showing a higher information score than when only considering the calibration questions. Overall, when considering a performance weighted decision maker (PWDM), this would account mainly for E1 and E2 responses, with weights below 0.005 for E3 and E4 on the final estimation.

Figure 1 shows the performance of each expert by calibration question. As observed, question 9 has been the most problematic with all experts performing well off range of the realization. Only in questions 5 all experts reported a 90% confidence range which contained the realization.

[1] http://www.ine.es

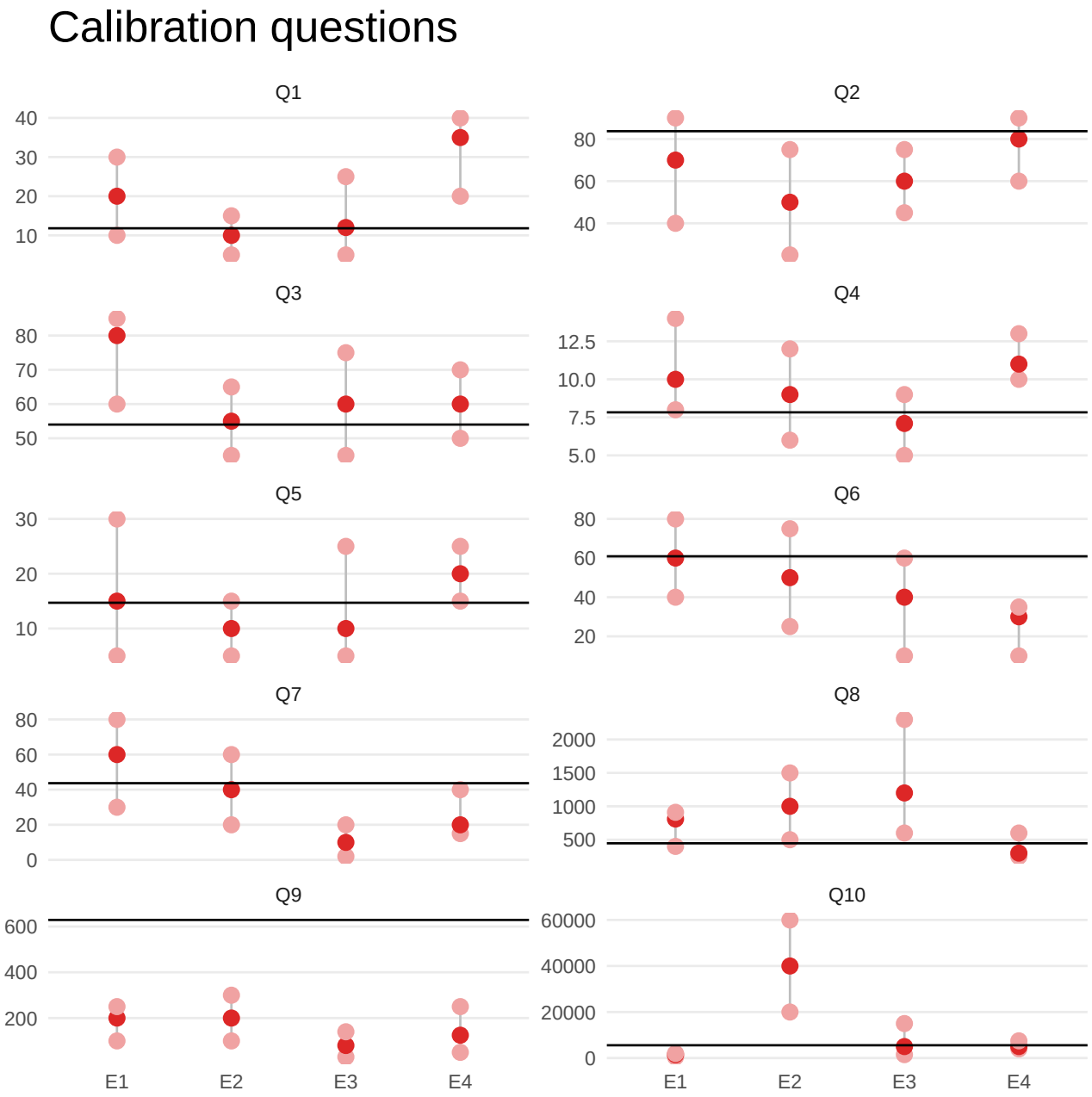



Figure 1: Experts' perfomance for 10 calibration questions. Percentiles 5, 50 and 95 are reported by expert. Horizontal line denotes realization.

## 4 Results

In this section we report the final predictions on the Humanities rankings based on experts' opinions. Five rankings are reported based on the type of decision maker chosen:

- **PWDM**. The Performance Weighted Decision Maker considers the calibration and information scores reported by the experts and weights their predictions accordingly using the weights shown in Table 2.
- **EWDM**. The Equal Weight Decision Maker considers the same level of expertise for all experts, disregarding their calibration performance.
- **PWDM Optimized**. As the PWDM, the optimized version considers also the estimates of the PWDM in its calculation.
- **PWDM Item**. The difference with the PWDM is that it considers experts' performance by question.

- **PWDM Optimized Item**. As the PWDM item, the optimized version considers also the estimates of the PWDM Item in its calculation.

| Expert ID | Calibration | Inf. Score (Cal) | Inf. Score |
|---|---|---|---|
| **PWDM** | 0.683 | 0.277 | 0.302 |
| **EWDM** | 0.683 | 0.116 | 0.105 |
| **PWDM Optimized** | 0.683 | 0.325 | 0.373 |
| **PWDM Item** | 0.493 | 0.292 | 0.31 |
| **PWDM Optimized Item** | 0.493 | 0.342 | 0.404 |

Table 3: Decision makers' performance indicators

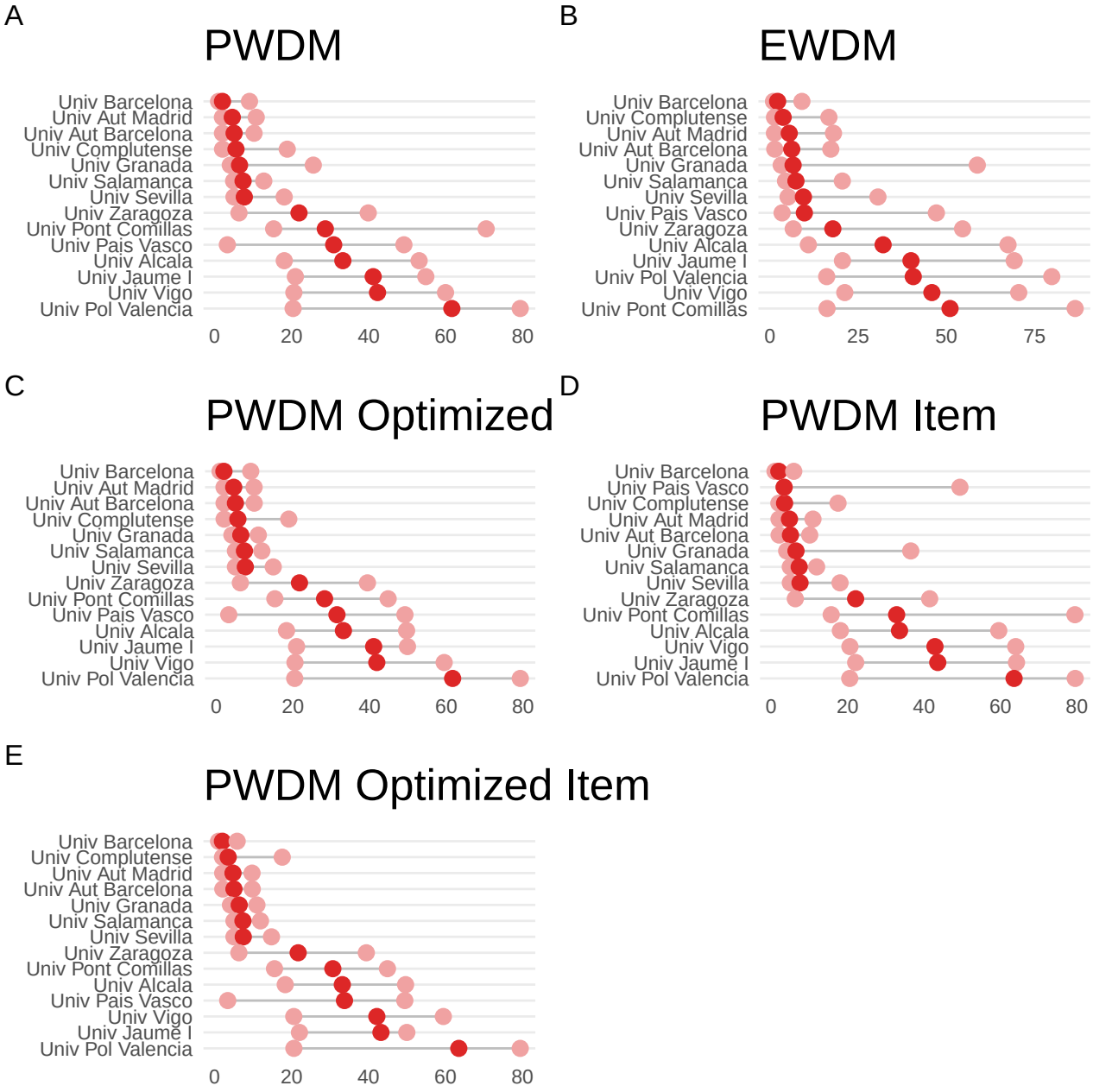



Figure 2: Predicted ranks in a Humanities ranking of Spanish universities based on different decision makers: **A**, performance weighted decision maker, **B**, equal weight decision maker, **C**, performance weighted decision maker optimized, **D**, performance weighted decision maker by item, **E**, performance weighted decision maker optimized by item.

Table 3 shows the decision makers' performance according to the same variables reported in Table 2. As observed, PWDM Optimized and PWDM Optimized Item are the two decision makers that better balance calibration scores with information scores, although the former shows better calibration scores and hence should be prioritized when considering predictions.

While all five decision makers consider the University of Barcelona as the top Spanish university in the Humanities (Figure 2), we do observe some differences on the rest of the positions. For instance, while PWDM, EWDM, PWDM Optimized and PWDM Optimized Item consider Univ Pais Vasco to have an intermediate position, PWDM Item, considers it to be the second of the 14 ranked universities. Also, with the exception of the EWDM and PWDM Item, we observe that the uncertainty threshold is much smaller with top positions than with lower position, something to be expected in university rankings.

# 5 Discussion and recommendations

The present study showcases an example on how Structured Expert Judgment can be applied to the field of research evaluation, specifically to performing university rankings in highly contested fields like the Humanities, where no overall consensus on the criteria used for assessment exists. While the actual outcome of the final ranking is not the main interest of the exercise, but the actual implementation of the methodology, we observe interesting patterns on the performance of the decision makers as uncertainty grows as we descent in the list of ranked universities. Furthermore, despite the unbalanced performance of experts (mainly E1 and E2 were considered for PWDM and its variants), there are not that many differences between the PWDM ranks and the EWDM rank.

Experts were chosen following to criteria: 1) expertise on asssessment of universities and national evaluation systems, and 2) expertise on assessment in the humanities. As a result, two experts of each group were considered, however, their performance was quite balanced and we did find certain degree of consensus on their final predictions. Having said that, we must note that their performance was quite poor, at least for two of the experts, which suggests that, if implemented seriously, a larger pool of experts should be considered to improve the reliability of results.

# References


Bougnol, M.-L., & Dulá, J. H. (2015). Technical pitfalls in university rankings. *Higher Education*, *69*(5), 859–866.

Collini, S. (2012). *What are universities for?* Penguin UK.

Cooke, R., et al. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand.

Hicks, D. (2004). The four literatures of social science. In *Handbook of quantitative science and technology research* (pp. 473–496). Springer.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

Robinson-Garcia, N., Torres-Salinas, D., Herrera-Viedma, E., & Docampo, D. (2019). Mining university rankings: Publication output and citation impact as their basis. *Research Evaluation*, *28*(3), 232–240.

Van Raan, A. F. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143.

# 6 Appendix A. Expert Elicitation Protocol

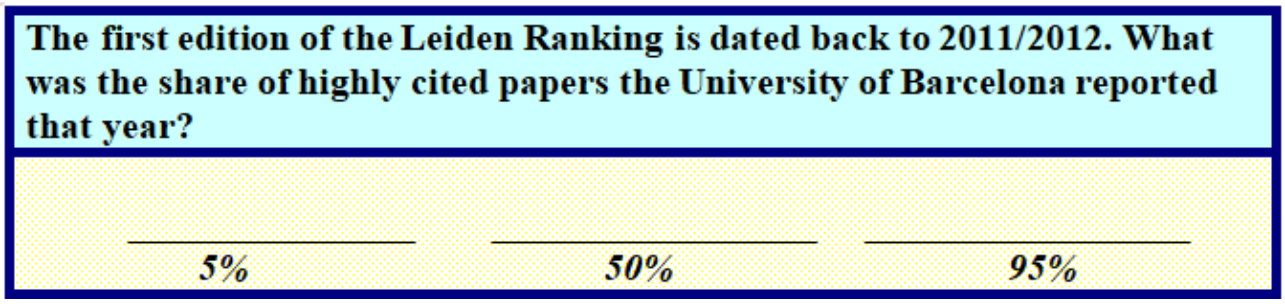
## 6.1 Introduction

Structured expert judgment is an accepted tool in risk analysis for supplementing data shortfalls, quantifying uncertainty and building rational consensus. It has been used in studies sponsored by the European Union, the US NOAA, EPA, Health Canada, CDC among many others, to characterize uncertainty in a wide variety of relationships not amenable to repeated experimentation. To pick a few examples, these include the effects of medical procedures, risks from nuclear power plants, and risks of invasive species.

A panel of experts quantify uncertainty with regard to variables of interest and calibration variables from the subject area. Experts are treated as statistical hypotheses and combined so as to maximize the statistical accuracy and informativeness of the "decision maker". Expert names are preserved to enable competent peer review but are not associated with responses in any open documentation. Expert reasoning is captured during the elicitation and becomes, where indicated, part of the published record. Elicitation is done by specifying percentiles of uncertain quantities, as illustrated below.

## 6.2 Elicitation format

You are presented with an uncertain quantity:




You are asked to quantify your uncertainty by specifying percentiles of your subjective uncertainty:

- The 50%-tile is that number for which you judge the chance $\frac{1}{2}$ that the true value is above or below
- The 5%-tile is that number for which the chance that the true value is BELOW IS 0.05 and the chance that the true value is ABOVE is 0.95.
- The 95% -tile is that number for which the chance that the true value is BELOW, is 0.95, and the chance that the true value is ABOVE is 0.05.

**ALWAYS: 5%-tile $<$ 50%-tile $<$ 95%-tile**
Suppose you respond as shown below:

- **5% percentile:** 8.0
- **50% percentile:** 9.5
- **95% percentile:** 10.0

This means that the true value is equally likely to be above or below 9.5; there is a 90% chance that it lies between 8.0 and 10.0.

A *good probability assessor* is one whose assessments capture the true values with the long run correct relative frequencies (**statistically accurate**), with distributions that are as narrow as possible (**informative**). Informativeness is gauged by 'how far apart the percentiles are' relative to an appropriate background (Shannon relative information).

Measuring statistical accuracy requires the true values for a set of assessments. The true value for the above question is 10.7. It falls above the 95%-tile. If the expert's assessments are *statistically accurate*, then in the long run, 5% of the answers should fall within this inter-percentile interval. Similarly, 90% of the answers should fall between the 5%-tile and the 95

In gauging overall performance, statistical accuracy is more important than informativeness. Non-informative but statistically accurate assessments are useful, as they sensitize us to how large the uncertainties may be;

highly informative but statistically very inaccurate assessments are not useful. Do not shy away from wide distributions if that reflects your real uncertainty.

If you have little knowledge about an item, this fact by itself does NOT disqualify you as an uncertainty assessor. Knowing little means that your percentiles should be 'far apart'. If other experts are more informative, without sacrificing accuracy, then they will exert more influence on the decision maker. But if there are no statistically accurate experts with more informative assessments, then the uninformative assessments accurately depict the uncertainty. That in itself is VERY important information.

## 6.3 Calibration questions

1. What share of the total faculty staff in Spain belongs to the field of the Arts and Humanities in 2017/2018?
2. What was the share of faculty staff with at least a sexenio in Arts and Humanities in 2017/2018?
3. There were 88 universities in Spain between 2015 and 2018, how many had at least one PhD programme in the field of the Arts and Humanities?
4. What was the threshold needed to access to a degree on English Studies at the University of Extremadura (the academic results range from 0 to 16) in 2018?
5. In 2009, what was the share PhDs in fields from the Humanities?
6. In 2009, out of all PhDs in fields from the Humanities, what share where working on a job which was highly related with their PhD?
7. In 2010, the Science fields received 122M euros in fundamental research projects. How much money did projects from the Social Sciences and Humanities received from the National I+D Plans?
8. In the year 2016/2017, how many Bachelor programmes were offered in Spain in the Arts  Humanities fields?
9. How many theses have been defended in Pompeu Fabra from 2010 to 2019 in fields related to the Arts Humanities?
10. How many publications in the Arts  Humanities fields were indexed in 2019 in Scopus (please remember you should not look for the actual data but provide your best estimate)?