Naïve Sampling and Format Dependence in Subjective Probability Calibration

Patrik Hansson (patrik.hansson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Anders Winman (anders.winman@psyk.uu.se)
Department of Psychology, Uppsala University
SE-751 42, Uppsala, Sweden

Abstract

Previous data suggest that probability judgments can often be fairly realistic, but production of probability intervals promotes extreme overconfidence bias (Juslin & Persson 2002; Juslin, Wennerholm, & Olsson, 1999; Klayman, Soll, González-Vallejo, & Barlas, 1999). We present a novel explanation of this format-dependence effect in terms of a *naïve sampling model*. The model assumes that people process distributional information in an unbiased manner, but they are naïve in the sense that they uncritically take sample properties as estimates of population properties. A Monte Carlo simulation demonstrates that the model predicts format dependence, with extreme overconfidence bias for interval production that practically disappears for probability assessments for the same intervals. In the last section this novel prediction is empirically verified by data from human participants.

Introduction

Imagine that you are about to ask your banker for advice about the house installment interest rate next year. In getting the banker to express his or her belief about future interest rates you could either ask him or her to *produce an interval* in which it is likely that the interest is going to fall next year or provide an interval and ask the banker to *assess the subjective probability* that the interest will fall within that interval. These two formats are merely different ways to express a belief (or a subjective probability distribution) about the interest rate next year; in the following referred to as *interval production* and *interval assessment*, respectively. The decision about what question-format to choose appears trivial, but as we shall see – it may be one of the more important personal financial decisions you make.

The realism or calibration of probability judgments varies with the assessment format, so called *format-dependence* (Juslin & Persson 2002; Juslin, Wennerholm & Olsson, 1999; Klayman, Soll, González-Vallejo, & Barlas, 1999). People often make fairly realistic assessments of the probability of events occurring/facts being true, such as "What is the probability that Burma has more than 10 million inhabitants?". Interval production ("Give the smallest interval within which you are 90% certain that the interest will fall next year.") in general produces extreme *overconfidence* bias, see Figure 1. For example, often across 100% intervals, about 40% of the true values fall within the intervals (i.e. rather than the 100% required for realism).

We present a new explanation of this phenomenon with a *naïve sampling model* and test a prediction derived from the model. The "naivety" stands for the assumption that people, in making these intervals, uncritically take sample properties as estimates of population properties (see Fiedler, 2000; Kareev, Arnon, & Horwitz-Zeliger, 2002, for similar ideas). With a Monte Carlo simulation we demonstrate that application of this naïve sampling process indeed produces more overconfidence for production than assessment of intervals. Thereafter, we verify that empirical data discloses the same basic pattern in a task with the corresponding structure.

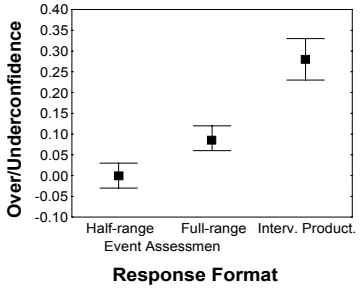

Figure 1: Format dependence with different formats applied to the same subjective probability distributions. With the half-range (two alternatives, forced choice-50-100% scale) format there is close to zero over/underconfidence, with the full-range (no choice, 0-100% scale) format there is marginal overconfidence, and with the interval production format there is extreme overconfidence. Reproduced from Juslin and Persson, (2002).

A Naïve Sampling Model

Consider assessment of a Subjective Probability Distribution (hereafter SPD) for an unknown quantity, for example, a decision maker's belief about the population of Burma. If the decision maker can retrieve the exact population figure from long term memory the SPD will represent precise knowledge; otherwise a plausible range of population fig-

ures will have to be inferred from other facts that are known about Burma. The decision maker's belief is summarized by the SPD in Figure 2A, where the *xx*th fractile of the SPD is the population figure *Y* at which the subjective probability that the true value equals to or is lower than *Y* is *.xx*. Figure 2A could, for example, represent that the decision maker is 25% certain that the population of Burma is equal to or lower than 25 millions (the .25 fractile) and 75% certain that it is equal to or lower than 60 millions (the .75 fractile). Together the .25 and .75 fractiles define a .5 interval around the judge's best guess in which the judge is .5 confident that the population of Burma falls.

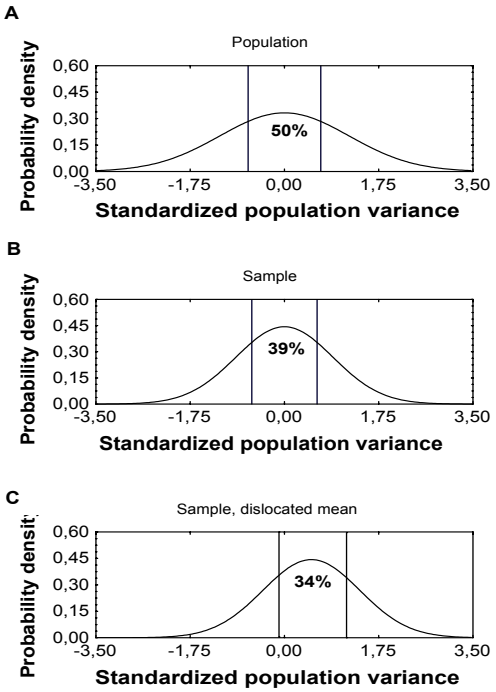

Figure 2: Panel A: Probability density function for the distribution of target values y_i in the reference class R_C cued by *C*. The values on the target dimension *Y* have been standardized to have mean 0 and standard deviation 1. The interval between the .75th and the .25th fractiles of the population distribution include 50% of the population values. Panel B: Probability density function for a sample of 4 exemplars with the sample mean at the same place as the population mean. The values on the target dimension *Y* are expressed in units standardized against population variance. The interval between the .75th and the .25th fractiles of the sample distribution include 39% of the population values. Panel C: Probability density function for a sample of 4 exemplars with the sample mean displaced relative to the population mean. The values on the target dimension *Y* are expressed in units standardized against population variance. The interval between the .75th and the .25th fractiles of the sample distribution on average include 34% of the population values.

With *interval assessment* the decision maker is provided with an interval and is required to assess the probability that the uncertain quantity falls within the interval. For example, the decision maker in Figure 2A may be asked for the probability that Burma's population falls between 25 and 60 millions and if the SPD in Figure 2A is expressed without error the probability judgment is .5.

Interval production requires the decision maker to state the smallest central probability interval which he or she is .xx certain includes the true value. If the decision maker in Figure 2A is asked for a .5 interval the answer should be "between 25 and 60 millions". Formally, the two methods are merely different ways of eliciting the same SPD and should produce the same result. The format dependence in Figure 1 suggests that interval production is particularly prone to overconfidence. Previous research has, however, compared formats that involve *assessment of one tail* of the distribution ("What is the probability that the population of Burma's exceeds 75 millions?") to *production of intervals that involve both tails* of the distribution (Juslin et al., 1999). In this study we directly compare *interval production* to *interval assessment*. Both formats involve a pre-stated interval and both tails of the distribution. To the best of our knowledge these formats have not been compared before. If the explanation of format dependence implied by the naïve sampling model is correct, we predict a format-dependence effect also for interval production and interval assessment, something that is tested in the last section. The algorithm for application of the naïve sampling model to interval production and interval assessment is described next.

Frequentistic algorithm. The *naïve sampling model* suggests that the SPD for the quantitative value *y* of an object *T* is assessed by retrieval of *n* similar exemplars $X_1 \ldots X_n$ from memory, where the sample distribution is directly translated into the required response format:

1. Reference class. One or several facts (cues) about *T* are retrieved from memory, jointly referred to as cue set *C*, defining a reference class R_C of objects in the environment. Keeping with our example, faced with the question about Burma's population the decision maker may retrieve that Burma (*T*) lies in Asia (*C*), which in turn defines a reference class (R_C) of Asian countries.

2. Sample. A sample of *n* exemplars X_i from reference class R_C with known values y_i are retrieved from memory providing a sample distribution. In our example the decision maker may retrieve the populations of *n* Asian countries (other than Burma) which provide a sample of populations for countries similar to Burma (i.e., in this example the similarity refers to the property of being an Asian country).

3. Naïve estimation. The sample distribution of X_i is directly taken as an estimate of the population distribution:

a. Interval assessment: Provided event *E*, the subjective probability *P* is *m*/*n*, where *m* is the number of retrieved exemplars X_i satisfying *E*. In our example, event *E* may be to have a population between 25 and 60 millions. The decision maker may retrieve (say) 4 Asian countries with known populations, 1 of which has a population between 25 and 60 millions, thus reporting probability $P = 1/4 = .25$.

b. Interval production. Provided a probability interval *.xx*, the (1-*.xx*)/2 fractile in the sample defines the lower limit of the interval and the *.xx* + (1-*.xx*)/2 fractile defines the upper limit of the interval, where the two fractiles together define

the *.xx* interval for *y*. When the decision maker in Figure 2A is required to provide his .5 probability interval for the population of Burma he or she may again retrieve 4 Asian countries, 1 (25%) of which has a population equal to or below 25 millions and 3 (75%) of which have a population equal to or below 60 millions. The .25 and the .75 sample fractiles suggest a .5 interval of 25 to 60 millions.

The cognitive processes are essentially the same with both assessment formats: A sample of similar observations are retrieved from memory and directly expressed as required by the format, as a probability (proportion) for interval assessment and as fractiles of the SPD for interval production. There is no bias in the process, only naivety in the sense that sample properties are taken as unbiased estimators of population properties. The implication is that only sources of error that are explicitly represented are considered.

Format-dependence. For infinite sample size the algorithms for interval assessment (Steps 1, 2, 3a) and interval production (Steps 1, 2, 3b) produce identical results: at small sample sizes they do not! Direct (or naïve) use of the sample distribution to assess probability (3a) provides a relatively unbiased[1] estimator of probability (population proportion), but direct use of the sample distribution to produce intervals (3b) generates severely biased and too narrow intervals for small samples. Specifically, the naïve sampling model implies that with interval assessment there are two reasons why the interval may not include the population of *T* (i.e., why *E* is not happening), both of which need to be considered when the probability assessment is made: **a**) The *sampling error* directly manifested in the sample, some countries within R_C satisfy the event *E*, others do not. For sampling with replacement sample proportion *P* is an unbiased estimate of population proportion *p* (the expected value of *P* is *p*). **b**) *The constrained population*: By definition, *T* (e.g., Burma) is never a member of R_C (Asian countries other than Burma: if the population of Burma is known there is no need for an inference!). The fact that R_C always excludes *T* adds a bias to the expected long run proportion that is not explicitly represented in the sample and thus not represented in the probability judgment. Considering error source **a** in the probability judgment but not source **b** produces a minor overconfidence bias with interval assessment.

With an *.xx* probability interval we normatively expect the event *E* to happen with probability *.xx* (e.g., we expect a proportion .9 of the values to fall inside the .9 interval). In regard to interval production there are four contributors to why *E* may not occur, only one of which is explicitly manifested in a sample: **a**) The *sampling error* directly manifested in the sample, some countries within R_C satisfy event *E*, others do not. With correctly estimated fractiles (e.g., large *n*) this occurs with probability 1-*.xx* for a *.xx* probability interval. This is illustrated in Figure 2A, where the proportion of values falling inside the interval is .5 and the error rate is .5. **b**) *The constrained population* (as detailed above). **c**) The *sample variance underestimates the population variance*. The sample interval needs to be corrected by n/(*n*-1) to become an unbiased estimator of the population interval. Even if the sample mean coincides with the population mean a failure to acknowledge this bias adds to the error rate. Figure 2B plots the expected probability density function for sample size 4, with the constraint that the sample and the population means are the same. In this case (assuming normal distributions) the .5 sample interval includes .39 of the true values (i.e., the error rate is .61 rather than .5). **d**) *Dislocated sample distribution*. At small sample size, the interval itself is likely to be dislocated relative to the population distribution. For example, at sample size 4 the .5 sample interval is expected to include 39% of the true values, only if the sample interval has not been dislocated by sampling error relative to the population distribution. The Monte Carlo simulations presented below suggest that if we take this sampling error into account, the sample interval only includes 34% of the true values.

The naïve sampling model implies that because only the first source of error is explicitly represented in the sample, only the first contributor to the error-rate is taken into account in the interval production. At small sample sizes this produces extreme overconfidence. In the next section, we verify our intuitions with a Monte Carlo simulation applied to a data-base identical to the source of the data presented in the empirical section (see Juslin, Winman, & Hansson, 2003, for further discussion of the *naïve sampling model*).

A Monte Carlo Simulation

The target variable estimated by the participants in the experiment reported below is population of a country. The database was defined by the 188 countries listed in the United Nations database (2002). The simulations were performed as follows: A country was sampled at random as the target country *T* with a population *y*. The continent of a country was used as the cue *C*. A sample of *n* exemplars X_i was sampled without replacement from the reference class R_C of exemplars from the same continent as *T* (excluding *T*). For the probability intervals 100, 75 and 50 the fractiles of the sample distributions were used to generate an *.xx* interval for the value of *y*, as detailed by Steps 1, 2 and 3b[2]. The proportion of intervals that included the true value was recorded. To obtain predictions for interval assessment the intervals produced in the first simulation were taken to define the events (the event being that the population of *T* falls within the interval). For each interval and *T* a new and inde-

[1] The qualification "relatively" unbiased refers to error source **b**, *constrained population*, discussed below that applies to the present application and adds a minor bias in the assessments. In the general case and under standard assumptions, of course, sample proportion is an unbiased estimate of population proportion.

[2] With finite sets of data when a distribution of values cannot be directly divided into fractiles there exist different methods for calculating these by approximating a continuous distribution. The present analysis relied on the empirical distribution function with interpolation. This method has the property of discriminating between limits of the intervals for the fractiles and small samples employed in this study. It is calculated as follows; Let *n* be the number of cases, and *p* be the percentile value divided by 100. *(n-1)p* (*(n-1)* times *p*) is expressed as *(n-1)p=j+g* where *j* is the integer part of *(n-1)p*, and *g* is the fractional part of *(n-1)p.* The fractile value is calculated as x_{j+1} if g=0 and as $x_{j+1} + g(x_{j+2} - x_{j+1})$ if g>0.

pendent sample of n exemplars X_i was obtained and the proportion of retrieved X_i falling within the interval was used to compute a probability, as described by Steps 1, 2, and 3a above. The proportion of target countries included by the intervals and the mean probability assigned to the same intervals were calculated for the three different intervals. The averages in Figures 3 A-B are based on 2 million iterations.

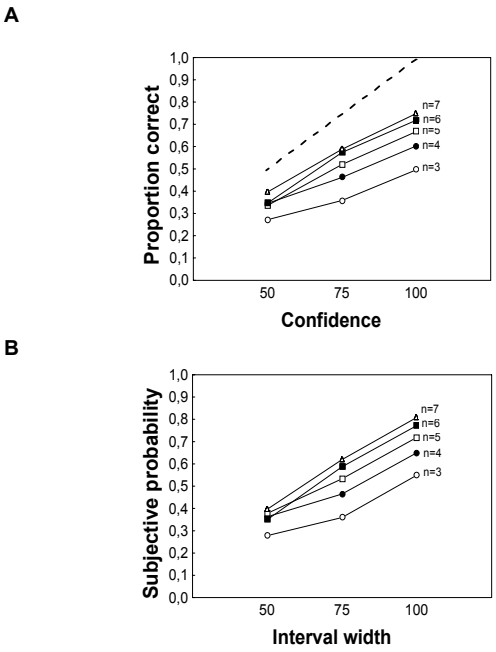

Figure 3: Panel A: Proportion of correct target values included in the intervals for the three different confidence levels with different sample size (n). The dotted line represents perfect calibration. Panel B: The probability ratings for each interval with different sample size (n).

The simulations in Panel A of Figure 3 reproduce the large overconfidence typically found in interval production tasks. For example, with sample size 3 only half of the target values are included by the 100% intervals. There is also as expected an effect of sample size, with smaller sample sizes leading to larger overconfidence. Panel B shows the subjective probability ratings for each interval. As can be seen these ratings coincide quite well with the actual proportions in Panel A. There is a moderate tendency towards overconfidence for the 100% intervals because the target country is prevented from inclusion in the sample of countries used to calculate the proportion included in the interval (this is the reason why we stated that naïve interval assessment is "relatively" unbiased as compared to the severely biased naïve interval productions, see Footnote 1).

Experiment

The experiment transposes the database used in the Monte Carlo simulation into a general knowledge task introduced to human participants. Because interval assessment has not previously been empirically investigated the purpose was to see if the patterns predicted by the naïve sampling model hold empirically. The procedure was basically the same as in the simulation in that one group of participants (the P-group) produced 50, 75 and, 100 probability intervals for randomly selected countries from the database. Another group of participants (A-group) made subjective probability assessments for the intervals produced by the P-group.

Method

Participants

Forty undergraduate students (no specialized skill in geographical or statistical knowledge were required for participation), twenty in each group, (16 female and 24 male) with an average age of 24.8 for the A-group and 25.5 for the P-group, attended. The participants were paid 99 SEK (app. 10 US $) each and the task lasted about 1-1½ hour.

Materials and Apparatus

The experiment was carried out on a PC. A database of 188 world countries listed in the United Nations database (2002) where used as stimuli. For the P-group a computer program picked countries randomly from the database. For the A-group the intervals produced by the P-group were used as stimuli.

Design and Procedure

A randomized between-subject (P vs. A) complemented with a within-subjects control design was used (As explained below, the A-group also performed interval productions after they had made their interval assessments). The interval productions for both groups were made under three different probability levels; .5, .75 and 1.0 and the order of these levels were varied within each group. Half of participants in each group produced intervals in an ascending .5-.75-1.0 (blocked) order; the second half produced them in a descending (blocked) order).

The participants carried out the tasks independently without feedback. Each participant assessed 40 intervals on each probability level resulting in a total of 120 intervals in the interval production task. The participants in the A-group made the same number of judgments. For the P-group the computer generated questions in the following format, for example:

> *Produce the (smallest) interval within which you are 75% certain (probability 0.75) that the population of Burma falls:*
> *Between__X__ and__Y__millions*

For the A-group the program generated the following statement, for example:

> *The population of Burma lies between X and Y millions.*
> *Assess the probability that the statement above is correct?*
> *0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%*

where X and Y defined an interval previously produced by a participant in the P-group. The program accessed the database in the following way: Each participant in the P-group received an independent different random sample of countries. The A-group assessed probabilities on the already

produced intervals of these samples so that Participant 1 in the A-group made probability assessments on intervals produced by Participant 1 in the P-group, and so on. All of the 120 intervals were presented in random order for the A-group. For the P(within)-group the program accessed the database in the same manner as for the P-group. In this way each of the participants' in the A-group and P(within)-group assessed and produced intervals for the same countries.

Results

Figure 4 presents the proportion correct (Panel A) and the interval-range (Panel B) for interval production[3]. There is clear difference between the probability intervals, both in the proportions falling within the interval and the interval ranges. As predicted by the naïve sampling model, the intervals are sensitive to distributional information and increase for higher probabilities.

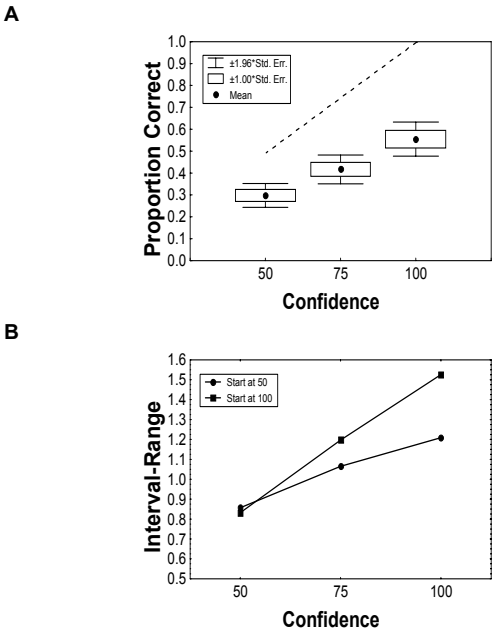

Figure 4: Panel A: Mean proportion of correct population figures included in the produced intervals for the three probability levels (with 95% confidence intervals, *N*=40). The dotted line is the proportions required for perfect calibration. Panel B: The interval-range plotted for each probability for the groups creating the intervals in ascending vs. descending order (*N*=10).

The *over/underconfidence* score *ou* is calculated as $r - c = \pm ou$ for interval productions where *r* is the probability level (i.e. 50, 75, & 100) and c is the proportion of true population figures falling in the interval. For interval assessment the over/underconfidence is computed $\bar{r} - \bar{c} = \pm ou$, where $\bar{r}$ is the mean probability assessment and $\bar{c}$ the proportion of true population figures falling within the interval. In both cases overconfidence bias is represented by a positive score, underconfidence bias by negative score and good realism or calibration is represented by a zero score.

Figure 5A summarizes how the response formats influence the participants' over/underconfidence in their judgments. For comparison with data, the over/underconfidence biases predicted from the Monte Carlo simulation of the naïve sampling model in are presented in Figure 5B.

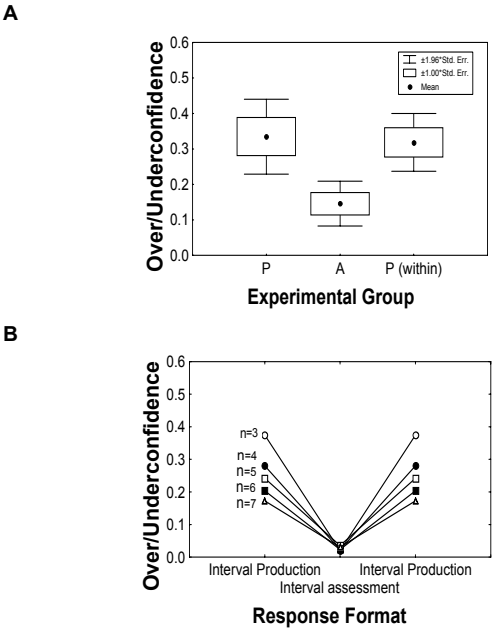

Figure 5: Panel A: Mean over/underconfidence bias for the two different assessment methods (with 95% confidence intervals, *N*=20). P=interval production, A=Interval assessment, and P(within) = Interval productions by the A-group after the interval assessments. Panel B: over/underconfidence scores for the different assessment methods produced by the naïve sampling model at different sample size *n*.

As predicted, there is significant difference between interval production and interval assessment in that the interval production leads to more overconfidence (P vs. A, Figure 5A, $F(1,38)$; 8.98, MSE=0.03, p=0.004). In order to verify that this effect is not confined to a between-subjects comparison, the A-group performed their own interval productions after they had completed their interval assessments (P-(within) in Figure 5A). The same pattern is verified also in the within-subjects comparison: when the A-group produced intervals; they, so to speak, fell back and became significantly more overconfident when they made their interval productions as compared to the interval assessments (A vs. P(within)), Figure 5A, $F(1,38)$; 10.73; MSE=0.02, p=0.002). The difference between the two interval production groups (P vs. P(within) in Figure 5A) is not significant ($F(1,38)$=.06, MSE=.046; p=.812). The main conclusion is that the format-dependence effect can be demonstrated both in a within- and a between-subjects comparison, but also in a within-items comparison. The same pattern is seen in Fig-

[3] The interval-range was calculated proportionally to population size by dividing the difference between the upper and lower limit with the midpoint of the intervals.

ure 5B: With interval productions the naïve sampling model produces high overconfidence bias but much more accurate calibration with the interval assessment format.

Discussion

In this paper we provide a tentative explanation of format dependence in subjective probability calibration with a naïve sampling model: Naïve use of sample properties provide a relatively unbiased estimate of probability, but a severely biased estimate of intervals. Moreover, as predicted if people retrieve sample distributions, the interval-ranges show that the participants appropriately make wider or narrower intervals for the different probability levels, suggesting that they do retrieve distributional information (Figure 4B).

As for the format dependence we find, as in earlier studies (Juslin et al., 1999; Juslin et al., 2002; Klayman et al., 1999), that the response format has profound effects on the conclusions. Interval production suggests extreme overconfidence bias, while subjective probability assessment does not. The present study shows that this holds also with interval assessment. This phenomenon is reproduced when the naïve sampling model is implemented as a Monte Carlo simulation (see Figure 3A and Figure 5B). It is obvious that the empirical data discloses the same basic pattern as the Mote Carlo simulation of the naïve sampling model. However, the model predicts close to zero average overconfidence bias for interval assessment but the data from the experiment show approximately 0.1 in the average overconfidence bias for interval assessment. This phenomena could be explained if we assume that there exist a correlation between the samples retrieved by the participants' in the two different groups (i.e. if the different participants in the two groups retrieved exactly the same samples, that is, a correlation of 1 between retrieved samples there would be no format dependence). The simulations of the model in this paper, however, only considers zero correlation between retrieved samples for the different formats (for a more detailed discussion and simulations that includes correlations between samples, see Juslin, Winman & Hansson, 2003).

The sample sizes in turn could be taken to reflect people's limited knowledge in a domain, for example, of the populations in different world countries. One could, of course, argue that the experiment and the simulation presented in this paper only tackles one knowledge domain (i.e. geographical knowledge). We are, however, convinced that the rational behind the model can be applied to all knowledge domains containing continues quantities, for example, estimation of interest-rate, blood-pressure, stock-values, distance, speed etc. The basic ideas are further easily conjoined with exemplar architectures to predict probability judgments that also take similarity into account (Juslin & Persson, 2002).

According to the explanation proposed by the naïve sampling model the overconfidence phenomenon with interval production is not caused by biased cognitive processing in any more straightforward or trivial sense (Kahneman, Slovic & Tversky, 1982). There is no explicit information processing bias, only limited knowledge. In other words, reliance on small samples which are not corrected relative to populations generates the overconfidence in interval production. This view is also inherently more consistent with the observations that experts (i.e., presumably with larger samples) are often better calibrated (see Kahneman et al., 1982).

If we return to the opening example about how to query your banker by now the answer should be straightforward: you should provide a pre-stated interval and let the banker assess the probability that the interest falls within this interval. In this way you could steer clear of an extremely overconfident advice and avoid unexpectedly high house-installments next year…

Acknowledgments

This research is supported by the Swedish Research Council.

References

Fiedler, K. (2000). Beware of Samples! A Cognitive-Ecological Sampling Approach to Judgment Biases. *Psychological Review, 107*, 659-676.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format Dependency in Subjective probability calibration. *Journal of Experimental Psychology: Learning, memory and Cognition, 28*, 1038-1052.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563-607.

Juslin, P., Winman, A., & Hansson, P. (2003). *The Naïve Intuitive Statistician: A Sampling Model of Format Dependence in Probability Judgment.* Manuscript in preparation.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgments under uncertainty: Heuristics and Biases.* New York: Cambridge University Press.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the Misperception of Variability. *Journal of Experimental Psychology: General*, *131*, 287-297.

Klayman, J., Soll, J. B., González-Vallejo, C. & Barlas, S. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*, *79,* 216-247.