An Experimental Study of the Effectiveness of Three Debiasing Techniques*

B. Kemal Büyükkurt
Marketing Department, Faculty of Commerce and Administration, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec H3G 1M8, Canada
Meral Demirbag Büyükkurt
Decision Sciences and MIS Department, Faculty of Commerce and Administration, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec H3G 1M8, Canada

ABSTRACT

Subjective probability distributions constitute an important part of the input to decision analysis and other decision aids. The long list of persistent biases associated with human judgments under uncertainty [16] suggests, however, that these biases can be translated into the elicited probabilities which, in turn, may be reflected in the output of the decision aids, potentially leading to biased decisions.

This experiment studies the effectiveness of three debiasing techniques in elicitation of subjective probability distributions. It is hypothesized that the Socratic procedure [18] and the devil's advocate approach [6] [7] [31] [32] [33] [34] will increase subjective uncertainty and thus help assessors overcome a persistent bias called "overconfidence." Mental encoding of the frequency of the observed instances into prespecified intervals, however, is expected to decrease subjective uncertainty and to help assessors better capture, mentally, the location and skewness of the observed distribution. The assessors' ratings of uncertainty confirm these hypotheses related to subjective uncertainty but three other measures based on the dispersion of the elicited subjective probability distributions do not. Possible explanations are discussed. An intriguing explanation is that debiasing may affect what some have called "second order" uncertainty. While uncertainty ratings may include this second component, the measures based on the elicited distributions relate only to "first order" uncertainty.

Subject Area: Decision Theory.

INTRODUCTION

As representations of expert opinion under uncertainty, subjective probability distributions regarding the current and/or future values of some uncertain variables constitute an important part of the input to decision aids such as decision analysis, expert systems, and decision support systems. Subjective probabilities also play a key role in areas such as judgmental forecasting and auditing. The courses of action to be recommended by such decision aids, and their predictions and/or diagnoses, depend heavily on the proper assessment of the relevant subjective probabilities. Any substantial error or bias in the elicited probabilities may be reflected in the output of these decision aids, potentially leading to biased decisions.

A stream of research on human judgments under uncertainty during the last two decades has compiled a long list of persistent biases associated with such judgments [16], which, in turn, suggested biases in the elicited subjective probability distributions. Recognizing the implications of these findings for the users of the decision aids mentioned above, researchers focused their attention on procedures which would help assessors to overcome their own judgmental biases [9] [18] [25]. The basic assumption of these attempts was that the causes of the judgmental biases could be identified and then appropriate corrective procedures could be formulated. As presented in the comprehensive review of the literature on overconfidence and hindsight biases [9], various ingenious methods of debiasing have

*This research was partially funded by an operating grant (A6743) to the second author from the Natural Science and Engineering Research Council of Canada. The authors are listed alphabetically and contributed equally to this study.

been less than successful. An encouraging conclusion of this review was that mechanical attempts of debiasing without prior psychological theory are not likely to be productive whereas those corrective procedures which change the psychological nature of the task and the assessors' mental approach to it will probably be effective.

The objective of this study is to test the effectiveness of three such debiasing techniques, namely, the Socratic [18] and devil's advocate [6] [7] [31] [32] [33] [34] approaches, and a mental encoding aid in reducing the effects of three types of persistent biases. The first bias to be studied has been called overconfidence [1] [21] and is typically indicated by subjective probability distributions which are too tight, given some normative criteria. The second bias (which also leads to tight distributions) is the assessors' insensitivity to nonrandomness and the small size of self-selected samples [17]. The final bias is the insensitivity to the skewed nature of observed distributions and the tendency to make subjective probabilities conform to a normal model [41]. Consequently, the shape of the observed nonsymmetric distributions may not be adequately captured by those assessors who are prone to this bias.

LITERATURE REVIEW AND HYPOTHESES

The findings in the literature on human judgments under uncertainty indicate that such judgments are prone to some systematic biases mainly because of the mental heuristics that humans use [12] [13] [16] [27]. Some of the findings were criticized from a methodological point of view, arguing that subjects in these studies were misled by simplistic and unrealistic experimental circumstances of little significance beyond the laboratory [4] [22]. It is also possible to argue that subjects were not good probability assessors because they lacked substantive expertise (knowledge regarding the subject matter of interest), and/or normative expertise (ability to express expectations in probabilistic form) [24] [44]. The reported biases, however, seem to persist even when judgmental tasks are meaningful and important to the subjects in terms of potential outcomes of the decisions and when the subjects are well trained in statistics [17].

Using the Devil's Advocate Approach to Debias Overconfidence

Overconfidence regarding general-knowledge items of extreme and moderate difficulty has been a persistent systematic bias in numerous studies [1] [21] where the assessed subjective probability distributions were consistently tighter than what the assessors' level of uncertainty would imply. That is, dispersion of the assessed distributions underrepresented subjects' uncertainty. Overconfidence was most prevalent with tasks of greater difficulty.

As presented in detail in the comprehensive review by Lichtenstein, Fischhoff and Phillips [21], debiasing efforts to reduce overconfidence in probabilistic judgments included explicit warnings, specific instructions to spread out the tails of the assessed subjective probability distributions, increased motivation, training, and restructuring of the cognitive task. In general, warnings and instructions could not eliminate overconfidence. The effects of training, however, were mixed. Restructuring the cognitive task by asking the subjects to think about and then write down the reasons that could contradict their previously stated judgment reduced overconfidence substantially [19]. This devil's advocate approach, in which subjects are encouraged to search for and consider information discrepant with their current plans and beliefs, was also effective in numerous laboratory experiments on prediction performance [6] [7] [31] [32] [33] [34].

The devil's advocate approach is likely to be effective especially when the assessor treats the subjective probability distribution of the random quantity of interest as a predictive (conditional) distribution, as in the regression context, and attempts to assess subjective probability as a function of the perceived covariates of the random quantity. This means that the random quantity is treated as a dependent variable whose probability is induced from a mental model of cue-criterion relationships. Nisbett, Krantz, Jepson, and Fong [25] argued that mental models as such play a crucial role in inductive reasoning. Similary, Kahneman and Tversky's [17] discussion of variants of uncertainty implied that the subjective probability associated with a random quantity can be inferred from a causal system by taking an inside view of the system, or reasoned by referring to related knowledge. For such uncertain events where a mental model is operating, the devil's advocate approach can focus on the perceived cue-criterion relationships and the decision analyst can ask the assessor to generate instances where such relationships do not hold. The higher the perceived correlation between each of the cues and the criterion variable, the higher should be the confidence placed on the final prediction [13]. Consequently, factors that increase the assessor's uncertainty about the strength of the correlations between the predictors and the criterion should increase the uncertainty associated with the predicted criterion. Thus, the first of a series of four hypotheses to be addressed in the present study can be stated as follows:

> H1: If the subjective probability of a random quantity is induced from a mental model of cue-criterion relationships, then the devil's advocate approach will increase subjective uncertainty associated with the criterion by increasing the uncertainty associated with the cue-criterion relationships.

Using the Socratic Approach to Debias Insensitivity to Nonrandomness and Small Sample Size

Another persistent human bias in judgments under uncertainty concerns the lack of appreciation of the impact of nonrandom and small samples on subjective estimates. This bias seems to be due to the use of what has been called the availability heuristic [17]. Those assessors who resort to this heuristic judge the probability of an event by the ease with which instances can be recalled. Thus, subjective probability of instances which are readily available in long term memory because of factors such as vividness and recency are usually overstated. A simple debiasing approach in this case is to help the assessors think about the previously observed values of the random quantity of interest as an instance of sampling [17] [26]. Then the assessors can be reminded that small and nonrandom samples, especially those which are self selected without a statistically sound sampling plan, are likely to be unrepresentative of the population of interest, and therefore lead to biased estimates. Rather than stating this directly to the subjects, the researcher can use the Socratic procedure [17] to guide the respondent through subtle questions to the desired conclusion. Such a debiasing technique should decrease overconfidence which may be due to insensitivity to sample size and nonrandomness. Therefore, the second hypothesis is stated as follows:

> H2: For those tasks where the subjective probability of a random quantity may be affected by the "availability" heuristic, the Socratic procedure will sensitize the assessor to nonrandomness and potentially small size of the self selected sample, and thereby increase subjective uncertainty.

Aiding Mental Encoding of Instances to Debias the Tendency to Force a Normal Model

The third bias examined in this study deals with the apparent attempt of assessors to force the structure of a normal distribution on their subjective probability distributions, possibly because the normal distribution is overly stressed in statistical training [41]. It may be possible to reduce this bias by discussing with the assessor the significance of basing probability assessments on perceived uncertainty rather than on an assumed model [41]. Another corrective procedure could involve encoding of the observed instances in memory. For example, the assessor could be instructed to break the potential range of the random quantity of interest into prespecified small intervals, and then keep a mental record of the frequency of occurrences in each. If the intervals are properly specified, then it should be easier to mentally capture the general shape and location of the observed distribution using this mental aid. If this procedure enhances proper mental representation of the random quantity and its recall, then subjective uncertainty associated with the random quantity should decrease for those who use the aid. Thus, two hypotheses can be stated:

H3: Those who mentally encode the frequency of observed instances in terms of small, prespecified intervals will be less uncertain about the random quantity than those who do not use such an aid.

H4: Those who mentally encode the frequency of observed instances in terms of small, prespecified intervals will be able to better capture the general shape and location of distribution of the random quantity of interest than those who do not use such an aid.

METHODOLOGY

Experimental Design, Subjects, and Equipment

To test the above hypotheses, a laboratory experiment was conducted in which each assessor's subjective probability distributions regarding several random quantities were elicited in an interactive session by using a personal computer and an electronic mouse. As discussed in detail later, the subjects expressed their judgments of uncertainty by moving the cursor on the screen to designated areas and then simply pressing the mouse buttons. The factors of the resulting 2x3x6 within subjects experimental design were debiasing (debiasing versus no debiasing), nature of task (three different assessment tasks), and method of assessment (six methods, as described below). The second factor was replicated within subjects.

The method of assessment was included in the design as another factor to increase external validity of the results through "deliberate sampling for heterogeneity" [5, p. 75]. Since previous studies clearly demonstrated that assessed distributions are sensitive to the method of elicitation [36] [38] [40] [41] [42], generalizability of the effectiveness of the debiasing efforts should increase if the effects are observed across different elicitation methods.

Subjective probability distributions were elicited interactively by using either a cumulative probability or a relative likelihood method. Cumulative probability assessments included one of three methods: (1) a numerical method, that is, a numerical expression of uncertainty in terms of a number between 0 and 100, (2) a pie chart, a graphical version of the probability wheel [36], and (3) a graphic rating scale, that is, a horizontal bar with poles labeled as "impossible" on the

left and "certain" on the right. For those who used the pie chart, the relative size of a subject-adjusted pie in a circle represented subjective probability. Those who used the graphic rating scale simply selected a point along the bar to indicate their degree of subjective uncertainty. Relative distance of the selected point from the poles of the bar indicated subjective probability.

Relative likelihood assessments (see, for example, [2]) involved breaking the subject supplied range for the random quantity of interest into small intervals, and then presenting such intervals pairwise to the assessors and asking them to indicate the relative likelihood associated with each. Just like the cumulative probability methods discussed above, three versions of the method were used: numerical method, pie chart, and bar chart. In the numerical method, the subjects adjusted a number displayed just below the pair of intervals by using the buttons of the electronic "mouse," increasing or decreasing an initial value of "1" by pressing the left or right button, respectively. The final value selected by the subject indicated the relative likelihood of the event displayed on the left in comparison to the event on the right. The remaining methods also involved expression of relative likelihood by using the mouse buttons. In the pie-chart method, a full circle was initially divided into two equal parts. The subjects adjusted the size of the two parts such that the sizes corresponded to subjective probabilities associated with the displayed pair of intervals. In the bar chart, the relative likelihood of the pair of intervals was expressed by adjusting the relative heights of two bars. Thus, a total of six elicitation methods were used in the experiment.

The subjects of the experiment were senior or M.B.A. level university students majoring in business administration who had taken at least one course in probability and statistics. Seniors participated in this study as part of a decision analysis course during the last week of the semester in which the course was taken. M.B.A. students were familiar with the introductory concepts of decision analysis such as conditional payoff tables and decision making under uncertainty using prior information expressed in terms of subjective probabilities. Twenty-seven of the final effective sample size of 50 were MBA students.

The use of students as experimental subjects is a controversial issue in research involving the elicitation of subjective probability distributions in particular, and in the social sciences in general. The results of the experimental studies involving the elicitation of subjective probabilities will be meaningful only if the assessors have normative and substantive expertise [24] [44]. That is, the assessors should be knowledgeable about the subject matter of interest, and be able to express subjective uncertainty in terms of probabilities. Since the experimental tasks (discussed below) were relevant to and within the experience domain of the subjects, and since they were already knowledgeable about the elicitation of subjective probabilities and their use in decision analysis, the two requirements (to provide meaningful and coherent subjective probabilities) were not drastically violated in this experiment. To improve the quality of the data in this regard and reduce the potentially negative effects of lack of experience in probability elicitation, a warm-up elicitation task was included in the experiment, as discussed later. Furthermore, relatively "bad" assessors were excluded from the study by eliminating six subjects whose responses implied incoherent subjective probabilities (for example, in the cumulative probability method, judging the probability associated with a high value of random quantity of interest to be lower than the probability associated with a low value).

The use of student subjects in experiments may also jeopardize the external validity (generalizability) of experimental results [23]. Specifically, if the purpose

of the research is to generalize the results to the performance of a real world task (for example, assessing the responsiveness of an advertising function within a budget range), then the the external validity becomes questionable when the student subjects do not have the necessary characteristics and background (market experience) to carry out the task. This is not a problem in the current study because such a generalization is not sought. The relevance of the experimental task to the student subjects as a generalizability issue has also been discussed by other researchers [3] [15] [28] [30] [33].

The extent to which the subjects are involved in the experimental task [33] and how hard they work at it [9] [40] are also important in determining the validity of the experimental results. Although the subjects were apparently enthusiastic about the experimental task, measures of subject involvement were not taken in this study.

Experimental Procedure

Each elicitation session started with a brief description of the nature of the research project and instructions regarding the use of the mouse. The research project was introduced as "quantification of uncertainty," and the subjects were asked to recall the use of subjective probabilities in decision analysis in courses they had recently taken or were currently taking. Then, the subjective nature of probability judgments was explained. It was stressed that the purpose of the study was not to test the subjects' knowledge. The subjects were reminded that there were no right or wrong probabilities associated with the random variables to be presented and were encouraged to freely express their own feelings of uncertainty while at the same time being as consistent as possible. Next, each subject went through a computerized training session which was designed according to the guidelines mentioned in previous research (see e.g., [41], [42]). Following an introduction to the elicitation method they would use throughout the session, the subjects went through a warm-up elicitation task in which their subjective probability distributions regarding "the number of days per year it snows in Montreal" was assessed.

After the warm-up task, each subject was randomly assigned to one of the cells of the experiment by the computer program and completed three elicitation tasks. In the first task, the random quantity of interest was the number of instructors without Ph.D. degrees employed by the university where the subject was enrolled. This is a task where the subjective probability judgments are likely to be affected by the instances the subjects can recall and therefore possibly subject to the availability heuristic [17]. If the subject bases the probability judgments only on the self-observed sample without attempting to account for the fact that his/her sample of instructors may be biased, then the final probability judgment may be biased even if the recall of self-observed sample is perfect. This is true because the subject's limited and nonrandom sample may come from a few faculties within the university where the proportion of instructors who hold Ph.D. degrees may be different than the proportion in the university as a whole.

The second task involved an almanac question: What is the population of Peru? [1]. A prior pilot study with 60 subjects whose background was similar to the subjects of the experiment suggested that those who were not familiar with this random quantity of interest treated the population of Peru as a dependent variable and attempted to predict its value from some perceived correlates such as geographic size of the country, education level, gross national product, birth rate, religion, and climate. In line with the arguments in the previous section, they

seemed to have a mental model (schema) of determinants of population size and relied on this model to predict Peru's population when they could not judge the population size directly.

The final task requested the subjects to assume the role of a bank teller and observe 30 interarrival times from a highly skewed distribution, simulating the arrivals of users to the service window. Then, the subjective probability distribution corresponding to the observed distribution was elicited.

For each task, the subjects supplied the shortest possible range within which they would be certain to observe the random quantity of interest. If one of the cumulative probability assessment methods was used, then various values of the random quantity within the subject specified range were displayed sequentially and the subjects expressed the probability that the value of the random quantity would be less than or equal to each displayed value. A subset of the displayed values corresponded to 5, 25, 50, 75, and 95 percent of the upper value of the range. The remaining displayed values were obtained by adding one percent to or subtracting one percent from each value in the mentioned subset. Thus, three subsets of five values each (a total of 15 values) were displayed in each task. These can be regarded as three approximate replications of the same five probability judgments. The order of presentation was randomized within each subset of five values of the random quantity. If one of the relative likelihood methods was used, the subject supplied range was first divided into five nonoverlapping intervals. Next, all possible pairs of intervals (10 in all) were presented sequentially on the screen and the subjects expressed their relative likelihoods as described above. The order of presentation of pairs of intervals and left-right position on the screen of each interval for a given pair was randomized.

Debiasing

About half of the subjects were exposed to debiasing efforts. In the first task, where the uncertain quantity of interest was the number of instructors without Ph.D. degrees, the Socratic procedure [17] was used. First, a series of general statements was displayed on the screen to caution the subjects about the potential biases in statistical estimation due to nonrandom and small samples. Then, the subjects simply answered "yes" or "no" to questions regarding the effects of small sample size and nonrandom samples on sampling error. If a normatively wrong answer was given, then the right answer was displayed with the underlying rationale. Next, the total number of professors at the university and a list of the various schools (faculties) of the university were displayed. A series of questions were asked, encouraging the subjects to consider whether their sample of professors could be regarded as random and sufficiently large, given the total number of professors at the university and given that they have taken courses only from a few of the schools. Some of the questions were: How many courses have you taken so far? How many different professors have you met so far? How many of your professors had Ph.D. degrees? From what schools of the university have you taken courses?

The debiasing efforts for the second task of estimating the population of Peru took a devil's advocate approach and focused on cue-criterion (independent variable-dependent variable) relationships. First, a list of independent variables, which were judged by 60 university students to be relatively reliable predictors of population size, was displayed. Next, the subjects were cautioned that the relationship between each of these predictors and population size may not be as strong as some

might believe. To demonstrate the point, 10 pairs of countries were presented with data on population size and each of the predictors, such that the pairs constituted counter examples of the presumed correlations. For each pair, subjects were asked to examine a two by two table and then draw a conclusion from it. Their verbal conclusions were confirmed by the experimenter who was present throughout the session. Although the experimenter was instructed to correct the wrong conclusions, her assistance was not needed. Note that this debiasing approach is different from the approach used by Koriat, Lichtenstein and Fischhoff [19]. They asked their subjects to list the reasons why their responses might be wrong. Apparently, this encouraged the subjects to address their memory differently from what is customary in confidence assessment tasks. In the present study, the subjects were not encouraged to engage in such an active memory search. However, their cognitive role involved active interpretation of the displayed counter examples.

In the final task, the corrective procedures emphasized encoding of the observed interarrival times in memory such that the skewed nature and the general location of the observed distribution could be captured and the subjects' attempt to mentally mold their subjective probability distributions according to a normal model could be avoided. Dividing the potential range of the interarrival times into five equal intervals, the subjects were told to count and keep in memory the number of instances in each interval. Next, they were given the opportunity to manually calculate the relative percentage of cases in each interval by recalling the absolute frequency of instances in each interval.

Upon completion of the elicitation sessions, the subjects were asked to fill out a questionnaire where they rated their uncertainty regarding each of the random variables of the study on a seven-point semantic differential scale. The two poles of the scale were labeled "very certain" and "very uncertain." Finally, the subjects provided feedback regarding the computerized elicitation session and potential improvement areas in the method of elicitation used.

ANALYSIS

The first step in the analysis of the interactively generated data was to obtain a subjective probability density for each task for each subject. If one of the cumulative probability assessment methods was used in data collection, then probability judgments regarding the three approximate replications of each value of the random quantity were averaged. Such averages for all values within the subject supplied range were then used to construct the associated cumulative probability distribution, which was later transformed into a probability density (pd). If a relative likelihood method was utilized in expressing uncertainty judgments, each of the five intervals in the subject supplied range was regarded as a reference interval. For each interval, a pd was obtained (see [2]) by including only those relative likelihood judgments involving the reference interval and the remaining intervals. Averages of the probabilities across corresponding intervals of all five pd's provided the subjective pd for that task. Note that if a psychometric point of view is taken [38], then it can be argued that such an average of multiple measures of uncertainty will provide a more reliable measure than a single measure.

Four dependent variables were defined to test the hypothesized effects of the three debiasing methods on subjective uncertainty. While subjects' uncertainty ratings constituted one of the dependent variables, the remaining three were based on the elicited subjective pd's. Variance of each subjective pd was considered a good candidate in this context, since variance has traditionally been used in the

Bayesian literature as an indicator of relative uncertainty [2] [43]. Another possible measure is the coefficient of variation since it reflects psychologists' findings that human judgment processes focus on relative rather than absolute deviation [20]. The fourth dependent variable was entropy of each subjective pd. Letting the random quantity of interest be denoted as θ, and π be a pd on θ, then the entropy of π for discrete values θ_i of θ is defined by the following equation:

$$E(\pi) = -\Sigma \pi(\theta_i) \cdot \log \pi(\theta_i).$$

If $\pi(\theta_i) = 0$, then, $(\pi(\theta_i) \cdot \log \pi(\theta_i))$ is defined to be zero [2]. Entropy, as an information theoretic measure, indicates the amount of uncertainty inherent in the pd [29]. The minimum value of entropy is zero, and it corresponds to the case of certainty where the pd collapses on a specific value θ_k such that $\pi(\theta_k) = 1$ and $\pi(\theta_i) = 0$ for $i \neq k$. Maximum entropy is equal to $\log(n)$ for $\pi(\theta_i) = 1/n$. Note that this is the discrete counterpart of the uniform distribution which reflects "the most uncertain" form of subjective pd [2].

An exploratory data analysis stage preceded the multivariate analysis of variance (MANOVA) of the four dependent variables. In order to meet the normality assumption of MANOVA, a log transformation of variance, and a square transformation of the entropy measure were used. Uncertainty ratings and coefficients of variation were not transformed. Also, as presented in Tables 1 through 3, Bartlett's sphericity test indicated that the correlations between the dependent variables warranted MANOVA rather than univariate ANOVA. Finally, Box's *M* test for each of the three tasks using debiasing and method of elicitation as two experimental factors suggested that the assumption of equal variance and covariance matrices was violated. An examination of the cell variances revealed that much of the variation was due to the method of elicitation, which is not surprising given the vast literature about the effects of method of elicitation on subjective probability distributions. Therefore, the method of elicitation was collapsed as an experimental factor. As presented in Tables 1 through 3, Box's *M* statistic for each task implies that the null hypothesis of equal covariance matrices is not violated when debiasing is the only experimental factor. Therefore, further data analysis was carried out with debiasing as the only independent variable, collapsing across method of elicitation.

Multivariate and univariate *F*-tests regarding the effects of debiasing on the dependent variables of the study are summarized in Tables 1 through 3, which relate to hypotheses H1, H2 and H3, respectively. As far as the effectiveness of the Socratic procedure is concerned, the multivariate *F*-test based on Wilk's λ in Table 1 indicates that there are statistically significant differences between "debiasing" versus "no debiasing" groups ($F(4,45) = 3.235$, $p = .05$). This difference is due to subjects' ratings of uncertainty (univariate $F(1,48) = 13.429$, $p = .001$). As hypothesized, the Socratic procedure increased perceived uncertainty, but this was not captured by any of the three measures based on the assessed subjective pds. This is indicated by the corresponding nonsignificant univariate *F*-statistics in Table 1.

As presented in Table 2, although the multivariate F-test associated with the effects of the devil's advocate approach was not significant ($F(4,44) = 1.494$, $p = .22$), univariate *F*-tests on each of the four dependent variables replicated the results regarding the effects of the Socratic procedure. The devil's advocate approach increased subjects' ratings of uncertainty ($F(1,47) = 5.658$, $p = .002$), but

Table 1: MANOVA results regarding the effectiveness of the Socratic procedure.

Multivariate Tests

Test	Test Statistic	*F*	*df*	*P*-value
Bartlett's test of sphericity	45.579	—	6	.00
Box's *M*	9.412	.855	10,10357	.58
Wilks' λ	.777	3.235	4,45	.02

Univariate *F*-Tests

	Mean Square			
Dependent Variable	Debiasing	Error	F^a	*P*-Value
Uncertainty rating	4.267	.318	13.429	.001
Log of variance	.313	.653	.479	.492
Coefficient of variation	.009	.027	.347	.559
Squared entropy	.004	.011	.339	.563

[a]Degrees of freedom for univariate *F*-tests = 1,48.

Table 2: MANOVA results regarding the effectiveness of the devil's advocate appproach.

Multivariate Tests

Test	Test Statistic	F^a	*df*	*P*-value
Bartlett's test of sphericity	14.507	—	6	.02
Box's M	16.038	1.453	10,9563	.15
Wilks' λ	.880	1.494	4,44	.22

Univariate *F*-Tests

	Mean Square			
Dependent Variable	Debiasing	Error	F^a	*P*-Value
Uncertainty rating	2.989	.529	5.658	.02
Log of variance	1.075	.758	1.419	.24
Coefficient of variation	.000	.027	.000	.99
Squared entropy	.007	.009	.735	.40

[a]Degrees of freedom for univariate *F*-tests = 1,47.

did not affect the remaining measures based on the elicited pds. Similarly, mental encoding of the observed instances into prespecified intervals reduced perceived uncertainty ($F(1,46) = 17.635$, $p = .00$) as presented in Table 3. However, the effects of debiasing were not reflected in the remaining measures of uncertainty based on the elicited distributions.

As for the last hypothesis of the study, the results suggest that mental encoding of the observed instances as instructed in this experiment enables the individual to better capture the location of the distribution of interest, but not its skewness. The average absolute deviation of the means of the elicited distributions from the "true" mean of the observed distribution was smaller for the "debiasing" than

Table 3: MANOVA results regarding the effectiveness of mental encoding aid.

Multivariate Tests

Test	Test Statistic	*F*	*df*	*P*-value
Bartlett's test of sphericity	43.197	—	6	.00
Box's M	14.247	1.275	10,9461	.18
Wilks' λ	.701	4.579	4,43	.04

Univariate *F*-Tests

	Mean Square			
Dependent Variable	Debiasing	Error	F^a	*P*-Value
Uncertainty rating	9.399	.533	17.635	.00
Log of variance	.016	.043	.369	.55
Coefficient of variation	.017	.007	2.428	.13
Squared entropy	.007	.006	1.171	.29

[a]Degrees of freedom for univariate *F*-tests = 1,46.

"no debiasing" condition (approximate $t(37) = 1.422$, $p = .09$).[1] However, no significant differences were found in these means when the average absolute deviation of the skewness of the elicited distributions from the "true" skewness was the dependent variable ($t(48) = 1.04$, $p = .15$).

CONCLUSIONS AND DISCUSSION

This experimental study investigated the effects of the Socratic procedure, the devil's advocate approach, and a mental encoding aid in the context of interactive elicitation of subjective probability distributions. It was hypothesized that the first two debiasing techniques would increase and the last technique would decrease subjective uncertainty. While these expectations were confirmed by subjects' ratings of perceived uncertainty, no significant differences were found between debiasing and no debiasing conditions when three other measures based on the dispersion of the elicited pds were used as the dependent variables.

A possible explanation for the result that the measures based on subjective pds did not parallel the pattern observed in the ratings of uncertainty is that the pds were not accurately assessed, and therefore, poorly represented subjective uncertainty. This is possible, but not highly likely, because the subjects were familiar with the statistics and decision analysis concepts crucial in the assessment of subjective pds, and the tasks the subjects worked on were within their domain of experience. Another possible explanation is that high between subject variability in terms of the assessed pds may have masked the effects of the studied debiasing techniques. A carefully conducted repeated measures design, if it can overcome potential demand artifacts, may attempt to check the soundness of this explanation. Yet, if the uncertainty ratings and the three other indicators of uncertainty measure the same construct, then one could expect high variability in the ratings of uncertainty, too. Therefore, debiasing effects would be masked also for the ratings of uncertainty,

[1]Since an *F*-test on sample variances rejected the null hypothesis of equal variances, an approximate *t*-statistic with adjusted degrees of freedom was calculated using the separate rather than pooled variance estimate.

which was not the case. Another explanation, which we believe is a plausible one, is that ratings of uncertainty and the measures based on elicited pds do not measure the same construct. This would be a convincing argument for those who suggest that elicited subjective probabilities which obey the additivity axiom may not be sufficient to capture subjective uncertainty, mainly because they ignore the strength of the evidence on which the probability judgments are based [8] [10] [11] [35] [36] [37] [39]. As a remedy, different approaches have been proposed to relate the fuzziness of the probability judgments, typically called second order probability, to the strength of the related evidence. In the current context, the subjects' ratings of perceived uncertainty may have included the second order uncertainty, but the three remaining dependent variables which were basically related to the dispersion of the subjective pds reflected only the first order uncertainty. If this explanation is valid, then it can be concluded that debiasing techniques can increase or decrease assessors' doubts about the strength of the evidence they need to recall, and thereby affect assessors' second order uncertainty. An empirical test of this conclusion can be carried out in a replication which includes a proper operationalization of second order uncertainty as a dependent variable in addition to others that measure first order uncertainty.

The sample evidence was not strong enough to confirm the hypothesis that the recommended mental encoding aid would help the assessors overcome the tendency to force a normal model on the assessed distributions. Although those who used the aid seemed to mentally capture the skewness of the observed pd better than those who did not, the difference was not statistically significant. A possible explanation is that the "bank teller" task, as a laboratory task, was so well structured and compact temporally [14] that both the control and the treatment groups could identify the skewness of the distribution, although somewhat better in the latter group. Another explanation is that the number of intervals suggested to the debiasing group to encode the observed instances was not sufficient to reflect the long right tail of the observed skewed distribution. Since both of these explanations are plausible, further studies should avoid the mentioned problem areas. These results also draw attention to the issues inherent in prespecifying the size and number of the intervals of the random quantity of interest if the discussed mental encoding aid is going to be considered a potential debiasing technique. If the number of intervals is too small, then the ability to capture the shape of the distribution is limited. However, if the number of intervals is too large, then this increases the burden on the assessor's long term memory for the observed instances.

The findings of this experiment may not seem encouraging to those practitioners who are enthusiastic about using debiasing techniques in subjective probability elicitation. However, it would be premature to conclude that the Socratic and devil's advocate approaches, and the mental encoding aid, are not effective debiasing tools, and can therefore be neglected in elicitation sessions. Further research is needed to examine the above rival explanations before these tools can be eliminated as ineffective. This is especially important since these debiasing techniques did indeed affect the ratings of subjective uncertainty in this study as hypothesized. Furthermore, the mental encoding aid seems to help the assessors better capture the location of the distributions that they observe. Therefore, the practitioners can identify the uncertain quantities they would like to monitor over time, decide on the number of intervals for each, and then begin to encode the data in their memory as they observe the new values for each uncertain quantity.
[Received: September 12, 1988. Accepted: October 3, 1989.]

REFERENCES

[1] Alpert, W., & Raiffa, H. A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press, 1982, 294-305.

[2] Berger, J. O. *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag, 1985.

[3] Cats-Baril, W. L., & Huber, G. P. Decision support systems for ill-structured problems: An empirical study. *Decision Sciences*, 1987, *18*(3), 350-372.

[4] Cohen, J. Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, 1981, *4*, 317-370.

[5] Cook, T. K., & Campbell, D. T. *Quasi experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.

[6] Cosier, R. A. The effects of three potential aids for making strategic decisions on prediction accuracy. *Organizational Behavior and Human Performance*, 1978, *22*, 295-306.

[7] Cosier, R. A., Ruple, T. L., & Aplin, J. C. An evaluation of the effectiveness of dialectic inquiry systems. *Management Science*, 1978, *24*, 1483-1490.

[8] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 1976, *38*, 325-39.

[9] Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press, 1982, 422-444.

[10] Gardenfors, P., & Sahlin, N. Unreliable probabilities, risk taking, and decision making. *Synthese*, 1982, *53*, 361-386.

[11] Good, I. J. On the principle of total evidence. *British Journal for the Philosophy of Science*, 1967, *18*, 319-321.

[12] Hogarth, R. M. Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, 1975, *70*, 271-294.

[13] Hogarth, R. M. *Judgment and choice*. Chichester, England: Wiley, 1980.

[14] Howell, W. C., & Burnett, S. A. Uncertainty measurement: A cognitive taxonomy. *Organizational Behaviour and Human Performance*, 1978, *22*, 45-68.

[15] Jenkins, A. M. *A program of research for investigating management information systems*. Internal report, Indiana University, 1982.

[16] Kahneman, D., Slovic, P., & Tversky, A. *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press, 1982.

[17] Kahneman, D., & Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 1972, *3*, 430-454.

[18] Kahneman, D., & Tversky, A. On the study of statistical intuitions. *Cognition*, 1982, *11*, 123-141.

[19] Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 1980, *6*, 107-118.

[20] Lathrop, R. G. Perceived variability. *Journal of Experimental Psychology*, 1967, *73*(4), 498-502.

[21] Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press, 1982, 306-334.

[22] Lopes, Lola L. Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology*, 1982, *8*, 626-636.

[23] Lynch, John G., Jr. On the external validity of experiments in consumer research. *Journal of Consumer Research*, 1982, *9*(3), 225-239.

[24] Murphy, A. H., & Winkler, R. L. Probability forecasting in meteorology. *Journal of the American Statistical Association*, 1984, *79*(387), 489-500.

[25] Nisbett, R. E., Krantz, D. H., Jepson, C., & Fong, G. T. Improving inductive reasoning. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press, 1982, 445-452.

[26] Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 1983, *90*(4), 339-363.

[27] Nisbett, R. E., & Ross, L. *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall, 1980.

[28] Pracht, William E., & Courtney, J. F. The effects of an interactive graphics-based DSS to support problem structuring. *Decision Sciences*, 1988, *19*(3), 598-621.

[29] Rosenkranz, R. D. *Inference, method, and decision: Towards a Bayesian philosophy of science*. Boston, MA: Reidel, 1977.

[30] Safizadeh, H. M. The internal validity of the trade-off method of conjoint analysis. *Decision Sciences*, 1989, *20*(3), 451-461.

[31] Schwenk, C. R. Effects of inquiry methods and ambiquity tolerance on prediction performance. *Decision Sciences*, 1982, *13*, 207-221.
[32] Schwenk, C. R. Some effects of planning aids and presentation media on performance and affective responses in strategic decision-making. *Management Science*, 1984, *30*, 263-272.
[33] Schwenk, C. R. Devil's advocacy and dialectical inquiry effects on prediction performance: Task involvement as a mediating variable. *Decision Sciences*, 1984, *15*(4), 449-462.
[34] Schwenk, C. R., & Cosier, R. A. Effects of the expert, devil's advocate, and dialectic inquiry methods on prediction performance. *Organizational Behavior and Human Performance*, 1980, *26*, 409-424.
[35] Shafer, G. A. *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press, 1976.
[36] Spetzler, C. S., & Stael von Holstein, C. S. Probability encoding in decision analysis. *Management Science*, 1975, *22*(11), 340-358.
[37] Spiegelhalter, D. J. A statistical view of uncertainty in expert systems. In William A. Gale (Ed.), *Artificial intelligence and statistics*. Reading, MA: Addison-Wesley, 1986, 17-56.
[38] Wallsten, T. S., & Budescu, D. V. Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 1983, *29*(2), 151-73.
[39] Wallsten, T. S., Forsyth, B. H., & Budescu, D. V. Stability and coherence of health expert's upper and lower subjective probabilities about dose-response functions. *Organizational Behavior and Human Performance*, 1983, *31*, 277-302.
[40] Von Winterfeldt, D., & Edwards, W. *Decision analysis and behavioral research*. Cambridge, England: Cambridge University Press, 1986.
[41] Winkler, R. L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967, *62*, 776-800.
[42] Winkler, R. L. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 1967, *62*, 1105-1120.
[43] Winkler, R. L. *Introduction to Bayesian inference and decision*. New York: Holt, Rinehart and Winston, 1972.
[44] Winkler, R. L., & Murphy, A. H. Good probability assessors. *Journal of Applied Meteorology*, 1968, *7*, 751-758.

B. Kemal Buyukkurt is Associate Professor of Marketing at Concordia University, Montreal, Canada. He holds a Ph.D. from Indiana University, and a B.A. and M.B.A. from Bogazici University, Istanbul, Turkey. Dr. Buyukkurt has published in the *Journal of Consumer Research* and *Journal of Marketing Research*. His current research interests include probabilistic choice models, measurement of subjective uncertainty and preferences, and interactive scaling.

Meral Demirbag Buyukkurt is Assistant Professor of Decision Sciences and Management Information Systems at Concordia University, Montreal, Canada. She holds M.B.A. and Ph.D. degrees from Indiana University, and a B.A. from Bogazici University, Istanbul, Turkey. Dr. Büyükkurt has published in *IIE Transactions* and *Journal of Marketing Research*. Her current research interests include measurement of subjective uncertainty and the use of graphic methods in decision making.