Uncertain intervals and interval uncertainty are not the same: On the credibility of credible intervals

Karl Halvor Teigen
University of Oslo, Norway
and
Magne Jørgensen
Simula Research Laboratory, Oslo, Norway

Running head:
Credibility of credible intervals

Correspondence to:
Karl Halvor Teigen
Department of Psychology
University of Oslo
P.O.B. 1094, Blindern
N-0317 Oslo
Norway
tel: +47 22 84 51 87
fax: +47 22 84 51 75
e-mail: k.h.teigen@psykologi.uio.no

Abstract

Estimates of confidence intervals for general knowledge items are usually too narrow. We report five experiments showing that people have much less confidence in these intervals than dictated by the assigned level of confidence. For instance, 90% intervals can be associated with an estimated confidence of 50% or less (and still lower hit rates). Moreover, interval width appears to remain stable over a wide range of instructions (high and low numeric and verbal confidence levels). This leads to a high degree of overconfidence for 90% intervals, but less for 50% intervals or for free choice intervals (without an assigned degree of confidence). To increase interval width one may have to ask exclusion rather than inclusion questions, for instance by soliciting "improbable" upper and lower values (Experiment 4), or by asking separate "more than" and "less than" questions (Experiment 5). We conclude that interval width and degree of confidence have different determinants, and cannot be regarded as equivalent ways of expressing uncertainty.

Uncertain intervals and interval uncertainty are not the same:
On the credibility of credible intervals

Laypeople and experts alike are often called upon to formulate estimates or make predictions about imperfectly known quantities, like: How many experimental subjects do I need to achieve reliable results? How much will I have to pay for a decent flat? How many states will vote for the Democratic candidate in the next presidential election? How many weeks will it take to finish the present paper? And how long time will it take to receive an answer from the journal editor?

Answers to such questions are often fraught with considerable uncertainty. This uncertainty can be expressed in two ways: (1) By adding a probabilistic modifier to the most likely estimate. This can either be a numeric expression ("it is 90% probable"), or a verbal phrase ("it is not completely certain"). (2) By using an interval rather than a point estimate ("you need 100-200 participants," "we need 4-6 weeks to complete the manuscript"). Sometimes, intervals are indicated by lower or upper limits only ("we need at least four more weeks"; "we will be finished before Easter").

The most complete uncertainty descriptions are achieved by a combination of probability and interval estimates, forming estimates that have variously been labeled *credible intervals, subjective confidence intervals, fractile assessments*, *uncertainty intervals,* or *probabilistic prediction intervals*. For instance, project managers are often encouraged to predict both the most likely effort of a new project (in work-hours) and the 90% prediction interval (minimum and maximum limits that will include the correct value with 90% certainty) (Moder, Phillips & Davis, 1995). More informally, authors submitting manuscripts to *Psychological Science* are told by the editor: "We *hope* to inform you of its status *within 8 to 12 weeks*" (italics ours). In this case the interval is not identified with a specific probability, but with a phrase most language users will understand as indicating (a) a polite, positive wish, and (b) a non-trivial, but not very high probability.

From a formal point of view, probability levels and interval magnitudes are compensatory, in the sense that high uncertainties can be expressed either by a low probability level or by a wide interval estimate. Narrow intervals ("8-12 weeks") can be compensated for by low probabilities ("a hope", "a 50% chance"), whereas high probabilities ("90% certainty") can be warranted if the interval estimates are wide enough (e.g., "less than six months"). Formal equivalence, however, does not necessarily mean psychological equivalence. The aim of the present study is to investigate the determinants for interval

magnitudes, and specifically the role played by probability levels. Will an increase in probability level be associated with a corresponding increase in interval magnitude, and vice versa?

Most studies of interval estimation have asked people to produce intervals associated with an assigned probability level, usually probabilities approaching certainty (90%, 98%, or 99%). The common finding of such studies is that the intervals people produce are far too narrow. Actual hit rates (the frequency of correct values falling inside the interval) are in many studies less than .50, leading to 50% or more "surprises", instead of the 1% - 10% that would be expected from a well-calibrated judge (Alpert & Raiffa, 1982; Klayman, Soll, Gonzáles-Vallejo & Barlas, 1999). Such results are usually described as evidence of "overconfidence". In fact, interval estimates have become the most popular and robust way of demonstrating overconfidence in textbook accounts (Bazerman, 1994; Russo & Schoemaker, 1989).

The concept of overconfidence is also used to describe the results from a different research paradigm, where people are asked to produce subjective probability estimates for their chosen answers to two-choice questions, or similar "discrete propositions" (Lichtenstein, Fischhoff & Phillips, 1982). These confidence estimates are subsequently compared to the proportion of correct answers. Overconfidence refers here to the fact that confidence estimates often exceed hit rates. For instance, respondents may say they are 80% certain that their answers are correct, while having only six out of ten correct answers (for reviews, see Arkes, 2001; Keren, 1991; Lichtenstein, Fischhoff & Phillips, 1982; McClelland & Bolger, 1994). Studies that have compared judgments of discrete events with interval estimates have found that interval estimates produce more overconfidence (e.g., Seaver, Winterfeldt & Edwards, 1978), a phenomenon referred to by Juslin, Wennerholm, and Olson (1999) as "format dependence" in subjective probability calibration.

The task given to participants in studies involving discrete propositions differs in important ways from the task in interval estimation studies. The first task requires participants to choose the most likely proposition before making a confidence estimate; probability judgments are in this case the dependent variable. In the second task, level of probability is usually prescribed by the experimenter, the dependent variable being the magnitude of the interval rather than the level of confidence. A participant producing intervals may not feel he is expressing his confidence, but rather his error margins. It may be misleading to describe too narrow intervals as "overconfidence", as long as questions about confidence have not been directly asked.

Assigned versus estimated confidence

The present experiments were designed to incorporate questions about confidence into the traditional interval estimation paradigm. This will enable us to compare prescribed confidence levels, used for generating credible intervals, with probability estimates generated in response to such intervals. To simplify the description of the studies, we will adopt the following terminology. Probability values proposed by the experimenter (as in most studies of credible intervals) will be referred to as *assigned confidence, AC*. This is distinguished from *estimated confidence, EC*, which refers to probability values produced by participants (as in studies of confidence in discrete propositions). We make a similar distinction between *estimated intervals, EI* (generated by participants) and *assigned intervals AI* (generated by the experimenter, or calculated by the subjects according to specific instructions, e.g., the most likely estimate plus/minus 50%). The possible combinations of assigned and estimated intervals and confidence levels are illustrated in Table 1.

<Insert Table 1 about here>

The standard procedure for studying "overconfidence" in credible intervals has been to ask for *estimated intervals* associated with high levels of *assigned confidence* (upper right cell in Table 1). In the present studies, results from this procedure will be compared with its mirror image, questions of *estimated confidence* associated with *assigned intervals* (lower left cell). We will also vary the levels of assigned confidence, to study its effects on EI. The assumed compensatory relationship between confidence and interval magnitude suggests that higher levels of assigned confidence should lead to wider interval estimates, and similarly, that wider assigned intervals should be associated with higher confidence estimates.

We first report the results from three studies (Experiments 1, 2, and 3), showing that AC and EC are not the same. Assigned confidence (AC) typically leads to the generation of too narrow EIs, but when these or similar intervals are assigned, people report different and usually lower confidence estimates (EC). Moreover, the magnitude of EI seems to be rather constant across a wide variety of AC (Experiment 2, 3, and 5). Similar EIs will also be reported under conditions where no confidence is assigned (Experiments 4 and 5), corresponding to the lower right cell in Table 1.

As a background for understanding these effects, we will review some potential psychological mechanisms that may affect (1) confidence estimates and (2) interval estimates in judgments under uncertainty.

Potential determinants of confidence estimates

Confidence ratings, especially those leading to overconfidence, have been explained in a variety of ways, ranging from a tendency to favor positive above negative evidence (Koriat, Fischhoff & Lichtenstein, 1980), to a lack of complete, immediate and accurate feedback (Arkes, 2001). Overconfidence has also been explained as an artifact, due to a biased sampling of questions (Gigerenzer, Hoffrage & Kleinbölting, 1991), or as a regression effect, due to random errors and unreliable measures (Erev, Wallsten & Budescu, 1994; Soll, 1996). A common theme going through several, otherwise different, theoretical accounts is the idea that the rater does not have direct access to the certainty of any particular proposition, but has to make indirect assessments based on more or less valid probability cues. Some of these cues are derived from memory (e.g., observed frequencies, category memberships, specific items of retained information), others can be of a more perceptual nature, and still others are heuristic or inference based (e.g., reflecting assumptions about the development of trends, and of the controllability and predictability of the future). Whereas some of these cues are positive, indicating confidence, other cues may be negative, suggesting uncertainty. As claimed by support theory (Tversky & Koehler, 1994), any probability estimate may be regarded as a compromise between evidence favoring a target proposition, and evidence favoring *other* propositions (implying that the target proposition may be wrong). One method for reducing overconfidence implies a conscious effort to argue against one's better judgment, "stop to consider that you might be wrong" (Arkes, 2001; Slovic & Fischhoff, 1977).

Following Kahneman and Tversky's (1982) distinction between internal and external uncertainty, we may think of confidence judgments as reflecting (1) the judge's subjective expertise, i.e., an individual's degree of trust, or lack of trust, in his or her own knowledge, and (2) the degree of variability believed to be associated with the target value. Prediction problems, as reflected in the "planning fallacy" (Buehler, Griffin & Ross, 1994) may be primarily due to an underestimation of the external uncertainties involved. For general knowledge items, which are the subject of the present investigation, attributions to external uncertainty are usually not applicable (there is not much variability associated with the birth year of Wolfgang Amadeus Mozart). In this case, degree of confidence must reflect a balance between the individual's "internal" arguments for and against a particular piece of knowledge,

suggested by the experimenter or generated by the individual. Too high probabilities may be due to an overly strong conviction in the accuracy of a particular statement, but they could also reflect an inability to imagine, or to grant the existence of alternative possibilities.

Potential determinants of interval estimates

The magnitude of credible intervals will be dependent on many of the same determinants as confidence estimates, discussed in the previous paragraph, including (1) a person's trust in his own knowledge, and (2) the degree of variability believed to be associated with the target value. Yet confidence estimates and interval estimates differ by having different foci. Awareness of missing knowledge (high uncertainty) can be expressed directly in terms of low probabilities. Interval estimates require in addition that deviating alternative outcomes are imaginable. If not, too narrow credible intervals will ensue. This has been repeatedly demonstrated for probabilistic prediction intervals for real tasks (Connolly & Dean, 1997; Jørgensen, Teigen & Moløkken, 2004), as well as for general knowledge questions (Alpert & Raiffa, 1982; Juslin, Wennerholm & Olson, 1999; Soll & Klayman, in press). In these cases, it is not enough to admit ignorance in a negative sense, as a lack of conviction in one's own best guess. In addition, uncertainty has to be expressed in a more bold or "positive" way by admitting the possibility of quite remote 'minimum' and 'maximum' values.

Intervals may be determined by other considerations that have even less to do with confidence. One is an implicit demand for communicative informativeness. This is a special case of the Gricean maxim of quantity (Grice, 1975), indicating that in a conversational context, intervals (like other parts of the information) should be as informative as possible, and thus not exceed a certain size even under conditions of relative ignorance. For instance, even if I had no idea of the time it takes to review a manuscript, I would hesitate to say that it is probably done in "less than ten years". I would prefer to make a more "informative" guess, say 2-3 months, perhaps adding that I don't really know. Yaniv and Foster (1995, 1997) have shown that people's preferences for "fine-grained" and precise values can lead to inaccurate estimates by a process of informativeness-accuracy tradeoff.

Intervals may also be affected by two strategies that can be used in any categorization task: An inclusion strategy, where the question is whether a target object should be accepted as belonging to the class; and an exclusion strategy, where the task is to reject or eliminate those items that do *not* belong to the class. These two strategies are not entirely complementary. Yaniv and Schul (1997) found that inclusion instructions led to a much smaller range of acceptable items than exclusion instructions. Instructions asking respondents

to mark alternatives "that are likely to be the correct answers" marked, on the average, 18% of the alternatives, whereas those who were given eliminations instructions (marking alternatives "that are not likely to be the correct answer"), marked 49.9% of the set, thereby implying that 50.1% were "likely". Interval estimates, where participants are asked to identify lower and upper limits for the category of correct answers ("the population of London is between …….. and …….. millions"), can be construed as a inclusion process, the question being how large and how small populations that can be accepted as belonging to the set of potential London populations. Soll and Klayman (in press) have recently suggested that when interval estimates are formulated as range judgments, they tend to be treated as a single (fuzzy) judgment, dominated by a single search of the relevant information available. This is believed to create a more narrow range than interval estimates that are formulated as two separate questions (one about the low and one about the high interval limit), which encourages people to make two separate searches, one for a low and another for a high value.

Thus narrow intervals can reflect various aspects of overconfidence, but they can also be the result of other processes (informativeness and inclusion strategies) that are conceptually distinct from confidence. This may lead a judge to propose narrow intervals that are not associated with a great deal of confidence. Thus estimated confidence (EC) may prove to be lower than the high assigned confidence levels that prompted the generation of uncertainty intervals in the first place.

EXPERIMENT 1

Experiment 1 was designed to compare interval estimates obtained by the traditional method of credible intervals, with corresponding confidence estimates of assigned intervals (the two diagonal cells in Table 1). We predict that people will be more overconfident in the first case, where their performances (hit rates) are compared to assigned, high levels of confidence, than in the second, where they are allowed to produce their own confidence estimates.

Method

Participants

Altogether 83 students were recruited in the reading room for social science students at the University of Oslo. They were randomly allocated to two conditions by receiving different questionnaires, and received an instant lottery ticket for participation.

Questionnaires

All questionnaires contained 10 general knowledge questions asking the participants to give quantitative estimate of a variety of subjects, including the population of Spain, the height of the City Hall building in Oslo, and the annual number of suicides in Norway.

After giving their most likely estimates, participants in Condition 1 (n = 44) were asked to produce 90% confidence intervals for each estimate. This was specifically described as a minimum and a maximum value that were sufficiently far apart to contain the true value in nine out of ten cases. E.g. "The population of Spain is between ….. and ….. millions, with 90% certainty".

Participants in Condition 2 (n = 39) were asked to calculate minimum values by subtracting 50% from their most likely estimates (E), and maximum values by adding 50%, yielding mean relative interval widths of 1.00 (MRI = (max-min) / E). So if the estimated value were 1000, the computed interval would be between 500 and 1500. They were then asked to estimate their degree of confidence (as a percentage) that the true answer would fall within this range.

Results and Discussion

The questions proved more difficult than intended. The hit rates in Condition 1 (estimated intervals) ranged from 5% (the number of medals won by USA in the Sydney Olympics) to 55% (population of Spain). The average participant had 2.35 rather than 9 correct answers, amounting to a mean overconfidence of 66.5% (90% - 23.5%).

The 50-150% intervals calculated by participants in Condition 2 turned out to be rather close to the interval estimates of Condition 1. Incidentally, six were wider and four were narrower, none of the differences being significant. In line with this, the hit rates were also similar, the average participant having 2.34 correct answers. Their confidence estimates were, however, much less than 90%. Mean confidence ratings ranged from 34% (length of a Jupiter year, in days) to 66% (population of Spain), with a grand mean of 52.5%. This amounts to an overconfidence of 29.1% (52.5% - 23.4%).

The experiment shows that credible intervals, with an *assigned* confidence of 90%, correspond to a much lower degree of *estimated* confidence. In both cases, the intervals were evidently too narrow, and the participants were clearly overconfident, but much less so in the estimated than in the assigned confidence condition. This makes it difficult to determine a specific magnitude of overconfidence, as the degree of confidence appears to be a function of the way the question is asked. If a high degree of confidence is required in the outset,

participants fail to compensate for it by producing wide enough intervals. With assigned intervals, participants appear willing to admit a higher degree of uncertainty.

EXPERIMENT 2

Experiment 1 used only one confidence level (90%), and one interval width (plus/minus 50% of the most likely estimate). Experiment 2 was intended to extend this design, by comparing (1) estimated intervals given in response to different assigned confidence levels, and (2) estimated confidence given in response to assigned intervals of different width.

From normative considerations, one would expect higher levels of confidence to require wider intervals, and, conversely, that wider intervals should be associated with higher confidence estimates. However, the too narrow intervals generated in Experiment 1 indicate that participants do not pay sufficient attention to the fact that they are supposed to be 90% certain. This result replicates findings from most previous studies of interval estimates (Alpert & Raiffa, 1982; Lichtenstein, Fischhoff & Phillips, 1982, Soll & Klayman, in press). In Experiment 2 we ask whether people are sensitive to variations in assigned confidence levels. Previous studies show that when the same participants have been asked to produce intervals corresponding to more than one level of confidence, for instance 50% intervals (ranging from the first to the third interquartile) in addition to the more usual 98 or 99% interval, they change the ranges accordingly (Alpert & Raiffa, 1982; Seaver, Winterfeldt & Edwards, 1978). It does not follow, however, that this pattern of response will be observed in a between-subjects design where different participants are assigned to different confidence levels. In a recent study, Jørgensen, Teigen and Moløkken (in press) asked four groups of computer science students to estimate the range of work hours they thought would be required to complete a programming task, with assigned confidence levels of 50%, 75%, 90% and 99% in the four groups, respectively. The median relative interval width turned out to be .8 in the 50% group and .7 in the four other groups, implying that level of confidence had little or no effect on the interval estimates. The present experiment was designed to replicate this finding with a set of more easy general knowledge items. In addition, we wanted to explore the reverse relationship, namely to what degree intervals of different magnitude will lead to different confidence estimates.

Method

Participants

Participants were 75 students attending a course in computer science at the University of Oslo. They were randomly allocated to six different conditions (with 11-15 participants in each) by receiving different variants of the same basic questionnaire.

Questionnaires

All questionnaires contained questions about the traveling distance between Oslo and ten other, generally well known Norwegian cities and townships (the correct distances ranging from 215 km to 2104 km).

After giving their most likely estimate, E, participants in Conditions 1-3 were asked to give minimum-maximum estimates corresponding to confidence levels of 99%, 90%, or 75%, respectively.

Participants in Conditions 4 were instead asked to suggest the probabilities for the actual distance to fall within plus/minus 10% of their most likely distance estimate (for instance, with an estimate of 300 km, they would have to evaluate the interval from 270 to 330 km). Participants in Condition 5 were given the same instructions with a plus/minus 25% interval, whereas Condition 6 were asked to evaluate plus/minus 50% intervals (corresponding to an interval from 150 to 450 km in the above example).

For each participant, mean relative interval widths (MRI) were computed based on (max-min) /E. For instance, with an E of 200 km and min/max values of 150 and 250, respectively, MRI = .50. In Conditions 4-6 MRIs were, by definition, .20, .50, and 1.00.

Results

These estimation tasks were clearly easier than the tasks in Experiment 1, resulting in more narrow intervals in the three first conditions (MRI around .50) and higher hit rates (50% or above). Yet we can observe the same pattern of results: Estimated intervals suggesting a high degree of overconfidence, which is reduced in the EC conditions (participants in Condition 5 are well calibrated, whereas participants in Condition 6 are actually underconfident).

Insert Table 2 about here

Confidence levels, interval widths (MRI) and hit rates for the six conditions can be compared in Table 2. Results from the first three conditions indicate that assigned confidence level has little, if any effect on interval estimates. The 99% confidence interval in Condition 1, and the 75% confidence interval in Condition 3 lead to almost identical relative intervals.

The hit rates are also the same in all conditions, giving evidence of too narrow intervals (overconfidence) in all conditions where interval estimates are the dependent variable.

Different intervals, on the other hand, led to different confidence estimates. Large assigned intervals (MRI = 1.00) lead to higher confidence than more narrow intervals, and, necessarily, to higher hit rates. Estimated confidences are much lower than the assigned confidences of Conditions 1-3, even in Condition 6 where the intervals are almost twice as large as those suggested by participants in the first three conditions. Assigned intervals in Condition 5 are comparable to estimated intervals in Condition 2, with a slightly lower hit rate. Yet the mean estimated confidence of these intervals in Condition 5 is around 45%, about one half of the 90% level of Condition 2.

EXPERIMENT 3

The previous study indicated that the magnitude of credible intervals can be essentially unaffected by variation in confidence levels from 99% to 75%. This called for a replication in a different domain, and with even wider variations in confidence levels, like 90% versus 50% confidence. Also, if interval estimates vary only little, or not at all, with prescribed confidence, one would expect similar intervals when people are simply asked to produce intervals with no prescribed level of confidence (corresponding to the lower right cell of Table 1). In both previous experiments, estimated confidence was lower than assigned confidence values. We would accordingly expect lower confidence estimates (less overconfidence) after an interval has been suggested than when the interval is produced in response to a prescribed level of confidence.

Method

Participants

Participants in this experiment were 237 students following a course in introductory psychology at the University of Oslo. They were randomly allocated to five conditions by receiving different variants of the same basic questionnaire, which they completed in a break between lectures.

Questionnaires

The questionnaires asked for interval estimates of birth years for five famous characters from world history (Mohammed, Newton, Mozart, Napoleon, and Einstein), and the years of death for five other famous persons (Nero, Copernicus, Galileo, Shakespeare, and Lincoln).

Participants in the two first, assigned confidence conditions were asked to state intervals that they believed would contain the true answer with 90% confidence (Condition 1) or with 50% confidence (Condition 2).

Participants in Condition 3 were asked to state intervals of their own choice, i.e., not linked to a specific level of confidence. They were then asked to give confidence estimates (percentages) indicating their probability that these intervals actually contained the true answer.

Participants in the two last, assigned interval conditions were asked to suggest intervals of 50 years (Condition 4) or 20 years (Condition 5), which they believed would contain the true answer. They were then asked to give confidence estimates (percentages) that the correct answer would actually be within the suggested interval.

Results and Discussion

Participants in Conditions 1-3 generated intervals ranging from around 50 years for Lincoln and Einstein, to 150-200 years for Nero and Mohammed, with increasing intervals for persons belonging to the more distant past. Average intervals were in these three conditions of similar magnitude, as can be seen from Table 3. Intervals from Condition 1, which were supposed to correspond to a 90% confidence, were only slighter wider than 50% confidence intervals of Condition 2, in fact five were wider and five were narrower, none of the differences being significant. The hit rates were quite low, making participants in both interval conditions (and especially in the 90% condition) appear highly overconfident.

Insert Table 3 about here

Intervals in the “free choice” condition did not differ systematically from the intervals given in response to assigned confidence levels. But when these participants were asked to describe their own confidence in these intervals, the mean estimates ranged from 32% (Nero) to 50% (Einstein). They were, in other words, less confident than participants in the assigned

confidence groups, who should, by definition, be either 90% or 50% confident in all their estimates.

The 50 years and especially the 20 years assigned intervals were clearly narrower than the estimated intervals. The 20 years interval led, as expected, to lower confidence than 50 years intervals for all ten birth- and death year estimates ($p < .01$, sign test), and also to a very low hit rate.

These results confirm the general finding from the first two experiments in yet another domain: Estimated confidence in 90% credible intervals is not equal to 90%, but much lower. Interval size does not seem to be much influenced by assigned confidence level, and will remain about the same even without instructions to match a particular level of confidence. Confidence (as a dependent variable) is, however, influenced by interval magnitude. Most participants in this study appeared to be highly overconfident, but there was no fixed degree of overconfidence. Overconfidence is massive with a high prescribed confidence level, less so with a 50% level and even lower when the confidence is not prescribed.

EXPERIMENT 4

Experiment 3 showed that people produce credible intervals of the same magnitude under quite different instructions. Intervals corresponding to 90% confidence were similar to "free" intervals, where level of confidence had not been specified. At the same time, confidence in these free intervals was much lower than 90%. Experiment 4 was designed to replicate this finding in another domain. To avoid biased sampling of items (Gigerenzer, Hertwig & Kleinbölting, 1991), they were this time randomly drawn from a finite, well-defined universe, namely the population of European capitals.

As suggested in the introduction, narrow intervals can be a result of an inclusion strategy, where participants are only searching for "likely" estimates. In the present experiment, this procedure is explicitly contrasted with instructions to produce "unlikely" estimates, namely one value that is clearly too low, and another that is clearly too high to be true. This was supposed to create wider intervals almost by force, partly because it asks the participants to bring up two separate, opposing values, and partly because it clearly encourages an exclusion strategy, the task being to name numbers outside rather than inside the expected range. With this procedure we would also expect to achieve high confidence estimates (it is after all very likely that the true value can be found between the two unlikely

extremes). Still, the estimated confidence may not necessarily attain the 90% or 98% levels that are typically required in studies of estimated credible intervals.

Method

Participants

Participants were 94 students at the Universities of Oslo and Tromsø, who were paid NOK 100 ($12) for completing this and several other unrelated judgment tasks. They were divided in two equal groups by receiving different variants of the same basic questionnaire.

Questionnaires

Two sets of ten European capitals were prepared by draws according to a table of random numbers, from the complete list of 45 European countries and their capitals (One World – Nations Online, n.d.).

List A contained (in alphabetical order) the following capitals: Andorra la Vella, Berlin, Budapest, Chisinau, Kiev, Lisbon, Minsk, Moscow, Rome, and San Marino. The list also included the name of the respective countries.

List B contained these capitals: Bern, Bucharest, London, Madrid, Paris, Riga, Sarajevo, Tallinn, Tirana, and Vaduz, along with the appropriate country names.

Participants in Group 1 (credible intervals) received one list and were asked to give a lower and an upper population estimate for each city, with 90% confidence, explained as the interval within which the correct number would fall in nine out of ten cases. Half of the participants (n=25) received List A, and the other half (n=22) received List B.

Participants in Group 2 (free intervals) received the same two lists, and were asked to give lower and upper estimates of their own choice along with their own level of confidence. As an example, 90% confidence was explained in the same way as to participants in Condition 1. One half of the participants (n=22) received List A, and the other half (n=25) received list B.

When this task was completed, participants in both groups received the second list, but with a different instruction, namely to suggest two "*improbable"* population figures for each city, one of them being clearly too low and the other being clearly too high, but without further specifications of the degree of improbability involved. Finally, they were asked to indicate how confident they were (as a percentage between 0 and 100) that the actual number would fall between these two figures. Again, 90% confidence was taken as an example, explained as the interval that would include the true number in nine out of ten cases.

Results and Discussion

Participants in Group 1, who were asked to produce 90% confidence intervals, produced too narrow intervals for all cities on both lists, with an average of 3.85 (rather than 9) correct answers. The two samples of cities gave very similar hit rates, as shown in Table 4. When asked to produce free intervals (Group 2), they offered even more narrow intervals, including only 2.72 correct answers. The difference between Group 1 and 2 is significant, t (95) = 2.41, $p < .02$. But these participants were at the same time much more willing to admit their uncertainty, giving confidence estimates around 50%. Thus the degree of overconfidence was reduced from 51.5% in Group 1 to 23.8% in Group 2.

Insert table 4 about here

When asked to give "improbably" low and high population figures, participants in both conditions generated much wider intervals, containing the correct population figures in 6.53 out of ten cases. Despite the "improbability" of the high and low values, the participants were far from sure about their success in capturing the correct population figures, and gave confidence estimates around 75%, i.e., not far away from their actual hit rates. Paradoxically, these intervals were much *wider* than the intervals intended to correspond to a 90% confidence level in the first part of the experiment, despite a lower reported self-confidence.

A closer analysis of errors (actual values outside the confidence intervals) showed both over- and underestimation, for instance the population of Bucharest and Moscow were typically underestimated, whereas the populations of Bern and Andorra (small capitals) were typically overestimated. This can be explained partly as a regression effect, but some large, well-known capitals (Paris, Rome) were also overestimated. Overestimations were in this experiment generally two to three times more common than underestimations, over all conditions, perhaps reflecting a belief that capitals, being important cities, should also have high populations (May, 1986).

EXPERIMENT 5

The purpose of Experiment 5 was (1) to study intervals produced in response to verbal probability phrases rather than numeric probabilities, (2) to compare intervals produced by

different elicitation methods, (3) to measure overconfidence/underconfidence by comparing actual hit rates to estimated hit rates, and (4) to study the effects of feedback on subsequent performance.

1) The previous experiments showed that similar intervals were produced in response to different levels of probability, the width of 90%, 70%, and 50% intervals being of approximately the same magnitude. In the present experiment, verbal phrases were used instead of numeric probabilities. Verbal phrases are in daily life more common and perhaps more meaningful than numbers. We have all used phrases like "I believe that …" or "I am quite certain that …", but may find it more inconvenient or artificial to use numeric expressions like "I am 75% (or 90%) sure". In the present study we asked participants to produce probability intervals either based on the phrase "I believe that ..", or "I am quite certain that …". These phrases do not represent specific probabilities, as all attempts to translate verbal phrases into numbers have concluded that such phrases are vague, and can be used to characterize a wide segment of the probability dimension (for overviews, see Budescu & Wallsten, 1995; Teigen & Brun, 2003). Yet there will be a general agreement on group level that some phrases indicate higher probabilities than others. We assumed that "quite certain" would indicate probabilities close to 1, whereas "I believe that" is less specific and include probabilities in the whole range from .5 and upwards.

2) Soll and Klayman (in press) found that participants who were asked separate questions about the lower bound and the higher bound of the probability interval (the two point method), produced wider intervals (and hence, less overconfidence) than participants who were using the more conventional range method. The range method refers to questions like "I am 80% sure that Charles Dickens was born between …….. and …….". In the two-point method, this interval is broken down to two questions, namely "I am 90% sure that Charles Dickens was born after ……." and "I am 90% sure that Charles Dickens was born before ……".

Soll and Klayman asked the same participants to estimate both lower and higher bounds. In the present experiment, questions about lower and higher bounds were given to separate groups of participants. We did not formulate specific predictions about the effect of this method (the study was planned and conducted before Soll and Klayman's research was known to us). One could argue both ways:

When participants are asked to concentrate exclusively on the lower bound, or on the upper bound, they may recruit different magnitude information, leading to either quite low or quite high estimates. Single boundary questions (between-subjects design) make this procedure even more different from the range method, and thus any difference between the range method and the two-point method is likely to become more prominent. This makes it reasonable to expect less overconfidence with the single boundary method than with the range method.

On the other hand, the need to produce informative statements could pull in the opposite direction. If I would like to tell you something meaningful about Charles Dickens' birth year, using the range method, an early lower bound could be to some extent be counterbalanced by a late higher bound (this would also be the case if I were using the two-point method in a within-subjects design). In contrast, a single early lower bound leaves it completely open where the most likely value is assumed to fall. Moreover, we use the same probability phrase for range estimates as for single bound estimates. This should encourage wider intervals by the range method. In Soll and Klayman's study, the two-point probabilities were 90%, while the range probability was stated to be 80%, to compensate for the fact that 80% of the distribution falls between the 10^{th} and the 90^{th} percentile.

3) Calibration is in most studies of overconfidence measured by comparing mean confidence estimates with average hit rates. This procedure rests upon a belief that these two estimates are (or should be) comparable. But individual confidence estimates suggest a concept of personal probabilities for unique events, whereas the hit rate is a purely frequentistic concept. In many studies (like the preceding experiments in this article), this problem is circumvented by explicitly defining confidence in terms of frequencies (for instance by saying that 90% confidence means 9 out of 10 correct answers). Yet there is some evidence that mean confidence estimates and estimates of the number of correct responses do not always correspond, not even when done by the same subjects (Sniezek & Buckley, 1991). Frequency estimates (of number of correct responses) are typically more realistic than average confidence estimates. This may be due to an explicit focus on frequencies, and also by allowing the respondents to take a more detached "outside view" on their own performance. It may be easier to admit: "I am often wrong," after a set of answers have been performed, than to claim: "I am probably wrong," after each answer. In the present experiment, participants were not asked to give individual confidence estimates per item, but were instead asked to estimate their own number of hits after they had completed the set of ten items.

4) The fourth issue addressed in the present study is the effects of feedback on subsequent performance. When overconfident estimators are informed about the correct values, they can conclude that the intervals should have been wider, or that the confidence level should have been lower. In addition, they will have learned something about the typical magnitudes characterizing objects in this particular domain. All these lessons can, in principle, be carried over to a new task within the same field, leading to improved performance and better calibration. In an earlier study (Jørgensen & Teigen, 2002) we found predicted effort intervals of the time taken to complete various software programming projects improve over time (but rather slowly). It appeared easier for participants to learn to lower their confidence to an appropriate level than to increase their interval magnitudes.

In the present study, participants received feedback on their first task (a set of ten interval estimates), enabling them to compare their estimated number of hits to their actual hit rates. They were then given a second estimation task, with ten new items drawn from the same universe (European capitals). A comparison of task 1 and task 2 will show how participants have modified their performance, in the direction of (a) improved accuracy, (b) adjusted intervals (wider or narrower, dependent upon degree of overconfidence / underconfidence), or (c) adjusted confidence (more realistic hit rate estimates).

Method

Participants

Participants were 354 students (81 men, 235 women, 38 did not report sex), attending a course in introductory psychology at the University of Oslo. They were randomly divided in six groups, by receiving different version of the questionnaire.

Questionnaires

The questionnaires contained the same two lists of European capitals that were used in Experiment 4. Participants in Condition 1 were asked to estimate population intervals for all cities by the range method (lower and higher bounds). In Conditions 2 and 3, they were asked to suggest either a lower boundary, or an upper boundary, but not both. Level of confidence was manipulated by asking half the subjects to describe intervals that they *believed* would include the true answer, whereas the other half were asked to give intervals that they were *quite certain* would contain the correct number.

Thus, questions about the population of each city were asked (to different subjects) in six different ways:

1a: I *believe* that London has between …….. and ……inhabitants
1b: I am *quite certain* that London has between …… and …….. inhabitants
2a: I *believe* that London has more than …….. inhabitants
2b: I am *quite certain* that London has more than …….. inhabitants
3a: I *believe* that London has less than …….. inhabitants
3b: I am *quite certain* that London has less than …….. inhabitants

After completing the first set of ten estimates, the participants were asked to estimate their aggregated number of hits, by completing the statement: "I think I have ……. correct answers". They were then allowed to open a second envelope containing a list of the true population figures, which they were to check against their own estimates, computing their actual number of hits.

The envelope also contained a questionnaire with a second list of capitals, to be completed in the same way. Half the participants in each of the six groups received the A list (starting with Andorra la Vella) as their first task, followed by the B list (starting with Bern), whereas the other half received the two lists in opposite order. The lists proved to have the same level of difficulty, with 4.93 correct answers to list A and 4.92 correct answers to list B (averaged over all conditions and presentation order), so the performance on these two lists were pooled.

Manipulation check

We tested the assumption that *I am quite certain* reflects a higher probability than *I believe*, by presenting both phrases to an independent control group, consisting of 30 employees in a government agency (ranging from secretaries to lawyers). Participants in this condition were asked to rate *I am quite certain that* and *I believe that* (along with the filler item *I guess that*) on 0-100% visual analogue probability scales. *Quite certain* achieved a mean rating of 85%, whereas *believe* was given a mean rating of 68%, confirming that these two phrases are associated with different levels of confidence.

Results

Confidence levels

Participants in the moderate confidence conditions (*n*=171), who were asked about what they "believed", estimated their number of hits to be 4.59 and 5.44 on task 1 and task 2, respectively (averaged over all groups). The mean estimated hits in the high confidence conditions (*n*=178), where participants gave intervals they were "quite certain" about, were almost identical, 4.65 (task 1) and 5.43 (task 2). On an a priori basis, one would think that being "quite certain" implies an expectation of having most, if not all, answers correct. Yet participants in the high confidence condition were not very confident in their own performance, estimating their number of correct items to be around 5 rather than close to 10. If "quite certain" implies a higher degree of confidence than "believe", one would further expect this group to propose wider intervals to make sure that their estimates were, in fact, correct. But the number of correct estimates (actual hit rates) were only slightly higher in the high confidence group than in the moderate confidence group, mean hit rates being .47 (task 1) and .60 (task 2) in the moderate confidence condition, versus .51 (task 1) and .61 (task 2) in the high confidence condition. This lack of difference between two verbal levels of confidence confirms our findings with numeric confidence levels. The intervals are similar, and the expectations are much the same, regardless of confidence level. Since in the present studies, the verbal phrases did not make a difference, results from these two conditions were pooled.

<insert table 5 here>

Elicitation method

Table 5 shows mean predicted hits and mean actual hits for participants in the three elicitation conditions. Hit rates are clearly higher in the single estimate conditions (lower boundaries or higher boundaries only) than in the range condition, both for task 1 and task 2. Participants in the single estimate conditions also believed that they had more correct answers.

This is a result of too narrow uncertainty intervals in the range condition. Mean upper and lower limits were calculated for each of the 20 cities in the three conditions. Three participants with extremely high upper boundaries (100 millions or 1 billion inhabitants for all cities) were excluded from this analysis, to prevent outliers to have a disproportionate effect on the averages. Mean upper and lower interval limits for all conditions are reported in Table 6. The table shows that lower limits are consistently lower and upper limits are consistently higher in the single limit conditions than in the range condition, yielding 56.2% wider intervals for Task 1 and 31.2% wider intervals for Task 2. Two-way repeated measures

ANOVAs, with elicitation method (single vs. range) and task (task 1 vs. task 2) as the two factors, show highly significant main effects of elicitation method, for lower limit estimates, F (1, 19)= 7.15, p = .015, upper limit estimates, F (1, 19)= 46.9, p < .001, as well as for interval widths, F(1, 19) = 30.0, p < .001.

<insert Table 6 about here>

Overconfidence/underconfidence

Estimated hit rates were modest and did not show a general pattern of overconfidence. Participants in the range condition overestimated their own performance on Task 1 with 47%, which was reduced to 13% on Task 2. Participants in the upper boundary conditions were generally underconfident, whereas participants in the lower boundary condition were slightly overconfident on Task 1 and clearly underconfident on Task 2. This is evident from a comparison of mean estimated and actual hits shown in Table 5, and also from the number of overconfident versus underconfident participants in the three conditions, reported in Table 7 (for Task 1 only). Soll & Klayman (in press) found a sex difference in overconfidence, males being more confident in their estimates than females. The present sample consisted of about 75% women. They estimated their own hit rates consistently lower than did the men (estimated hits on Task 1 = 4.24 vs. 5.55, and on Task 2 = 5.14 vs. 6.26, both differences being significant at the .001 level). Male students had also, on the average, more correct answers.

Effects of training

Hit rates as well as degree of calibration, i.e., the overall correspondence between hit rates and estimated hits, improved from Task 1 to Task 2 (Table 4, upper vs. lower half). This improvement seems chiefly due to a general downward adjustment of population estimates, lower limits being adjusted downwards from Task 1 to Task 2 with nearly 50%, and upper limits with about 25%. It will be recalled that many estimates, particularly for small capitals, were originally much too high. Feedback on Task 1 informed participants, among other things, that some capitals of small European countries have less than 100 000 inhabitants, whereas capitals of large countries have populations of at least 1 million, but rarely exceeding 10 millions inhabitants.

Absolute interval width (in millions of inhabitants) remained fairly constant from Task 1 to Task 2, but relative interval widths increased, as seen from Table 6, last column. This

may be regarded as a side effect of the general downward adjustment of population estimates. A population estimate of 3 millions, plus/minus 1 million, yields a MRI of .67. After a downward adjustment from 3 to 2 millions, uncertainty intervals of the same absolute size would yield a MRI of 1.00.

Insert Table 7 about here

Feedback did have an effect on confidence (estimated hit rates of Task 2), dependent upon the participants' tendency to overestimate or underestimate their own hit rates in Task 1. Mean changes in hit rates for underconfident, well calibrated and overconfident participants are shown in Table 7. Underconfident participants adjusted their estimates upwards with an average of 1.90 (from 3.90 to 5.80 estimated hits). Accurate participants also adjusted their estimates upwards, but to a lesser extent, whereas overconfident participants adjusted their estimates slightly downwards, with an average of .26 (from 4.30 to 4.04 estimated hits). A 3 x 3 ANOVA on the changes reported in Table 6 reveals no effect of condition, $F(2, 338) = .61$, n.s., but a highly significant effect of confidence, $F(2, 338) = 20.38$, $p < .0001$. Post hoc analyses (Tukey) show that all three confidence groups are different at $p < .01$, and separate one-way ANOVAs yield significant main effect in all three conditions. Thus, we can conclude that feedback might make initially overconfident subjects slightly less optimistic, but that the encouraging effects of feedback on underconfident and even on well-calibrated subjects are much stronger.

Discussion

The four main findings of the present experiment are:

1. Verbal level of confidence has no apparent effect on the width of credible intervals and estimated hit rates. This is in line with our general finding that estimated uncertainty intervals remain largely the same regardless of level of probability.

2. People can believe that they have only a modest number of correct guesses (3-6 out of ten), despite being "quite certain" about each guess.

3. Feedback made overconfident participants slightly less certain, whereas underconfident and accurate participants became clearly more confident. This asymmetry is in line with a finding by Bruine de Bruin (2002), who showed that participants in a basket ball game adjusted their chances slightly downward after each miss, but much more upwards after each hit. Feedback may in the present experiment have had some effect on the intervals, not in

an absolute sense, but by increasing their relative widths. However, the main effect of feedback was a shift towards lower and more realistic population estimates. The participants seemed to have realized that many capitals, especially in small European countries, had fewer inhabitants than they originally thought. Put differently: participants seem to use feedback primarily to improve their domain knowledge, and less so to improve their own way of handling uncertainty. By scoring their own responses to Task 1 they learnt more about city populations than about judgmental strategies.

4. Single limit estimates produce *much wider* intervals than the more traditional range method. People seem to have no problem with estimates like "London has more than 1 million inhabitants", or "London has less than 15 million inhabitants", yet few would say "London has between 1 and 15 million inhabitants". They would prefer "between 5 and 10 millions," or even more restricted ranges.

This finding has obvious practical implications. In areas where intervals tend to be too narrow (most areas studied so far), more realistic intervals can be obtained by asking separate judges to produce separate lower and upper limits.

This finding was not predicted, but is clearly in line with Soll and Klayman's (in press) results on the two-point method, where the same judges produce lower and upper limits in response to two separate questions. In their view, this is because the two questions invite informants to sample their knowledge twice. Thus, questions about lower boundaries make ideas about low populations more accessible, whereas questions about upper boundaries facilitate ideas about high populations, preparing the ground for low or high estimates through a kind of priming procedure (similar to the process believed to account for many anchoring phenomena, according to Mussweiler and Strack, 2000).

These results are less compatible with the informativeness interpretation (Yaniv & Foster, 1997), according to which wide intervals are avoided because of their lack of communicative precision. A very low lower limit (or a very high upper limit) may convey even less information than a wide interval. Intervals, even wide ones, make it possible for the listener to infer the most likely value (the interval midpoint), whereas a single estimate of the lower limit is less helpful. If we are told that London has more than one million inhabitants we have still no clue to the expected population, 1.5 and 10 millions being equally good guesses. Yet a single limit estimate, even a low one, may perhaps *appear* less vague, simply because it consists of one rather than two numbers.

In addition, the single limit questions (as well as the questions asked in the two-point approach by Soll and Klayman) are clearly formulated as tasks of exclusion rather than of

inclusion. By saying that London has *more* than 1 million inhabitants, I indicate that a population of 1 million is *outside* the category of likely populations. If I say that the true answer is *less* than 10, 15, or 20 millions, I similarly imply that these values are *too high*.

GENERAL DISCUSSION

More than thirty years of research has shown that people tend to produce too narrow uncertainty intervals. The present study adds to this body of research by showing (1) that estimated confidence does not match the assigned confidence of credible intervals, but is usually much lower, and (2) that the magnitudes of estimated intervals stay fairly constant over a wide range of assigned confidence, whereas confidence estimates are more likely to vary with interval size.

insert Table 8 about here

Estimated confidence lower than assigned confidence

The five experiments reported here offer ample opportunities to compare assigned levels of confidence to the estimated confidence generated in response to intervals of similar magnitudes. As an example, let us look at conditions where respondents have been asked to produce 90% confidence intervals (for Experiment 5, the closest equivalent to 90% would be Task 1 range judgments in the "quite certain" condition). This instruction resulted in interval estimates with a mean relative width from .23 (Exp. 2) to about .50 (Experiments 1 and 5). For Experiment 3 (year of birth or death) the concept of relative width do not make sense, as the date scale has no natural zero point. In this experiment, an assigned confidence of 90% yielded mean interval estimates of 99 years. Next, we look for conditions where participants were asked to estimate the confidence of intervals of roughly comparable magnitude. This is the assigned interval condition with a MRI of 1.00 in Experiment 1, the assigned interval condition with MRI= .50 in Experiment 2, the free interval conditions in Experiment 3 and 4, and the range condition of Experiment 5. The mean confidence estimates in all these conditions are summarized in Table 8. They show that 90% intervals can be "translated" into intervals with a mean MRI of .50-1.0, which is then "back-translated" into a confidence estimate of 40%-50%. The situation can be compared to an exchange bureau where you are paid one Euro per dollar, but when you are returning your Euros, you can only get half a dollar for each. Under such circumstances, one may well ask what is the "true" exchange rate of dollar and Euro?

Table 8 also shows hit rates for the selected intervals. The hit rates in the assigned confidence conditions range from about 23% to about 46%, which is clearly below 90%. From these figures, respondents appear to be massively (44%-67%) overconfident. Hit rates in the equivalent interval conditions were in the same range (last column in Table 8). But when these are compared to confidence estimates from the same conditions, much of the overconfidence appears to be gone. One condition (Experiment 2) shows evidence of underconfidence; in the other conditions, overconfidence is down to 12%-39%.

Stable intervals

The second main conclusion to be drawn from the present set of studies is that estimated intervals remain stable over a wide range of instructions. In Experiment 2, 99% and 75% levels confidence yielded intervals of equal magnitudes. In Experiment 3, 90% confidence yielded only slightly wider intervals than 50% confidence. This contrasts sharply with the normative requirements. With a normal distribution of errors, we should expect the high probability interval in both these cases to be more than twice as wide as the low probability interval. In Experiment 5, with verbal probability phrases, "quite certain" intervals were no wider than "believed" intervals. Furthermore "free intervals" (with no assigned level of confidence) proved to be equal to the 90% confidence interval in Experiment 3 (but somewhat more narrow in Experiment 4).

This confirms a previous finding of effort predictions (Jørgensen, Teigen & Moløkken, 2004), where several different confidence levels yielded almost identical min-max estimates of work hours. It is also in line with Yaniv and Foster's (1997) finding that participants who were asked to generate 95% confidence intervals obtained the same number of hits as those who were asked to provide interval estimates that they merely "felt comfortable communicating". Thus we are forced to conclude that when people produce an uncertainty interval, it is not (or only in a slight degree) based on probability considerations. Hence it may be more fair to ask respondents simply to produce an interval, without specifying the degree of confidence that it is assumed to reflect.

Earlier studies using the "fractile" method (asking people to produce more than one interval) indicate that people are under some circumstances able to take into account that intervals corresponding to a confidence of 98% must be wider than intervals corresponding to 80% or 50% (Alpert & Raiffa, 1982; Juslin, Wennerholm & Olson, 1999). These studies have used a within-subjects design, highlighting the difference between high and low probabilities. Our studies show that this effect tends to disappear in a between-subjects design, where

intervals cannot be directly compared. Kahneman (2003) has argued that intuitive judgments are best studied in between-subjects designs, because such designs provide fewer cues about the target attribute that the experimenter intends to test. Thus we can conclude that people may realize the difference between the width of a high and a low probability interval when they are explicitly compared, but this requires analytical, deliberate considerations, which are not so readily accessed when only one type of intervals are asked for.

Variable confidence?

There is some evidence indicating that confidence estimates are more sensitive to variations in intervals than vice versa. In Experiment 2, assigned relative interval widths of 1.0 led to higher confidence than smaller intervals (Table 2). In Experiment 3, assigned 50 years intervals led to higher confidence than 20 years intervals (Table 3) (but free choice intervals, which were even wider, did not lead to higher confidence estimates). Finally, the very wide intervals generated in response to the upper and lower "improbable" values in Experiment 4, as well as the upper and lower single bound values in Experiment 5, were associated with higher confidence than those produced by the range method. Yet the variations in confidence were in all these cases less prominent than the much wider variations in actual hit rates produced by the interval manipulations. Thus, the present set of studies does not allow for any definite conclusions about the effect of interval size on confidence judgments.

Interval estimates and confidence estimates have different determinants

In the introduction, we claimed that uncertainty about quantities can be described in two, exchangeable and compensatory ways, namely in terms of wide or narrow uncertainty intervals, and/or in terms of high and low confidence in these intervals. The results from the present experiments suggest that these two indicators of uncertainty are, in practice, not so readily exchangeable. Interval size may be a meaningful way of expressing *external* uncertainty, where we know from experience or from theory that outcomes can vary between certain limits. With a throw of two dice, I can predict sums between 2-12 points for sure, or between 3 and 11 points with 95% confidence. With *internal* uncertainty, intervals make less sense (the birth year of Mozart will remain the same, but I don't know when). It may still be meaningful to prefer a coarse, but hopefully correct estimate ("some time around 1750", or "in the 18th century") to a sharp, but inaccurate one. Yaniv and Foster (1997) have argued that "graininess" or precision of uncertainty judgments involves a trade-off between two competing objectives: accuracy (which favors imprecise estimates) and informativeness

(which demands preciseness). These two objectives may be better served by a vague, but not completely uninformative interval, accompanied by a moderate level of confidence (e.g., 50% certainty), than with a completely uninformative statement that can be issued with 99% certainty.

A second problem with intervals for describing internal uncertainty is how to select the upper and lower bounds. It may be reasonable to place Mozart simply within the 18th century, but perhaps more problematic to state that he was born between 1700 and 1800, i.e., providing exact numbers to describe the inexactitude of one's state of knowledge. These numbers cannot be explained in the same way as the outcome range of a dice throw can be. In fact the only justification I have for suggesting these very different dates is that I have almost no clue to what I am talking about. Such lack of knowledge seems more easily and naturally communicated by "meta-cognitive" statements, referring to confidence ("I am just guessing", "I may be wrong", or "I am 50% sure"), than by suggesting intervals with wide, explicit bounds.

If we grant that too narrow intervals may be a result of a wish to communicate informative statements, why are people sometimes willing to generate much wider intervals, as we found in Experiment 4 when asking for improbable figures, and in the single question conditions in Experiment 5? Soll and Klayman (in press) have suggested that the range method yields a narrow interval because it is basically conceived as a question about the most likely event. The wide interval questions in Experiment 4 and 5 differ from this in two important respects: they are formulated as two separate questions, and they are formulated as questions of exclusion rather than inclusion. Both these features indicate that the object of communication, and hence the kind of informativeness required, differs from that of the range method. It would be an object for further studies to investigate the relative contributions of these two factors. Separate questions about minimum and maximum values (rather than about "more than" and "less than" values) might change the two-point method from an exclusion into an inclusion strategy. This might lead to a more restricted interval even with a two-question format.

Are people overconfident?

Discrepancies between estimated confidence and actual hit rates of discrete propositions have been debated as revealing genuine overconfidence, methodological artifacts, or both (Hoffrage, in press; Klayman, Soll, Gonzáles-Vallejo & Barlas, 1999). Such discrepancies have been even more prominent in the area of credible intervals, where the same

methodological criticisms do not apply (Soll & Klayman, in press). But most studies of interval overconfidence appear to have compared hit rates to assigned rather than estimated confidence. The present studies replicate these findings, showing that 90% confidence intervals can yield hit rates of 25-40% rather than 90%. However, when we compare hit rates to *estimated* confidence, the degree of overconfidence is greatly reduced. It is also greatly reduced if we ask for 50% intervals rather than 90% intervals. This does not imply that interval overconfidence is a purely methodological artifact, but it makes it difficult to conclude that overconfidence is "substantial", or even that it is "variable". If you ask people about their age and they say "30" regardless of their true age, they might be guilty of gross overestimations if they are 20, minor overestimations at 29, and even underestimations (at 35). However, one may feel that overestimations (or underestimations) are not the most appropriate way of describing this phenomenon, which should be characterized as a stable preference for being 30, rather than a variable tendency to overestimate or underestimate one's true age.

Practical implications

Uncertainty intervals are used or recommended in a number of applied settings. Standard texts on project management (Kerzner, 2001; Moder, Phillips & Davies, 1995) typically require managers to submit 'most optimistic' and 'most pessimistic' completion times. These intervals are usually defined in terms of frequencies or probabilities, as values that will not be exceeded in more than 1 per cent or 5 per cent of the time. The research reported in the present paper should make us suspicious about such estimates. Not only because people tend to give range estimates that do not match their actual hit rates, but also because they may misrepresent their own confidence in these estimates. Rather than specifying a high confidence level before asking for interval estimates, it may be better to ask for unspecified intervals followed by confidence estimates in these intervals. To avoid overconfidence, separate assessments of lower and higher interval bounds may be an even more promising procedure.

The present results are based upon "artificial" general knowledge questions, which may appear far removed from optimistic and pessimistic predictions of real life events. They are, however, in good agreement with findings from a parallel set of studies concerning effort estimates in software development projects (Jørgensen, in press; Jørgensen & Teigen, 2002; Jørgensen, Teigen & Moløkken, 2004). In one experiment, 29 software professionals were asked to estimate completion times of 30 software enhancement tasks, which had already

been performed by a different company. The tasks were described in detail, and feedback about actual completion time was given after each estimate. Half of the participants were asked to produce 90% confidence intervals around their most likely estimate. They produced too narrow intervals, starting with a hit rate of 64% for the first set of 10 tasks, increasing to 81% for the last 10 tasks. The other half were instead asked to estimate their confidence in an assigned interval, with minimum and maximum values arbitrarily set as 50% and 200% of most likely estimate (based on recommendations in NASA, 1990, for uncertainty intervals of new projects). This led to intervals with hit rates of 67%-73%. More important, the estimated confidence in these intervals were quite realistic, ranging from 73% to 71%, and thus clearly lower than the 90% assigned confidence of the first group. Thus, we believe that the discrepancy between uncertainty intervals and interval uncertainty, demonstrated in the present article, reflects a general problem of uncertainty estimation, not restricted to a specific subset of laboratory tasks.

References

Alpert, M. & Raiffa, H. (1982). A progress report on the training of probability advisors. In D. Kahneman, P. Slovic and A. Tversky, *Judgment under uncertainty: Heuristics and biases* (pp. 294-305). Cambridge: Cambridge University Press.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (ed.), *Principles of forecasting* (pp. 495-515). Boston: Kluwer Academic Publishers.

Bazerman, M. H. (1994). *Judgment in managerial decision making*, 3rd Ed. N.Y.: Wiley.

Bruine de Bruin, W. (2002). *Self-serving bias in learning from experience: Practice makes perfect?* Unpublished manuscript, Department of Technology Management, Eindhoven University of Technology.

Budescu, D. V. & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *The Psychology of Learning and Motivation, 32*, 275-318.

Buehler, R. Griffin, D. & Ross, M. (1994). Exploring the 'planning fallacy': why people underestimate their task completion time. *Journal of Personality and Social Psychology, 67*, 366-381.

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

Erev, I., Wallsten, T. S. & Budescu, D. V. (1994). Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review, 101*, 519-527.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press.

Hoffrage, U. (in press). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: Fallacies and biases in thinking, judgment, and memory*. Hove: Psychology Press.

Jørgensen, M. (in press). Realism in assessment of effort estimation uncertainty: It matters how you ask. *IEEE Transactions on Software Engineering*.

Jørgensen, M. & Teigen, K. H. (2002). Uncertainty intervals versus interval uncertainty: An alternative method for eliciting effort prediction intervals in software development projects. *Proceedings of International Conference on Project Management*, (pp. 343-352). Singapore: ProMAC-2002.

Jørgensen, M., Teigen, K. H. & Moløkken, K. (2004). Better sure than safe? Overconfidence in judgment based software development effort prediction intervals. *Journal of Systems and Software, 70*, 79-93.

Juslin, P., Wennerholm, P. & Olson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1038-1052.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697-720.

Kahneman, D. & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*, 143-157

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

Kerzner, H. (2001). *Project management: A systems approach to planning, scheduling, and controlling*. New York: Wiley.

Klayman, J., Soll, J. B., Gonzáles-Vallejo, C. & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216-247.

Kuhn, K. M. & Sniezek, J. A. (1996). Confidence and uncertainty in judgmental forecasting: Differential effects of scenario presentation. *Journal of Behavioral Decision Making, 9*, 231-247.

McClelland, A. G. R. & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980-94. G. Wright and P. Ayton (Eds.), *Subjective probability* (pp. 453-482). Chichester, John Wiley.

May, R. S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, H. Jungermann, P. Lourens, and G. Sevon (eds.), *New directions in research on decision making.* Amsterdam: North Holland.

Moder, J. J., Phillips, C. R. & Davis, E. W. (1995). *Project management with CPM, PERT and precedence diagramming*. Wisconsin: Blitz Publishing Company.

Mussweiler, T. & Strack, F. (2000). Comparing is believing: a selective accessibility model of judgmental anchoring. In W. Stroebe and M. Hewstone (Eds.), *European Review of Social Psychology, 10* (pp. 135-167). Chichester, U.K.: Wiley.

NASA (1990). *Manager's handbook for software development*. Greenbelt, MD: Goddard Space Flight Center.

One world – Nations online. (n.d.). *Capitals and states of the world – Europe*. Retrieved September 9, 2002, from http://www.nationsonline.org/oneworld/capitals_europe.htp .

Russo, J. E. & Schoemaker, P. J. H. (1989). *Decision traps: Ten barriers to brilliant decision making and how to overcome them*. New York: Simon and Schuster.

Seaver, D. A., Winterfeldt, D. v. & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance*, *21*, 352-379.

Slovic, P. & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 544-551.

Sniezek, J. A. & Buckley, T. (1991). Confidence depends on level of aggregation, *Journal of Behavioral Decision Making, 4*, 263-272.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure, *Organizational Behavior and Human Decision Processes, 65,* 117-137.

Soll, J. B. & Klayman, J. (in press). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Teigen, K. H. & Brun, W. (2003). Verbal expressions of probability and uncertainty. In D. Hardman and L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 125-145). Chichester: Wiley.

Tversky, A. & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547-567

Yaniv, I. & Foster, D. P. (1995). Graininess of judgment under uncertainty: An informativenes-accuracy tradeoff. *Journal of Experimental Psychology: General*, *124*, 424-432.

Yaniv, I. & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making, 10*, 21-32.

Yaniv, I. & Schul, Y. (1997). Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making*, *10*, 211-220.

Acknowledgements

This research was supported by grant No. 135854/350 from the Research Council of Norway to the first author.

Thanks are due to Siri Skåre-Botner and Hege Undem Store for valuable assistance in conducting, scoring, and analyzing Experiment 5.

Table 1

Paradigms for studying subjective confidence in intervals

		Intervals	
		Assigned	*Estimated*
Confidence			
	Assigned	--	Credible intervals with specified confidence
	Estimated	Credibility of specified intervals	Credibility of free intervals

Table 2

Confidence levels, interval widths (MRI), and hit rates in Experiment 2

Condition	Confidence	MRI	hit rates	*n*
Credible intervals				
1	**99%**	.60	57.1%	11
2	**90%**	.46	51.5%	13
3	**75%**	.62	50.2%	12
Interval credibility				
4	53.0%	**.20**	31.5%	13
5	45.1%	**.50**	43.6%	14
6	62.4%	**1.00**	78.8%	11

Note: Confidence and MRI in bold are assigned values.

Table 3

Mean confidence intervals, mean confidence, and hit rates in five conditions, Experiment 3

Condition	Interval (years)	Confidence	Hit rate
50% confidence	84.9	***50%***	22.6%
90% confidence	99.0	***90%***	22.8%
Free choice	93.8	42.4%	27.2%
50 years interval	***50***	50.6%	25.4%
20 years interval	***20***	43.3%	10.7%

Note: Confidence values and intervals in bold are assigned values.

Table 4

Mean hit rates and mean confidence estimates for population intervals, Experiment 4

	Part 1				Part 2 (both groups)	
	Group 1		Group 2		Improbable values	
	hit rate	confidence	hit rate	confidence	hit rate	confidence
List A	37.9	**90**%	21.6	46%	66.7	78%
List B	39.1	**90**%	32.9	55%	63.9	72%
Total	38.5	**90**%	27.2	51%	65.3	75%

Table 5

Estimated and actual hits for task 1 (before feedback) and task 2 (after feedback) for participants in range and single limit estimates condition, Experiment 5

	Condition 1 Range estimates n = 121	Condition 2 Lower limits only n = 98	Condition 3 Upper limits only n = 136
Task 1			
Estimated hits	3.87	5.24	4.84
Actual hits	2.64	4.61	7.19
Task 2			
Estimated hits	4.13	5.59	6.50
Actual hits	3.64	7.12	7.50

Table 6

Mean lower and upper population limits (in millions) and credible intervals for range estimates and single limit estimates, averaged over 20 capitals, Experiment 5

	Lower limit	Upper limit	Interval width	
			Absolute	Relative
Range estimates (Condition 1)				
Task 1	2.37	4.31	1.94	.92
Task 2	1.27	3.29	2.02	1.28
Single limit estimates (Conditions 2 + 3)				
Task 1	2.04	5.07	3.03	.64
Task 2	1.04	3.79	2.65	1.02

Note. Absolute interval width = Upper limit – lower limit estimates. Relative intervals = Absolute intervals / interval midpoints.

Table 7

Mean changes in estimated hits (confidence) from Task 1 (before feedback) to Task 2 (after feedback) for underconfident, overconfident and well-calibrated participants, Experiment 5

	Condition 1 Range estimates		Condition 2 Lower limits only		Condition 3 Upper limits only	
	M	*n*	*M*	*n*	*M*	*n*
Underconfident	1.16	23	1.14	32	2.36	90
Well-calibrated	1.14	21	.39	18	.40	25
Overconfident	- .28	76	- .17	46	- .38	16

Table 8

Hit rates and mean estimated confidence for intervals of equivalent magnitude to 90% credibility intervals, all experiments

		EI Conditions		EC Conditions for equivalent intervals		
Study	Domain	Assigned confidence	Hit rates	Equivalent intervals	Estimated confidence	Hit rates
1	Almanac	90%	23.4%	Assigned MRI=1.0	52.5%	23.5%
2	Distances	90%	43.6%	Assigned MRI=.50	45.1%	51.5%
3	History	90%	22.8%	Free choice	42.4%	27.2%
4	Capitals	90%	27.2%	Free choice[a]	51.0%	38.5%
5	Capitals	Quite certain	26.4%	Range estimates	38.7%[b]	26.4%

[a] In this experiment, free choice estimates led to somewhat smaller intervals than the 90% confidence intervals, so these intervals are not strictly equivalent.

[b] Estimated confidence is based on estimated hit rates for all subjects in the range conditions.