

Overconfidence in interval estimates

Jack B. Soll

Department of Technology Management

INSEAD

Fontainebleau, France

Joshua Klayman

Center for Decision Research

Graduate School of Business

University of Chicago.

January 2003

Please do not quote or cite without permission of the authors

Abstract

Many studies over the last several decades have found that people are generally overconfident about the accuracy of their knowledge. This generalization has been overturned by a number of recent, carefully controlled studies. These studies show little or no overall bias when judges express confidence in a choice between two alternative answers to a question. Apparent overconfidence is due primarily to unsystematic error in judgments, combined with an unrepresentative selection of task items. However, Klayman et al. (1999), found that substantial overconfidence persisted under equivalently controlled conditions with a different type of confidence judgment. When judges are asked to provide intervals such that they are $x\%$ sure the correct answer is within the interval, the answer falls inside their interval much less than $x\%$ of the time. The present paper shows that, although unsystematic judgmental error may contribute to overconfidence, subjective confidence intervals are indeed systematically too narrow—sometimes only 40% as large as necessary to be well calibrated, depending on the domain of knowledge and the way in which intervals are elicited. We discuss the possible psychological mechanisms underlying this pattern of findings.

Overconfidence in interval estimates

Are people generally overconfident about their level of knowledge? Research has produced a considerable amount of evidence over the last several decades to suggest that they are. In these studies, when people say they are $x\%$ sure about a fact, they are typically right much less than $x\%$ of the time. However, the veracity and generality of the phenomenon of overconfidence have been increasingly subject to challenge in recent years. The presence of overconfidence seems to depend to a very large degree on how the questions are chosen, what they are about, and how confidence judgments are elicited. In the present study, we present evidence on how overconfidence varies with elicitation method, confirming that there is substantial bias in some, but not all, types of confidence judgments. The pattern of results across methods offers important clues to the psychological processes underlying confidence.

In most studies of confidence participants are given questions with a choice of two answers, one of which is correct. On each question the judge chooses an answer and then states his or her confidence, on a scale from 50% to 100%, that the choice is correct. “Who was born first, Charles Darwin or Charles Dickens?” “Dickens—75% sure.” In other studies, participants do not choose answers, but rather provide estimates of a quantity in terms of ranges or boundaries that correspond to a given degree of confidence. “When was Charles Dickens born?” “I’m 80% sure the answer is between 1750 and 1860.” In this paper, we examine the latter of these two types of judgments, which we call *interval estimates*.

Although interval estimates are less studied than binary choices, they are analogous to judgments that people commonly make in many contexts. When people decide when to leave for the airport, or how much to invest in stocks, or how much inventory to hold, they implicitly make judgments about a plausible interval for the time the ride will take, how much they will have in their account 20 years from now, or the rates of sales and production. Yaniv and Foster (1997) have found that people imply a rough sense of confidence in an interval estimate by choosing the precision with which they express information. “I think it was during the last half of the 19th century” implies a different degree of confidence than “I think it was around 1875.”

Studies by Klayman et al. (1999) and Juslin, Wennerholm, and Olsson (2000) found that interval estimates are prone to a great deal of overconfidence—much more so than binary choice questions. However, the processes underlying this type of judgment are not well understood. As a step in that direction, this paper examines interval estimates using different types of information and different types of questions. There are several surprises. Three methods of eliciting intervals that seem only modestly different produce very different results, with roughly 40 percentage points of overconfidence in one method and less than 20 in another. It appears that the more one breaks the interval estimate into more specific component judgments, the less overconfidence is observed. In addition, we provide different kinds of information intended to reduce some of the uncertainty in forming intervals. We were surprised to find that providing more objective information does little to reduce overconfidence. Finally, we demonstrate that unsystematic imperfections in judgment can translate into overconfidence in interval estimates. However, we also find evidence of significant overall bias: Under some conditions, intervals are less than half the size needed for good calibration.

A brief history of overconfidence

From about 1970 to 1990, studies of binary choice and interval estimates supported similar conclusions. Studies typically showed overconfidence, that is, the mean confidence

across a set of questions typically exceeded the percentage of correct answers. Binary choice tasks also showed a “hard-easy effect”: overconfidence attenuates, and sometimes reverses, as percentage correct increases. In addition, the higher the expressed confidence, the more overconfidence. For example, when judges are 90 to 99% sure in their answer to a binary choice question, they are typically correct about 70% of the time; when they are only 50% sure, they are correct around 55% of the time. The results for interval estimates have been similar, though perhaps stronger: For example, judges’ 90% intervals typically contain the correct answer less than 50% of the time (Klayman et al., 1999; Russo & Schoemaker, 1992). (The hard-easy effect has not been investigated for interval estimates.)

Researchers generally theorized that overconfidence results from biased retrieval and interpretation of evidence (e.g., Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980). Beginning in the late 1980’s, however, a different picture emerged concerning confidence in binary choice (see Budescu et al., 1997; Klayman et al., 1999; McClelland and Bolger, 1994 for syntheses). According to this new view, judges’ responses contain lots of error, but little or no bias. There are several sources of unsystematic error in subjective confidence, ranging from having accidentally had atypical experiences (Juslin, 1994; Soll, 1996) to unreliability in generating a number that goes with one’s feeling of confidence (Erev, Wallsten, & Budescu, 1994). The proportion of correct answers is also susceptible to variation depending on which particular questions happen to be sampled. Given that both confidence and accuracy are noisy measures of something else (the strength of the underlying information), mean-reversion effects are inevitable. That is, when confidence is high, it’s generally too high, and when it’s low, it’s generally too low—the typical pattern of miscalibration. Similarly, when accuracy is low for a set of questions it’s likely to be lower than the available information would lead one to expect, and when accuracy is high it’s likely to be higher than one would expect (Dawes & Mulford, 1996; Klayman et al., 1999; Suantak, Bolger, & Ferrell, 1996). This leads to overconfidence for “hard” question sets, and occasionally even underconfidence for “easy” question sets – the hard-easy effect.

A prediction of the noisy-but-unbiased view is that judges who are tested with a representative sample of questions across a wide variety of domains will on average be about right in their confidence. Because they are imperfect, they will naturally tend to be overconfident when they are very confident, but also underconfident when they are very unconfident. Recent research with binary questions has in general supported this view (e.g., Juslin, Winman, & Olsson, 2000; Klayman et al., 1999). Overconfidence may have been the predominant finding in earlier studies because of a tendency for experimenters to construct tests that favored the harder questions in any given domain. The hardest questions in a domain are naturally the most likely to be harder than a reasonable person would expect given his or her knowledge (Gigerenzer et al., 1991; Juslin, 1993; May, 1986).

What about judgments using the interval estimate paradigm? Klayman et al. (1999) found a very different picture. They used the same domains as for their binary choice questions, and again sampled randomly from all possible questions in each domain. The result: Fewer than 45% of answers fell within what were supposed to be 90% confidence intervals. Currently, little is known about when and why interval estimates produce such severe overconfidence.

Imperfection and bias, and how to tell the difference

The history of research on confidence in binary questions makes it clear that unsystematic judgmental error can affect results in ways that are difficult to anticipate and to

analyze. One important methodological safeguard is to be sure that samples of stimuli are unbiased. As in many recent confidence studies, we do this by sampling randomly from the population of available questions in a given domain of knowledge, and by examining multiple domains. A second important measure is to consider what one would expect from a population of noisy but unbiased judges, and to compare observed results to that standard rather than a standard of perfection.

Juslin et al. (2000) suggest one approach to a model of unsystematic error in confidence intervals. They extrapolate from the two choice task by postulating that, for example, a 90% confidence interval can be modeled as two binary decisions: 95% certainty about the lower limit and 95% certainty about the upper. As described earlier, the effect of unsystematic error is to make each of these estimates overly extreme. The size of the error varies with the judge's knowledge and reliability and with the predictability of the environment (see Juslin et al, 1997). A typical finding is that expressions of 95% certainty are associated with roughly 80% accuracy. Thus one might expect that a 90% confidence interval would miss about 20% at each end, and thus would include only 60% of answers. This would explain a lot of the overconfidence observed in interval judgments, although data from Juslin et al. (2000) and from Klayman et al. (1999) indicate that interval estimates show about .10 to .15 more overconfidence than predicted by this model.

There are several important caveats to the characterization of interval estimates as a pair of binary choices. As we will show later, judgments made separately for each endpoint do not show the same pattern as for a single judgment about an interval. More importantly, the primary explanation for miscalibration in binary choice does not apply to the usual interval judgment task. With binary choice questions, judges are given fixed alternatives and they respond with a range of different degrees of confidence. If the researcher then examines different subsets of responses, the mean-reverting effect discussed earlier applies. Assuming a noisy but unbiased judge, the higher the expressed confidence, the more likely it is to include some positive error, i.e., to be overconfident. So, 95% confidence judgments, being very high, are inevitably too high on average.

With interval judgments, the confidence level is fixed and the judge must determine the answer that fits it. The researcher considers all judgments, not a subset. Thus, mean-reversion effects are not expected. Across all these judgments, an unbiased judge should make a variety of errors, with judgments averaging out close to 95%. Thus, even if we accept the hypothesis that a 90% interval is made up of two 95% confidence judgments, we should not expect unbiased error to translate into overconfidence in the same way.

There is, however, a different way that unbiased judgmental noise might translate into overconfidence in interval judgments. If judges make unbiased errors in assigning dates, weights, prices, or whatever to the endpoints of their intervals, they will come out looking overconfident. To explain how this can happen, we present three mythical figures: a minor goddess by the name of Calibra, a sage named Serge, and a peasant named Peggy.

Calibra. Calibra, being only a minor goddess, is not omniscient. However, she is perfectly calibrated. Calibra's knowledge gives her a sense of how close any given value is likely to be to the truth. If you ask her the height of Mount Olympus, for example, her knowledge indicates that 300 meters is very likely to be too small, 1000 meters is probably still too small, 7000 meters is almost surely too big, and so on. Being a goddess, Calibra's sense of likelihoods corresponds precisely to a subjective probability density function (*spdf*) whose probabilities exactly match true, external probabilities of being right. When she says that she is

C% certain that the answer is $A \pm D$, she is right C% of the time. You can check this yourself. Over her indefinite lifetime, Calibra has made myriad judgments of this type and has kept precise records, and the correct answer is always revealed and recorded subsequent to her judgments. You can randomly draw a subset of judgments from the records at any level of confidence and check her hits and misses. You will find that your samples behave exactly as though they were drawn by randomly sampling a series of independent, random events of probability C.

Serge. Serge is a very sage sage. His subjective probability judgments are like those of the goddess Calibra, except, being mortal, he knows less and he is prone to error. He is completely unbiased, however; his errors are random and have a mean of zero. Specifically, when asked for an interval corresponding to C% confidence, he gives the same size of interval that perfectly-calibrated Calibra would give if she had Serge's knowledge, plus some random error. That is, Serge reports an interval of $A \pm (D + e)$, where e is a random, mean-zero error.

Peggy. Peggy is a peasant. She's not as wise as Serge is, but she's no dummy. Her confidence judgments do correlate positively with her chance of being correct. However, she makes larger random errors than Serge does, and she also is prone to bias. In particular, her intervals are systematically narrower than they ought to be.

We know that we will find Calibra to be perfectly calibrated, and Peggy will be overconfident on average. But what can we expect from Serge? As it turns out, Serge's random errors will also produce overconfidence. This makes it difficult to know the extent to which Peggy's overconfidence is due to bias rather than noise.

Why Serge is overconfident. Serge's unbiased errors translate into overconfidence because of a characteristic of the subjective probability density function (spdf) implied by his knowledge. The most likely single value is inside the interval, and the likelihood goes down as you get further from that peak. This seems likely to be true of very many spdf's. This implies that errors in setting the endpoints of the intervals have asymmetric effects depending on whether they are inward errors (making the interval narrower) or outward errors (making the interval wider). Consider the normal distribution shown in Figure 1. .

Suppose that we ask Serge to estimate the year in which Charles Darwin was born. The curve in Figure 1 represents the perfectly calibrated version of Serge's spdf. It's the curve that Calibra would have if she knew exactly what Serge knew. Calibra would say, for example, that there is an 80% chance that Darwin was born between 1800 and 1840 (interval I, shown by the solid, vertical lines). Serge, though, is prone to error. He might set an interval that's 10 years too narrow. So, he says he is 80% sure the year is between 1805 and 1835 (interval J). Calibra would have said there was 66% chance that the answer lies in that interval, and she's perfectly calibrated, so Serge is overconfident in this case.

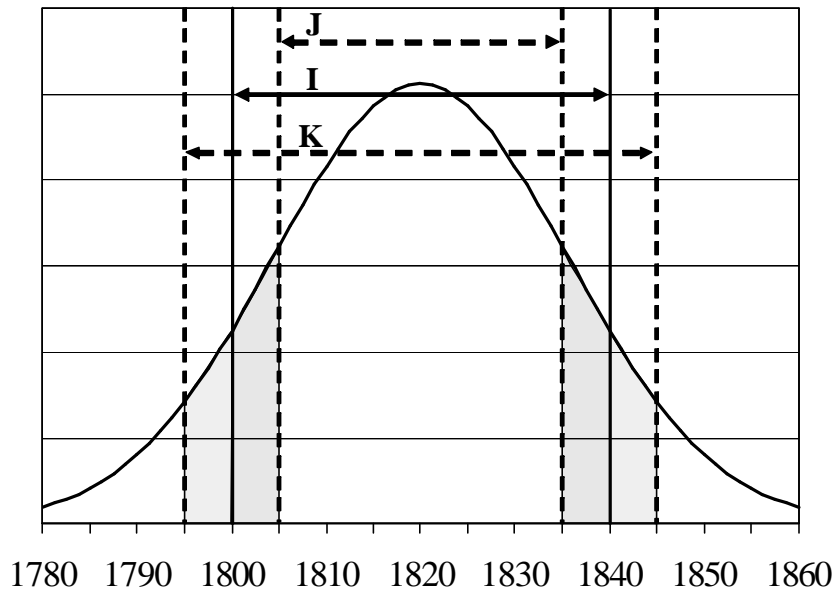


Figure 1. A hypothetical subjective probability function for an estimate of the year Charles Darwin was born. Intervals J and K represent opposite 10-year errors in estimating the interval, I, that contains 80% of the probability.

But Serge is not biased. He is equally likely to make the opposite error, making his interval 10 years too wide (interval K). Calibra would tell us that the probability that the answer lies between 1795 and 1845 is not 80%, as Serge reports, but 89%. This time, Serge is underconfident. Note, though, that the area mistakenly excluded by an inward error (the striped area) is greater than the area mistakenly added by an outward error (the gray area). Specifically, a ten-year error produces 14% overconfidence in one direction and 9% underconfidence in the other. On balance, then, unbiased errors about the size of the interval produce overconfidence. This example uses a normal curve, but this property is true of any curve that slopes upward toward a peak within the interval and downward on the other side.

Is Peggy really biased or is she just noisy? When we check the records on Peggy, we find that she is considerably more overconfident than Serge is. This could be because Peggy's errors are biased—on average, her intervals are narrower than her knowledge justifies. However, it could also be that Peggy's errors are merely larger than Serge's. In the above example, if Peggy's errors in each direction were 30 years instead of Serge's 10, they would produce 55% overconfidence and 18% underconfidence, and Peggy would be quite overconfident on average.

This raises the question of how to tell if an overconfident person is biased or just very noisy. Are Peggy's intervals the right size on average, or do they really tend to be too narrow? This is the question we want to answer for the participants in our studies. We use three methods for estimating what the "right size" is for judges like Peggy. Each method involves some simplifying assumptions and approximations. However, we can hope that the different methods converge on plausible estimates.

Compare the inferred and the observed accuracy of estimates. In this approach we pretend that each judge has a particular spdf for each question. We doubt that judges literally

carry probability density functions in their head, but as long as they have a sense that some answers are more probable than others, we can model judges as though they have *spdf*'s. We can estimate their *spdf*'s from the fractiles they give us. For example, if Serge tells us he is 90% sure that Darwin was born after 1800 and 90% sure he was born before 1840, and we assume his *spdf* is normal-distribution shaped, we can infer the curve shown in Figure 1, which has a mean of 1820 and a standard deviation of 15.6.

The *spdf* implied by fractile estimates in turn implies a distribution of expected surprises. There should be a 20% chance that the correct answer is more than 20 years away from 1820 (10% that is earlier than 1800 + 10% that it is later than 1840), a 52% chance that it is more than 10 years away, a 75% chance that it is more than 5 years away, and so on. Summing across all such probabilities for the curve shown in Figure 1, we find a mean expected absolute deviation (MEAD) of 12.4 years.

We can then look at a given judge's answers and compare two measures of surprise. The first is the observed mean absolute deviation (MAD) between the true answer and the median of her *spdf* for that estimate. The second is the mean expected absolute deviation (MEAD) implied by the fractiles she gives. For Calibra, MAD of course equals MEAD, except for some random variation due to luck of the draw in sampling questions. Importantly, the same is true for Serge, even though he's overconfident. As we noted before, given Serge's *spdf*, a ten-year mistake in one direction causes 14% overconfidence, while the same error in the other direction causes 9% underconfidence. However, equal and opposite mistakes have equal and opposite effects on MEAD (in this case, changing from 12.4 years to 10.8 or to 14.0 years, respectively). So if Serge's intervals are the right size *on average*, his MEAD and his MAD will also be the same size on average. As for Peggy, if her MAD exceeds her MEAD, that means she is receiving bigger surprises than she expects, and her intervals are too small. The ratio of MEAD to MAD is an estimate of the ratio of Peggy's average interval size to the size that represents her true accuracy.¹

Measure the correlation between interval size and hits. Greater unbiased error and narrower intervals both cause more overconfidence. However, the former has an additional effect that the latter does not. The greater the noise, the greater the correlation between the size of the interval and whether the correct answer falls in it (a "hit"). Calibra's intervals vary in size according to how much she knows about a particular question. However, her smallest 80% intervals contain the answer 80% of the time, as do her largest 80% intervals. Serge, on the other hand, also widens and narrows his intervals by accident. When he makes an inward error, his interval gets narrower and also gets less likely to catch the right answer. The opposite happens when he makes an outward error. Thus, for Serge there is a correlation between interval size and hits. For Peggy, we can identify the combination of noise + bias that would produce the combination of overall hit rate and size-to-hits correlation that we see in her responses.

How hit-rate and correlation map onto noise plus bias depends on the way that noise operates. We don't actually know this, so we use two different, plausible models for it. One model we use is

1 The assumption that MEAD / MAD is equal to the ratio of the observed interval width to the proper interval width is exact for a normal distribution with the mean in the center of the interval. We have also found via simulations that this is approximately true for a variety of skewed beta distributions. Since we do not know the shape of people's implicit *spdf*'s, this should in any case be regarded as an approximation.

$$(1) \quad W_i / W_i^* = z$$

where W_i is the width of the interval the judge gives for estimate i , and W_i^* is the well-calibrated size Calibra would have given. Variable z represents the judge's error in setting the interval size. In our simulations, we use a gamma distribution for z because it is bounded at 0, and, logically, intervals cannot be less than 0 in size. The other model we use is

$$(2) \quad \log(W_i / W_i^*) = z.$$

With logarithms, interval widths cannot be less than 0, so we can just assume a normal distribution for z . Further details are provided in the Appendix.

In the studies presented here, we find very high correlations (between .85 and 1.0) among our three methods of estimating each participant's average bias, \bar{z} . We present analyses using the MEAD/MAD measure, and note the few instances in which the other measures differed qualitatively.

Experiment 1

This study follows the approach used by Soll (1996), in which participants are not given specific, identifiable items, but rather are given only certain cues from which an estimate can be made. For example, a judge might be asked to estimate the price of an automobile, given that it had 6 cylinders, had been rated 4 out of 5 in road tests by *Consumer Reports* magazine, and was medium in size.

This approach provides a measure of subjective confidence intervals in a task that is not completely reliant on retrieval of information from memory. We went further toward reducing reliance on memory by also providing judges with information about the highest, lowest, and average value in the relevant population for each of the cues and the criterion. So, for example, judges were told that the automobiles they would be asked about came from the most recent *Consumer Reports* listing of sedans, that prices in that list ranged from \$11,325 to \$42,489, with an average of \$21,416, that the number of cylinders ranged from 4 to 8, with an average of 6, and so on. Judges of course still rely on memory for their impressions of the relations of cues to criteria, but not for the actual values of the cues or their ranges and averages. We were interested to see if reduced reliance on memory-based information would result in less overconfidence than had been observed in earlier studies of interval estimates.

A second advantage of providing objective cue information is that it provides objective standards for both accuracy and confidence. That is, we can compare the estimates of participants to statistical estimates using the same cues, and we can compare participants' confidence intervals to the statistical confidence intervals derived from regression models.

We elicited subjective intervals two ways. In one condition we asked for 80% intervals, in the other we asked for two separate 90% judgments, representing 10th and 90th percentiles. Researchers have generally assumed that judges intend the former to be equivalent to the latter, and that the two responses will be similar. We test those assumptions in the present study.

Methods

Participants. Participants were 32 undergraduate and graduate students from the University of Chicago, 9 male and 7 female in each of two conditions. They were recruited via notices posted around campus and were paid \$9 for the approximately 45 min it took to complete the procedure. Three participants in each of the two experimental conditions were replaced

because their estimates revealed that they had misunderstood the questions in one or more of the domains.

Procedure. After signing a consent form, participants were handed 1 ½ pages of printed task instructions, which they were permitted to keep with them during the procedure. The instructions included the following example:

... Child number 7, selected randomly from the files of the local primary school, weighs 86 pounds, is 51 inches tall, and has 11 permanent (adult) teeth. We...then ask you to make an estimate about how old Child 7 might be. We don't ask for a "best guess"; ...

The instructions continued with a description of one of the two elicitation methods:

... instead, we ask you to give us two numbers such that you are 80% sure that the correct answer lies somewhere between the two.

or

... instead, we ask that you give us a lower estimate such that you are 90% sure that the child is not younger than that. Then we ask you for an upper estimate such that you are 90% sure that the child is not older than that.

This was followed by an additional paragraph of explanation using the Child 7 example, including a translation of the required probabilities in terms of frequencies ("... you would expect on average to miss the answer on about 10 out of the 50 questions" or "... you would expect the correct answer to be lower than your lower estimate about 5 times and more than your upper estimate about 5 times"). The remainder of the instructions described how to locate and provide information using the computer.

The estimation task was presented using PC-type computers. Participants worked separately in a room that accommodated up to four participants at a time, separated by partitions that prevented seeing one another's screens. The task consisted of 50 questions, 12 from each of two domains, and 13 from each of two others. The opening screen of the task introduced the first domain. For example,

Hello! The next 12 questions will ask you to estimate the invoice price of different sedan-type automobiles (that is, the price the manufacturer charges the dealer.) We have randomly selected 12 out of the 67 sedans reviewed in Consumer Reports. Though we do not tell you the makes of the sedans, we give you some statistics about the year 2000 models, along with the low, high, and average for all 67 reviewed sedans. This information was drawn from the statistics available on the Consumer Reports website. These may not necessarily be the most useful statistics for making your estimates. Please hit "PAGE DOWN" to begin.

On each of the next 12 pages, an estimate in that domain was requested. Figure 2 provides an example.

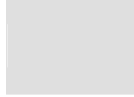
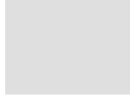

1	<u>Number of doors</u>	<u>Rating on road tests</u>	<u>Number of cylinders</u>	<u>Predicted reliability</u>	<u>Type of car</u>
	2 to 4 ave.= 4	1 to 5 ave.= 3	4 to 8 ave.= 4	1 to 5 ave.= 3	Small to Luxury ave.= Family
	this car: 4	3			Family
					
<p>I'm 80% sure that the invoice price of this car is between <input type="text"/> and <input type="text"/></p>					
<p style="text-align: center;">After entering your numbers, HIT "PAGE DOWN" FOR NEXT PAGE <Please do not return to previous estimates></p>					

Figure 2. Example of the display used to request estimates in the range-format condition of Experiment 1.

Gray boxes represented missing information. The cell marked “Dictionary” provided definitions for the criterion and each of the cues, and also the mean and range of for the criterion value (i.e., in Figure 2, for the full list of sedans from *Consumer Reports*). Figure 2 shows a screen from the range condition. Participants used the keyboard to enter numbers in each of the two outlined cells in the “between ___ and ___” line. The display in the two-point condition was the same, except that instead of completing the sentence “I’m 80% sure...,” there were two response lines that said, “I’m 90% sure that the invoice price of this car is at least [box]” and “I’m 90% sure that the invoice price of this car is at most [box].” Participants could change their responses prior to going on to the next question, but could not return to previously entered estimates.

After answering 12 or 13 questions, depending on the domain, the next screen asked participants to indicate, on a scales from 0 to 10, how much they knew about the domain and, also from 0 to 10, how important it was to them “to know a lot about” the domain.

After that, participants were presented with the next domain, using the same procedure. After completing a total of 50 questions from four domains, participants were asked, “Out of all 50 questions you answered, for how many of the 50 questions do you think the correct answer will turn out to be within the interval you gave?” They were also asked to provide some demographic information.

Design. Participants were assigned to one of two conditions. Those in the *range* condition were asked for an 80% confidence interval, and those in the *two-point* condition were asked for high and low 90% estimates, as described in the Procedures. Four domains of estimates were used, with each participant receiving all four. The order of domains was counterbalanced across subjects in a Latin Square design. The order of questions within domain was randomized at the beginning of the experiment, and was the same for all participants.

There were five possible cues available for each domain, shown in Table 1.

<u>Domain</u>	<u>Available cues</u>
This college's overall quality score	Graduation rate Academic reputation rating % students from top 10% of their high school Classes with over 50 students % of faculty who are full-time
The average box office gross of '90s movies starring this actress	Number of times she has been nominated for an Oscar Whether she uses her real name Number of films she has starred in Her age Number of children she has
Invoice price of this sedan-type car	<i>Consumer Reports</i> predicted reliability <i>Consumer Reports</i> rating on road tests Number of doors Size of car Number of cylinders
Winning % of this National Basketball Association team	Average turnovers per game Average assists per game Average rebounds per game Average blocks per game Average points per game

Table 1. Domains of estimates and the cues provided for each in Experiment 1.

Where source materials provided more than five potential cues, the five were chosen to be (a) easily comprehensible and (b) approximately representative of the distribution of cue validities among all the cues contained in the source. On any given trial, three of the five cues were presented. The selection of cues was such that each cue appeared five or six times in the 12 or 13 questions, in different combinations. This meant that different questions within a domain had different quality information, and thus should merit different interval sizes. The assignment of

which cues were given on which questions was constant across participants. In each case, the values on the cues were the actual values of a particular item in the source list. Thus, the distribution of cue values and the intercorrelations among cues approximated the population.

Due to a programming error, two of the 13 questions in the domain of college quality scores had to be eliminated from analyses; some participants saw all five cues for those items.

Results

Except where noted, results were analyzed using a multivariate analysis of variance with format (range or two-point) and gender as between-participants variables. For each participant, the dependent measures were sums or averages across the 12 or 13 trials in each domain, producing four measures per participant.

Hit rate. Overall, judges' 80% confidence intervals contained the correct answer 48% of the time. Intervals in the two-point format contained the correct answers more often than in the range format, $F(1,28) = 7.1, p = .013$, and there were significant differences among domains $F(3,26) = 5.9, p = .003$.² See Table 2. Female judges had a higher hit rate, $F(1,28) = 6.5, p = .016$. There were no significant interactions.

		<u>Interval size relative to norm</u>			
		<u>Hit rate^a</u>	<u>MEAD/MAD</u>	<u>Gamma simulation</u>	<u>Log simulation</u>
<u>Format</u>	<u>Range</u>	39	.45	.46	.44
	<u>Two-point</u>	57	.66	.71	.69
<u>Domain</u>	<u>Movie grosses</u>	39	.44	.46	.43
	<u>Basketball wins</u>	46	.57	.55	.54
	<u>Car prices</u>	51	.59	.63	.61
	<u>College ratings</u>	57	.62	.70	.68
<u>Gender</u>	<u>Female</u>	58	.70	.74	.72
	<u>Male</u>	40	.45	.47	.45

Table 2. Estimates of overconfidence and interval size relative to well-calibrated intervals for each domain of Experiment 1

^aAnswers that exactly matched an endpoint were counted as correct.

MEAD and MAD. Next, we applied the first method described earlier for estimating how the average size of subjective intervals compares to the average size of interval that would produce good calibration. For each estimate, we calculated the absolute difference between the center of the participant's interval and the correct answer. We then averaged these to obtain the mean absolute difference (MAD) for each domain for each subject. We also calculated the mean expected absolute difference (MEAD) that would result if the participant's announced endpoints

² In this and the subsequent experiments, an arcsin transformation was used in analyses of hit rates.

were the 10th and 90th percentiles of a normal distribution. We of course do not know that a normal distribution is a good representation of judges' spdf's. However, this provides a simple, parsimonious model by which to approximate the expected absolute deviation implied by the judge. The ratio of inferred, expected deviations to actual, observed deviations (MEAD/MAD) thus provides an estimate of the ratio by which subjective intervals compare to the size of well-calibrated intervals, on average. The MEAD/MAD ratio was higher in the two-point format than in the range format, $F(1,28) = 6.4, p = .018$, and varied by domain, $F(3,26) = 6.6, p = .002$. See Table 2. Males were more biased than were females, $F(1,28) = 8.8, p = .006$. There were again no significant interactions. Similar results were obtained with both of the simulation-based estimates of bias (see Table 2).

It is interesting to note that the ratio of MEAD to MAD was closer to one in the two-point condition for two reasons: MEAD was larger, $F(1,28) = 5.2, p = .031$ and MAD was smaller, $F(1,28) = 6.0, p = .021$.³ In other words, with the two-point format, judges were both more accurate and less confident. A different pattern underlies the gender difference. Males and females did not differ significantly on MAD (accuracy), only on interval width, $F(1,28) = 9.4, p = .005$.

Comparison to statistical estimates. Another way to gauge participants' confidence is to compare them to statistical estimates. We created linear regression models using each combination of cues seen by participants. These models were created using the whole population from which items were drawn (e.g., all 67 sedans reviewed by *Consumer Reports*, not just the 12 we asked about.) For each item participants saw, we obtained the statistical estimate from the relevant regression model, as well as that model's 80% confidence interval. As expected, the models' estimates were more accurate. Participants' MADs were 1.6 times as large than those of the corresponding models. But participants were also more confident. The models' intervals were 1.3 times the width of participants' intervals.

Self-ratings. Aside from providing interval estimates, participants answered three types of general questions about their task. Following each set of 12 or 13 items in a domain, participants indicated on 0-to-10 scales how much they knew about that domain and how important it was to them to know about it. Comparing individuals within each domain, neither of these measures nor their interaction was correlated with any of our performance measures. MANOVAs with format, gender, and domain as independent measures showed only a significant effect of domain on each measure, $F(3,22) = 10.3, p < .001$ for knowledge and $F(3,21) = 11.2, p < .001$ for importance. Comparing the four domains, ratings of importance were strongly correlated with hit rate ($r = .95, t(2) = 4.35, p < .05$). The correlation of hit rate with knowledge ratings was positive, but not significant ($r = .58$).

After completing all 50 trials, participants were asked to estimate how many of their intervals actually did contain the correct answer. Gigerenzer et al. (1991) asked a similar question following a series of binary choice questions. They found that the average retrospective estimate of the number of correct answers was close to the actual average number correct for a representative sample of questions from the domain. Our findings are partially consistent with theirs. Given that we asked for 80% probability intervals, participants should have been aiming

³ We wanted to include all domains in a single analysis, even though their scales were very different (e.g., millions of dollars, percentages). To roughly equate the scales, we standardized MAD and MEAD for each domain using the mean and variance of criterion values found among all items in the source list.

for 40 hits. In retrospect, they estimated that they had 31. In reality, they had 24. Thus, participants were still overconfident in retrospect, but less so than they were item by item. Even more interesting, we found a correlation of .47 ($p < .006$) between retrospective estimates and actual number of hits. Thus, participants seem to have had some insight after the fact into their degree of overconfidence.

Discussion

This study confirms the findings of Klayman et al. (1999) that confidence in intervals is very different from confidence in two-choice questions. Overconfidence is pronounced in the former case, and modest in the latter. Because unbiased variation in interval size can contribute to overconfidence, we use several methods to estimate specifically how the size of subjective intervals compares to what a well-calibrated judge would say. We find that subjective confidence intervals are indeed too narrow -- on average, less than 60% of the size they needed to be. It is interesting that providing objective cue information and explicit information about the range and mean of each of the cues and criteria did not greatly reduce overconfidence relative to what has been reported in the past (e.g., Juslin, et al., 2000; Klayman et al., 1999; Russo & Schoemaker, 1992). Apparently, overconfidence in interval judgments is not only a function of processes of retrieval of facts from memory.

We found a big difference between two ways of asking for intervals that appear on the surface to be nearly equivalent. When judges indicated they were 80% sure the answer was between ___ and ___, they were overconfident by 41%. When they separately indicated high and low two-point estimates about which they were 90% sure, they were overconfident by 23%. Similarly, in the range format, intervals were less than half the well-calibrated size, and in the two-point format, about two-thirds. Results also differed greatly by domain (see also Klayman et al., 1999). This finding highlights the risks of relying on any single domain, as a number of studies have done. We also found that men were almost twice as overconfident as women were. Women's estimates were not more accurate, but their intervals were more than 50% wider than men's.

Participants' self-ratings also showed some interesting effects. We found a very strong correlation between the average importance rating given a domain and the average hit rate in that domain. Given that this was a pot-hoc observation, that we did not see a similar pattern within domain, and that it is based on a sample of only four domains, we are cautious in interpreting this finding. However, it does suggest the hypothesis that participants monitor their state of knowledge or ignorance more carefully in domains they feel are important to know about. This could be one important contributor to differences among domains.

The fact that retrospective confidence is lower than the requested 80%, and is correlated with actual hit rate across individuals, suggests that people have some insight into their general state of knowledge that is not fully applied to individual estimates.

Experiment 2

The results of the first study established that interval estimates are indeed overly narrow. Three different methods provided similar estimates of bias in interval width, and arrived at similar conclusions regarding how bias varies across elicitation formats and domains. However, we made two assumptions in applying the methods that might be overly restrictive, which could potentially affect the results. First, for the range format, we assumed that the boundaries

participants provided excluded 10% of the probability at each tail of the spdf. This is not an issue in the two-point format, because participants were explicitly instructed to do so. Second, we assumed that the median of the spdf was halfway between the reported boundaries. A normal distribution is likely to be a good approximation if both assumptions hold true, but perhaps not otherwise.

In the present study, we obtained additional information from judges by asking them to provide explicitly their median estimate as well as separate low and high boundaries. This permits us to take into account the possibility that judges' intervals are in fact asymmetric (see O'Connor & Lawrence, 1989). Having three fractile points instead of two permits us to better approximate the underlying spdf, and to better estimate implied and actual deviations. We used beta functions for this purpose, because they can approximate a great variety of skewed distributions.

In Experiment 1, we found a surprising difference between range and two-point formats. Experiment 2 introduces another element, by requesting a median estimate. Prior research offers mixed suggestions about what effect this might have, if any. Intervals may be too narrow because of a natural tendency to anchor on a point estimate and adjust insufficiently for uncertainty. If so, then stating an explicit median estimate might have no effect, because the judge is already doing something similar covertly, or it might make intervals narrower by making the anchor more salient (Russo & Schoemaker, 1992). On the other hand, some studies (Block & Harper, 1991; Clemen, 2001; Juslin et al., 2000; Selvidge, 1980) have found that asking for a best guess in conjunction with intervals reduces overconfidence, perhaps because judges better appreciate their lack of good evidence if they encounter difficulty in generating a point estimate (Block & Harper, 1991).

Methods

Participants. Participants were 30 MBA students at INSEAD. They were solicited in a commons area on the main campus in Fontainebleau, France, and they received 50 francs and a lottery ticket for a task that took roughly 20 minutes. Two randomly chosen lottery winners received a bottle of Dom Perignon champagne. The experiment was conducted in English because the student population at INSEAD is international and all instruction is in English. We did not ask for gender or other demographic information.

Procedures. For this study we returned to the common practice of providing participants with named items and asking them to make estimates based only on their prior knowledge. We used five domains. Three of these domains were traditional "almanac questions," that is, general knowledge questions on arbitrarily chosen topics. We selected domains for which participants would have some knowledge base to draw on and for which there were non-selective lists of items from which to sample. These were the human fertility rates of different countries, the year in which a variety of devices and processes were invented or discovered, and the average daily high July temperature of major cities around the world.

In addition, we included two domains for which these participants could draw on direct, personal experience and knowledge. These were the enrollments in various courses at INSEAD, and the time required to walk from one place to another around Fontainebleau.

After reading one page of printed instructions and having an opportunity to ask questions, each participant received a printed booklet of thirty items from one of the five domains, with one item on each page. For each item, they were asked to make three estimates, corresponding to the

10th, 50th, and 90th fractiles of an spdf. For example, to guess the year that the telegraph was invented, participants completed the following three statements:

I am 90% sure that the year is after _____.
 I think the year is as likely to be after this year as before it: _____.
 I am 90% sure that the year is before _____.

They were permitted to take as much or as little time as they wished, but were told that people typically finish in about twenty minutes.

Design.

Participants were assigned alternately to each of the five domains. All participants in a given domain saw the same 30 items, but in a different randomized order. The items for each domain were selected at random from the corresponding source list. The order of the three estimate questions varied, with each possible order being assigned to one participant in each domain. For each participant the ordering of the three questions remained constant across items.

Results

Except where noted, results were analyzed using an ANOVA with domain (five levels), position of the median (first, second, or third estimate), and order of the 10th and 90th fractile estimates (10th before 90th or vice versa) as between-participant variables. Because there was only one participant in each cell, the three-way interaction was not analyzed.

	Hit rate ^a	MEAD/MAD	Gamma simulation	Log simulation
<u>Class enrollments</u>	0.49	0.55	0.62	0.65
<u>Invention dates</u>	0.60	0.55	0.71	0.66
<u>Walk times</u>	0.62	0.77	0.73	0.73
<u>July temperatures</u>	0.77	1.33	1.25	1.24
<u>Fertility rates</u>	0.83	1.59	1.36	1.32

Table 3. Estimates of overconfidence and interval size relative to well-calibrated intervals for each domain of Experiment 2.

^aAnswers that exactly matched an endpoint were counted as correct.

Hit rate. As in the previous study, judges were significantly overconfident. Their 80% intervals contained the correct answer 66% of the time. However, the 14% overconfidence observed here is considerably lower than the 40% and 22% found in the range and two-point conditions of Experiment 1. As in previous studies, overconfidence varied across domains, $F(4, 8) = 3.50, p = .06$ (see Table 3). There was a marginal effect of the order of 10th and 90th fractiles, $F(1, 8) = 3.74, p < .10$, qualified by an interaction between the position of the median and the 10th — 90th order, $F(2, 8) = 4.44, p = .05$. No other effects approached significance.

The order effects appear to have been produced by a single discrepant condition. Participants who gave their 10th fractile, then 50th, then 90th had only 39% hits. In contrast, the hit rates in the other five cells ranged from 60% to 83%.

MEAD/MAD. As in Experiment 1, the ratio of MEAD to MAD provides an estimate of the appropriateness of interval size. In Experiment 1 we measured the absolute deviation between the correct answer and the middle of the participant's interval for each item. In this study, we measure deviations from the median estimates provided by the participants. As before, we used the participant's fractiles to infer an spdf and its expected absolute deviation. This time, though, we have three estimates instead of two, allowing for the possibility of asymmetrical spdf's. For each set of three fractiles, we used the software package *@Risk* to infer the best-fitting beta function, and then estimated the expected absolute deviation from the median for that beta using Monte Carlo simulation.

The ratio of MEAD/MAD differed across domains, $F(4, 8) = 4.92, p < .05$. Although no other effects were significant, the pattern of means roughly tracks that for hit rates. This follows from the fact that across participants, MEAD/MAD is highly correlated with hit rate, $r = 0.80$. Overall, the average MEAD/MAD ratio of .96 appears much closer to unity than do the values of .47 and .68 for the range and two-point formats of Experiment 1.. Results using the estimates of interval bias from simulations are very similar to those using MEAD/MAD, as shown in Table 3.

Discussion

Experiment 2 confirms that subjective confidence intervals are too narrow. The degree of overconfidence was substantial (about 15%), but it was considerably less than we observed in Experiment 1, and subjective intervals were considerably less narrow. Looking across both studies, one could say that the larger the number of separate estimates required, the less overconfidence. When a single range was requested, intervals were roughly 45% of the well-calibrated size, when two separate boundaries were requested, they were roughly 70%, and when three fractiles were requested, they were roughly 96% of the appropriate size. Later we will offer further discussion of this observation. However, the comparison between Experiments 1 and 2 is not straightforward. They two studies were conducted using different domains, different information, and different populations. Therefore, the next study provides a direct test of the extent to which asking for a median estimate is indeed the reason for the relatively low overconfidence observed in this study.

The present study also shows an intriguing effect of the order in which the fractiles were elicited. Participants who provided estimates in the order 10-50-90 were more overconfident than those who provided estimates in different orders. We speculate that people may find this order most natural, which could lead to less deliberative, and thus more biased, information processing. However, we will refrain from interpreting this result further unless we can replicate it. Experiment 3 will provide an opportunity to do so.

Participants did not perform better in the domains for which they had direct, personal experience—class enrollments and walk times. Indeed, those two were among the most biased domains. Although we sampled only a few domains here, these findings support the conclusion that overconfidence is not restricted to general knowledge, “almanac” questions.

Experiment 3

This study uses two of the domains used in Experiment 2, namely July temperatures and the dates of inventions and discoveries. These domains were chosen because they were suitable for use with American participants without modification and because they differed considerably in the degree of overconfidence observed in Experiment 2.

Methods

Participants. The participants were forty-two students and staff at the University of Chicago, 25 female and 17 male. One participant in each of the two conditions was replaced because repetitive or arbitrary responses indicated that they were not attending to the task. Participants were solicited via advertisements posted around campus, and were paid \$10.

Materials. Forty-five items were selected from each of the two domains. These included the 30 items used in Experiment 2, plus 15 new items selected randomly from the same information sources. Estimates for each item were elicited using two different formats. The *two-point* format asked for only the high and low estimates, as in the equivalent condition of Experiment 1. These were always requested in the order of lower boundary, then upper boundary. The *three-point* format asked for two boundaries and a median estimate, as in Experiment 2. Estimates for these questions were requested in one of three orders: median, lower boundary, upper boundary; lower, median, upper; and lower, upper, median.

As in Experiment 2, printed booklets were prepared with each item on its own page. Each participant received two booklets. The first booklet contained the 30 items used in Experiment 2, in shuffled order, followed by the 15 new items in that domain, also shuffled. The second booklet contained only the 30 “old” items from that domain, in a different, shuffled order.

Design. Format (two-point or three-point) was manipulated within participants. Specifically, each participant made the first 45 estimates using one format and the last 30 using the other format. The order of formats was counterbalanced.

Three variables were manipulated between participants: domain (temperatures or discoveries), format order (two-point estimates before three-point, or vice versa), and position of the median estimate in the three-point format (first, second, or third).

There were 21 participants in each domain. Within each domain, 9 participants made two-point estimates first and 12 made three-point estimates first. In each sub-group, one third of the subjects were assigned to each of the different median-estimate positions.

Procedure. The study took place in three parts, requiring about 30 minutes, 5 minutes, and 20 minutes respectively. Introductory instructions indicated these time estimates, but participants were allowed to take as much or as little time as they wished. Part I consisted of 45 items from one of the two domains. In Part II, participants were asked to provide numerical ratings of three different types of cookies or of a recent restaurant meal, in order to provide a break and a numerical task that might reduce recall of previous estimates. These ratings were not analyzed. Part III consisted of the first 30 items, presented in a different order and in a different format.

Initial instructions described the three phases, and offered instructions for whichever format was to be used in Part I. The following example is for a participant in the domain of discoveries, using the two-point format:

For each of the 45 items in Part I, we ask you to make two estimates, by completing two statements. For example:

In what year was [] invented?

I am 90% sure that this happened after _____.

I am 90% sure that this happened before _____.

For the first statement, you should give us a year such that you believe there is a 90% chance that the discovery or invention happened after that. In other words, you believe there is a one in ten chance that the correct answer is earlier than this date.

For the second statement, you should give us a year such that you believe there is a 90% chance that the discovery or invention happened before that. In other words, you believe that there is a one in ten chance that the correct answer is later than this date.

For those participants with Part I items in the three-point format, the instructions included this additional estimate:

I think it's equally likely that this happened after or before _____.

placed in the appropriate position for that participant (i.e., before, between, or after the low and high estimates). These instructions were added:

For the third statement, you should give us a year such that you believe that the discovery or invention is about as likely to have happened after this date as before it. In other words, you believe there is an equal chance that your guess is too early or too late.

Participants were asked to consider each item in sequence, and were told they could change their answer to the current item, but could not go back to prior items.

After completing Part I, the experimenter collected the participant's booklet and handed them instructions and ratings forms for rating cookies or restaurants, which took place in another room. Following that, instructions were provided for Part III. These were the similar to those used for Part I, but now using the other format (three-point or two-point). The Part III instructions included the following:

In this final part of the study, we would like for you to revisit some of the questions that you answered in Part I....**Your answers should reflect your current beliefs about each question. Do not worry about how your answers compare to what you might have said the first time you encountered these questions.**

Results

Two types of analyses were possible given the design. In one approach, we perform ANOVAs using only the 45 questions presented in Part 1, ignoring the 30 that were repeated later, in Part 3. Domain (temperatures or discoveries), format (two-point or three-point), and gender are all between-participants variables. In another approach, we use only the 30 questions that appear both in Part 1 and Part 3 for each participant. In those, format is a within-participants variable, and there is the additional between-participants variable of which format was presented first. Initial analyses of both types showed significant main effects of gender and domain by gender interactions. Because participants were assigned to cells without regard to gender, males and females were not equally represented in each cell. To minimize confounds between gender and other effects, all analyses were redone using weighted least squares to approximate the results that would obtain with an equal distribution of males and females in each cell.

Very similar results were obtained with all four analytical methods. We present the results from the weighted, between-participants analyses, and we note the few qualitative differences found in alternative analyses.

Hit rate. Overall, judges' 80% confidence intervals contained the correct answer 55% of the time. Intervals in the three-point format contained the correct answers more often than in the two-point format, $F(1,34) = 4.7, p = .037$, and the two domains differed, $F(1,34) = 41.9, p < .001$. See Table 4. Female judges had a higher hit rate, $F(1,34) = 5.0, p = .033$. There was also a significant interaction between domain and format, $F(1,34) = 4.4, p = .043$. Format made a difference in the temperatures domain (with a hit rate of .60 in the 2-point format and .76 in the 3-point), but virtually no difference in the discoveries domain (.42 for both).

		<u>Percent within interval^a</u>	<u>Interval size relative to norm</u>		
			<u>MEAD/MAD</u>	<u>Gamma simulation</u>	<u>Log simulation</u>
<u>Format</u>					
	<u>Two-point</u>	51	.52	.58	.55
	<u>Three-point</u>	59	.67	.73	.68
<u>Domain</u>					
	<u>Temperatures</u>	68	.79	.83	.81
	<u>Discoveries</u>	42	.40	.48	.42
<u>Gender</u>					
	<u>Female</u>	59	.67	.72	.69
	<u>Male</u>	51	.52	.60	.54

Table 4. Estimates of overconfidence and interval size relative to well-calibrated intervals in Experiment 3.

^aAnswers that exactly matched an endpoint were counted as correct.

MEAD and MAD. As in the previous experiments, we calculated for each participant the ratio of expected deviations to observed deviations (MEAD/MAD). The MEAD/MAD ratio was higher in the three-point format than in the two-point format, $F(1,34) = 4.9, p = .033$, and varied by domain, $F(1,34) = 36.2, p < .001$. See Table 4. Males were more biased than were females, $F(1,34) = 5.06, p = .031$. There was a significant interaction between domain and gender, $F(1,34) = 4.7, p = .038$. Women had more appropriate interval sizes than men in the temperatures domain (ratios of .93 vs. .65), but women and men were virtually the same in the discoveries domain (both .40).

The advantage of the three-point format seems to derive from both smaller errors and larger intervals. MAD was about 10% lower in the three-point format, and intervals were about 13% larger, although neither of those components was itself statistically significant. The within-participants analysis looks a little different. There, the three-point format showed wider intervals, $F(1,34) = 6.7, p = .013$, but no significant advantage in MAD.

Women's intervals were almost 20% wider than men's, and their MADs were about 6% lower. The difference in interval widths approached significance ($p = .107$ in the between-participants analysis and .058 in the within-participants); for the difference in MAD, both F 's < 1.

Estimates obtained from simulations show results very similar to those found with MEAD/MAD, except that they also show a significant interaction between domain and format, $F(1,34) = 4.7$ and 6.8 for gamma and log models, respectively, $p = .036$ and $.013$. The reason for this can be traced to a similar interaction on MAD. The three-point format showed larger interval widths in both domains, but showed improved accuracy only in the domain of temperatures.

Order of estimates. In the three-point format, we always asked for the lower bound before the upper, but we varied the position in which we asked for the median estimate: first, between the two other questions, or last. We looked at the effects of order by examining only those questions asked in the three-point format, using the 30 questions that participants saw twice. Every participant answered those questions in the 3-point format, either in the first round or the second.

We performed between-participants ANOVAs with four independent variables: estimate order, domain, gender, and whether the 3-point questions were received before or after having given 2-point responses. The dependent variables were the same ones used in the other analyses. No main effects of estimate order approached significance. Hit rate and MEAD/MAD showed an order by domain interaction. Those measures and MAD also showed an order by gender interaction. In the temperatures domain, participants appeared to do better when the median came first or second. In the discoveries domain, they did better when it came second or third. Women did better when the median was second or third; men when it was first or second. Clearly, if order of estimates has effects, they are complex ones.

Discussion

Overconfidence was much lower in Experiment 2 than in Experiment 1. We wondered whether that was attributable to our having asked participants in Experiment 2 to provide a median estimate in addition to specifying high and low boundaries to their intervals. The results of the present study suggest that that is indeed a large part of the difference. To make a direct comparison, consider only the temperature and discovery domains of Experiment 2, and only exactly those 30 questions when answered in the 3-point format in the first round of the present study. (In other words, exclude the 15 new items added to the present study, and the second round of estimates.) Experiment 2 participants had 60% and 77% hits in discoveries and temperatures, respectively; in the present study the figures were 48% and 82%. (The figures shown in Table 4 are lower, because it happens that the 15 new items were more prone to overconfidence.)

Within the present study, we find that adding the median estimate improved hit rates considerably, although all the of the improvement was in one domain. We did not replicate the finding in Experiment 2 that providing estimates in the default, logical order of low-median-high produced more overconfidence. Indeed, we found no clear effect of where we put the median estimate in the sequence.

In sum, the present study supports the hypothesis that making additional, separate estimates about different parts of one's subjective probability distribution reduces overconfidence. Doing so seems both to improve the accuracy of one's estimates, and to increase the size of one's confidence interval. At the same time, we continue to find noticeable differences between populations and between domains. These differences suggest caution in generalizing from results, but they may also provide interesting clues to underlying processes.

How much overconfidence is due to noise?

In all of our studies, we find widespread overconfidence. As we explained at the beginning of this paper, overconfidence could result from unsystematic variations in interval size, even if interval were the right size on average. Clearly that is not what is happening. Using three different estimation methods, we find that people's intervals are smaller than they should be. Nevertheless, unsystematic error might still contribute to the overconfidence we observe. How much overconfidence is attributable to bias (i.e., narrow intervals) and how much to noise (i.e., unsystematic variation in interval size)?

To answer the question, we need to estimate what the participant's hit rate would have been if each one of their intervals were biased by a constant amount, without variation. Assuming that people's spdf's are normal, the 10th and 90th fractiles should be $\pm 1.28\sigma$ from the midpoint. So, we calculated what probability fell between fractiles set at $\pm \bar{z} * 1.28\sigma$ instead, where \bar{z} is the participant's average bias in that domain. For example, if a participant's average bias was .5, we checked what probability is included between $\pm (.5 * 1.28)\sigma$ in a normal distribution. The answer is 48%. We then looked to see if the participant had in fact had more or less than 48% hits in that domain. The difference between his or her actual hit rate and the expected rate with constant bias is our estimate of the biasing effect of noise.

In Experiment 2, participants gave three fractiles instead of two, so we were able to do better than just assume that each distribution was normal. In the case of asymmetric distributions, we performed the same process on each inferred beta-shaped spdf, moving each fractile in toward the median by the average bias, \bar{z} . As it turns out, this refinement produced almost identical results to those we got if we simply ignored the expressed median and assumed the distributions were normal, supporting the reasonableness of that assumption in the other studies.

Table 5 shows the results using MEAD/MAD as the estimate of bias. Results based on estimates from the simulations are similar. The first column shows the difference between the actual and constant- \bar{z} hit rates calculated for each person within each domain. Positive values indicate that variation in interval width contributed to overconfidence, while negative values imply that variation reduced overconfidence. The values range from -.03 to .02, suggesting that within-person variation had only a minor effect on overconfidence.

Variation from item to item within individuals is only one source of noise, however. There is also variation from person to person. If people are on average unbiased, with some positively biased and some negatively, that too can produce net overconfidence. So next we computed the effect of variation in \bar{z} across participants. For each domain, the individual-level \bar{z} 's were averaged. The expected hit rate was then computed, assuming that all participants consistently applied this same domain-level \bar{z} to all items in the domain. This was compared to the theoretical hit rate we calculated previously, assuming that each individual had a constant bias. The second column of Table 5 displays the difference. The effects are slightly greater in magnitude than the individual-level effects, ranging from .01 to .06.

Study	Condition	Effects of Variation			Total	Overconfidence due to Bias
		within individual	between individuals	between domains		
Experiment 1	Range	0.02	0.03	0.00	0.05	0.36
	2-point	0.00	0.03	0.00	0.04	0.19
Experiment 2	3-point	-0.01	0.05	0.07	0.11	0.03
Experiment 3	2-point	-0.03	0.01	0.01	-0.01	0.30
	3-point	-0.02	0.02	0.03	0.03	0.17
Klayman et al. (1999) ^a	mixed	0.02	0.03	0.01	0.06	0.47
	blocked	-0.03	0.04	0.03	0.05	0.36

Table 5. Effects of variation on overconfidence.

^aIn the mixed condition, participants saw 10 questions from each of twelve domains in mixed order. In the blocked condition, participants saw 40 questions from each of three domains presented in separate blocks.

One more source of noise is variation is between domains. If some domains are positively biased and some negatively, that can produce net overconfidence, too. So we next averaged the biases across both participants and domains to obtain an overall mean bias. The difference between the corresponding hit rate and the theoretical hit rate in the previous step reflects the amount of overconfidence due to variation across domains, which ranged from .01 to .07, as shown in the third column.

Summing all three sources, the total biasing effect of noise to the ranges from -.01 to .12. Any remaining overconfidence is attributed to bias, as shown in the table. After subtracting out the effect of variation, bias still accounts for a substantial amount of overconfidence in most of the experiments. Experiment 2 is an exception—we are not sure why. In that experiment there was an unusually small amount of overconfidence to begin with; the hit rate for 80% intervals was 66%. Most of the 14% overconfidence could be accounted for by variation, leaving only 3% due to bias. Overall, however, we find that overly narrow intervals account for most of the observed overconfidence.

Conclusions

People are grossly overconfident when they provide subjective confidence intervals for a quantitative estimate. This finding differs greatly from findings from studies of confidence in binary choice questions. With binary questions, apparent overconfidence mostly evaporates when one is careful to test a representative set of questions from multiple domains. With subjective intervals, in contrast, overconfidence of as much as 45% (in 90% confidence

intervals) persists even with representative sets of questions (Klayman et. al, 1999). The present studies produce three main findings concerning subjective confidence intervals: (1) Unsystematic variation contributes to overconfidence, but the main cause is that subjective intervals are too narrow; (2) the format by which subjective intervals are solicited has a large effect on overconfidence and interval size; and (3) men are more overconfident than women are.

We of course would like to understand when and why subjective confidence intervals are too narrow. We are not yet in a position to answer those questions definitively. However, the pattern of results in the present experiments offers some interesting suggestions. One obvious possibility is that narrow intervals are the result of anchoring (presumably on a best-guess point estimate) and insufficient adjustment. To some extent, though, this is merely a restatement of the phenomenon. One would still want to know why adjustment was insufficient, and more so in some formats than in others. At the same time, some of our results seem not to fit an anchoring story very easily. It is hard to know just where a judge's anchor lies, but it is likely to be near the median (i.e., where the answer is believed to be equally likely to be above or below). Making the median estimate explicit should make the anchor even more salient, or perhaps have no effect (because judges already start from their best guess). Instead, asking for an explicit median seems to increase adjustment; it increases the size of subjective intervals.

What, then, are the likely sources of bias? When judges collect information, either from their heads or from the outside world, they obtain only a limited sample of the potential information. Moreover, the sample they draw is probably not unbiased. Because of the nature of human associative memory, information that has many semantic connections with already-retrieved information will have a retrieval advantage, making the sample more consistent than a random sample of evidence would be. Confirmation biases (see Klayman, 1995) will further contribute to the excessive consistency of information and interpretations of its implications for the answer. People do not fully anticipate the consequences of confirmation biases (Klayman, 1995) nor of limited and biased samples more generally (Fiedler, 2000). They treat their sample of evidence too much as though it were a reliable and unbiased representation of the world.

The above explanation must be regarded as speculative, pending further studies of the cognitive processes underlying confidence judgments. However, the observed pattern of findings supports this general view. Consider first the comparison between binary questions and interval estimates. Klayman et al. (1999) argue that confirmation biases are expected to have more impact on the latter. Who has a longer average life expectancy, Argentines or Canadians? The question offers two explicit alternatives: Argentines live longer, or Canadians do. Thus, both possibilities are considered more or less equally, producing error-prone, but not heavily biased, judgments of the balance of evidence. In contrast, when asked for a single range estimate, the judge starts with some initial impression of, say, the longevity of Canadians and must then judge the appropriate precision of that impression. This is more akin to the situation in which the validity of a single hypothesis is being evaluated, which is when confirmation biases have the most effect (Klayman, 1995, McKenzie, 1999).

What about the format differences observed in the present studies? When judges are asked for separate low and high 90% judgments rather than a single 80% range, their intervals become wider and the correct answers are closer to the centers of their intervals. An evidence-sampling explanation fits this case. When asked to provide a lower bound, the judge is implicitly asked to consider what the lowest plausible answer could be (specifically, one which is 90% sure to be too low). This elicits a search for evidence that Canadians are prone to early death (e.g., due to the hardships of harsh weather conditions). The separate request for the highest plausible

answer elicits a search for evidence that Canadians are hardy (e.g., due to the rigors of harsh weather conditions). The result is two separate searches for evidence in two different regions of knowledge. This increases the range of evidence considered (making judgments more accurate) while making the variability of implications more apparent (increasing interval width.)

Now, add another request, namely for the point at which the correct answer seems to be equally likely to be above or below. This might be seen as an incitement to anchor, but in the context of the two other required estimates, it provokes a third search in yet another region of knowledge, again increasing accuracy and also the sense of the variety of implications of different evidence.

As Klayman et al. (1999) did, we also find systematic differences in overconfidence between individuals, between subpopulations, and between domains of knowledge. We don't have much evidence yet about what make some individuals consistently more overconfident than others, why men are more overconfident than women are, nor why certain domains are more prone to overconfidence than others. At the same time, certain factors that we thought might mitigate overconfidence showed no such effect. These include providing objective information to use in making the estimate and asking about domains that have been personally experienced.

Having appropriate confidence is important for making appropriate risky decisions, for knowing when to seek advice and information, and for communicating one's knowledge. Judging appropriate confidence is not easy. Aside from that, there is a general tendency toward overconfidence, which can be quite severe in some conditions, but is not universal. We are optimistic that future research into differences between formats, people, and domains will advance our understanding of the psychology of confidence and our ability to help people be better judges of confidence. In the meantime, we can say that if you want to obtain a well-calibrated estimate, ask for a comparison between two options, or else ask separately about different parts of the range of possible answers. Also, be careful what you ask about, and ask a woman.

Appendix

This appendix describes the models and procedures used to estimate interval-size bias based on the observed hit rate and the covariance between interval size and hits.

Gamma Model

This model corresponds to Equation 1 in the text, $W_i = z W_i^*$, where W_i is the width of the interval the judge gives for estimate i , and W_i^* is the well-calibrated size. Across a set of items, variable z has a mean of \bar{z} and some mean-zero variation, v_i , around that, so

$$(A1) \quad W_i = W_i^*(\bar{z} + v_i).$$

We define a binary variable H_i , such that $H_i = 1$ if the interval for item i contains the correct answer (a hit) and $H_i = 0$ otherwise. If there were no variation (no v_i), there would be no covariance between the width of the interval and whether or not there is a hit. By definition, for $p\%$ intervals, W_i^* is the size of interval for which the probability of $H_i = 1$ is p . W_i^* varies from item to item, being wider when knowledge is low and narrower when knowledge is high, so that the probability of a hit is always p . Since \bar{z} is also constant, $\text{cov}(\bar{z} W_i^*, H_i) = 0$.

In the presence of this variation, however, the probability of a hit does vary with interval size. A larger v_i means a larger interval for the same degree of knowledge, and thus a higher chance of a hit. On average, wider intervals will be associated with higher hit rates, that is, $\text{cov}(W_i, H_i) > 0$. By observing the size of this covariance in the data, we can get a clue as to the amount of bias and variation across a given set of items (typically, for a given participant in a particular domain).

We need, therefore, to find the mathematical relationship between the observable quantity $\text{cov}(W_i, H_i)$ and the theoretical quantities of interest, \bar{z} and v_i . Beginning with Equation A1, we have:

$$W_i = \bar{z} W_i^* + v_i W_i^*$$

$$(A2) \quad \text{cov}(W_i, H_i) = \text{cov}(\bar{z} W_i^*, H_i) + \text{cov}(v_i W_i^*, H_i).$$

To proceed, we need to make an additional, simplifying assumption, namely that H is a function only of \bar{z} and v_i , and not of W_i^* . This implies that W_i^* is independent of v_i as well as H_i and any other functions of v_i . Substantively, we are assuming that overconfidence does not vary with the ideal interval size, at least within a given person and domain. Early work on overconfidence suggested that people were more overconfident when they knew less, which would imply larger errors with larger intervals. However, recent research has shown that this “hard-easy effect” is mainly an artifact of how the hard and easy questions are selected (e.g., Dawes & Mumford, 1996; Gigerenzer et al., 1991; Klayman et al., 1999). Thus, the assumption that error is unrelated to well-calibrated interval width is reasonable. Given this independence assumption, plus the fact that \bar{z} is a constant, Eq. A2 simplifies to

$$\text{cov}(W_i, H_i) = \text{cov}(v_i W_i^*, H_i)$$

The next two steps are based on a standard definition of covariance, $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$, or equivalently, $E(XY) = E(X)E(Y) + \text{cov}(X, Y)$. Continuing, then,

$$\text{cov}(W_i, H_i) = \text{cov}(v_i W_i^*, H_i) = E(v_i W_i^* H_i) - E(H_i)E(v_i W_i^*)$$

$$(A3) \quad \text{cov}(W_i, H_i) = E(W_i^*)E(v_i H_i) + \text{cov}(W_i^*, v_i H_i) - E(H_i)\{E(v_i)E(W_i^*) + \text{cov}(W_i^*, v_i)\}.$$

By the definition of v_i , $E(v_i) = 0$, and the independence assumption implies that $\text{cov}(W_i^*, v_i) = 0$ and $\text{cov}(W_i^*, v_i H_i) = 0$, so Eq. A3 simplifies to

$$\begin{aligned} \text{cov}(W_i, H_i) &= E(W_i^*)E(H_i v_i) \\ &= E(W_i^*)\{E(H_i)E(v_i) + \text{cov}(H_i, v_i)\} \end{aligned}$$

$$(A4) \quad = E(W_i^*)\text{cov}(H_i, v_i). \quad [\text{because } E(v_i) = 0]$$

From Equation A1, it is easy to show that $E(W_i) = \bar{z} E(W_i^*)$. Therefore, we can rewrite Equation A2 as

$$(A5) \quad \frac{\text{Cov}(W_i, H_i)}{E(W_i)} = \frac{\text{Cov}(H_i, v_i)}{\bar{z}}$$

The quantities on the left side of Equation A5 can be estimated from observable behaviors. The numerator is simply the covariance between interval width and the binary variable that indicates hits and misses. The denominator is estimated by mean interval width.

We used Monte Carlo simulations to produce a variety of distributions of interval sizes, using a variety of different gamma functions for $(\bar{z} + v_i)$. Bounded at zero, gamma distributions have two parameters, shape and scale, that uniquely determine the mean (\bar{z}) and the standard deviation around it, \mathbf{s}_v . We then scored simulated hits and misses by matching each simulated interval to a standard normal distribution of correct answers. Luckily, different combinations of \bar{z} and \mathbf{s}_v produce unique combinations of hit rate and covariance that can be matched to observed results.

Log model

The logic for our alternative model is the same, but starting from a different assumption about how actual and ideal interval sizes are related, as shown in Equation 2 of the text, namely $\log(W_i / W_i^*) = z$. Again, we partition z into its mean, \bar{z} , and mean-zero variation around it, v_i . So,

$$\log(W_i / W_i^*) = \bar{z} + v_i$$

$$\log W_i = \log W_i^* + \bar{z} + v_i$$

$$\begin{aligned} \text{cov}(\log W_i, H_i) &= \text{cov}([\log W_i^* + \bar{z} + v_i], H_i) \\ &= \text{cov}(\log W_i^*, H_i) + \text{cov}(\bar{z}, H_i) + \text{cov}(v_i, H_i). \end{aligned}$$

As we discussed earlier, $\text{cov}(W_i^*, H_i) = 0$, so logically $\text{cov}(\log W_i^*, H_i) = 0$, too; $\text{cov}(\bar{z}, H_i) = 0$ because \bar{z} is a constant. Thus,

$$(A6) \quad \text{cov}(\log W_i, H_i) = \text{cov}(v_i, H_i).$$

This time, our Monte Carlo simulations used normal distributions for $(\bar{z} + v_i)$, with different means and standard deviations. We then matched the resulting distribution of intervals to a standard normal distribution of correct answers to calculate hits and misses and the covariance. Again, different combinations of \bar{z} and \mathbf{s}_v produce unique combinations of hit rate and covariance that can be matched to the observed hit rate and the observed covariance between log-interval-widths and hits.

REFERENCES

- Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior & Human Decision Processes*, *49*, 188-207.
- Budescu, D. V., Erev, I., Wallsten, T. S., & Yates, J. F., Eds. (1997). Special issue: Stochastic and cognitive models of confidence. *Journal of Behavioral Decision Making*, *10*, 153-285.
- Clemen, R. T. (2001). Assessing 10-50-90s: A surprise. *Decision Analysis Newsletter*, *20(1)*, (April), 2, 15.
- Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior & Human Decision Processes*, *65*, 201-211.
- 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.
- Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, vol. 107, pp. 659- 676.
- Gigerenzer, G., Hoffrage, U. & Kleinb lting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.
- Hoch, S. J. (1985) Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 719-731.
- Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, *5*, 55-71.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*, 226-246
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *10*, 384-396.
- Juslin, P., Wennerholm, P., & Olsson, H. (2000). Format-dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from a cognitive perspective*. New York: Academic Press (*Psychology of Learning and Motivation*, vol. 32), pp. 365-418.
- Klayman, J., Soll, J., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216-247.
- Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107-118.
- May, R. S. (1986). Inferences, subjective probability, and frequency of correct answers: a cognitive approach to the overconfidence phenomenon, in Brehmer, B., Jungermann, B., Lourens, P. and Sevon, G. (eds), *New Directions in Research on Decision Making*, Amsterdam: North-Holland.
- McClelland, A. G. R. & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1994. In Wright, G. & Ayton, P. (eds.), *Subjective Probability* (pp. 453-482). Chichester: Wiley.

- McKenzie, C. R. M. (1999). (Non)Complementary updating of belief in two hypotheses. *Memory and Cognition*, 27, 152-165.
- O'Connor, M. & Lawrence, M. (1989). An examination of the accuracy of judgment confidence intervals in time series forecasting. *International Journal of Forecasting*, 8, 141-155.
- Russo, J. E. & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, 33, 7-17.
- Selvidge, J.E. (1980). Assessing the extremes of probability distributions by the fractile method. *Decision Sciences*, 11, 493-502.
- 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.
- Suantak, L., Bolger, F. & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, 67, 201-221.
- Yaniv, I. & Foster, D. (1997). Precision and Accuracy of Judgmental Estimation. *Journal of Behavioral Decision Making*, 10, 21-32