Organizational Behavior and Human Decision Processes
Vol. 79, No. 3, September, pp. 216–247, 1999
Article ID obhd.1999.2847, available online at http://www.idealibrary.com on IDEAL®

Overconfidence: It Depends on How, What, and Whom You Ask

Joshua Klayman

University of Chicago

Jack B. Soll

INSEAD

Claudia González-Vallejo

Ohio University

and

Sema Barlas

Experian Direct Tech

Many studies have reported that the confidence people have in their judgments exceeds their accuracy and that overconfidence increases with the difficulty of the task. However, some common analyses confound systematic psychological effects with statistical effects that are inevitable if judgments are imperfect. We present three experiments using new methods to separate systematic effects from the statistically inevitable. We still find systematic differences between confidence and accuracy, including an overall bias toward overconfidence. However, these effects vary greatly with the type of judgment. There is little general overconfidence with two-choice questions and pronounced overconfidence with subjective confidence intervals. Over- and underconfidence also vary systematically with the domain of questions asked, but not as a function of difficulty. We also find stable individual differences. Determining why some

This research was supported by Grant SBR-9409627 from the Decision, Risk, and Management Sciences Program of the National Science Foundation. We thank Peter Juslin, Eldar Shafir, David Budescu, Robin Hogarth, Christopher Hsee, J. Edward Russo, William Ferrell, Terry Connolly, and an anonymous reviewer for their helpful comments on earlier drafts.

Address correspondence and reprint requests to Joshua Klayman, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637. E-mail:joshk@uchicago.edu.

people, some domains, and some types of judgments are more prone to overconfidence will be important to understanding how confidence judgments are made.

A variety of scientists, including meteorologists, statisticians, and psychologists, have been interested in measuring and explaining judgments of confidence and their relation to accuracy (e.g., see Budescu, Erev, Wallsten, & Yates, 1997; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Harvey, 1997; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994; Yates, 1990). Many of these studies report that people are systematically overconfident about the accuracy of their knowledge and judgment. That is, they tend to express confidence in their judgments that exceeds the accuracy of those judgments. At the same time, the extent to which overconfidence occurs seems to depend very much on the difficulty of the judgment task. With easy tasks, overconfidence seems to disappear, or underconfidence is observed. With hard tasks, overconfidence seems to be rampant.

Recently, a number of researchers have challenged the validity of both the conclusion that people are systematically overconfident and the finding that task difficulty affects the level of over- or underconfidence. In this paper, we present the results of three studies that take into account legitimate objections raised about past methods of research on confidence. We present new analytical methods that can be used to distinguish different sources of error and accuracy in confidence. Our findings confirm that there are systematic differences between subjective confidence judgments and observed accuracy. The more confident people are, the more overconfident they are, and, overall, confidence tends to exceed accuracy. These effects result from a combination of unsystematic imperfections in judgment and systematic effects of cognitive processes. Beyond this generalization, we find that there are systematic individual differences in over- or underconfidence as well as systematic differences between domains of questions—differences that are not a function of difficulty. We also find that results vary greatly with the way in which confidence judgments are elicited. Questions that require a choice between two alternatives elicit only a modest bias toward overconfidence; questions that request a subjective confidence interval elicit a very large bias. The inability of recent research to determine any simple explanation for the phenomena of confidence judgments probably reflects that they are multiply determined, and that some of the standard methods are not well suited to distinguishing different aspects of the process.

The Overconfidence Phenomenon

The most often used stimuli in studies of confidence are sets of two-choice questions, such as “Which of these nations has higher life expectancy, averaged across men and women: (A) Argentina, or (B) Canada?”[1] Participants may answer anywhere from 20 to 300 such questions. For each question, participants

[1] Argentina 71, Canada 77.

choose the answer they think is more likely to be right and indicate, on a scale from 50% to 100%, how sure they are that they have chosen correctly. Another method that has been used in several studies is to ask participants about a single, numerical estimate such as "How many calories are there in 1/2-cup of bread pudding?"[2] Participants might, for example, be asked to estimate fractiles (e.g., stating a value for which there is a 25% chance that the correct answer is higher) or to provide a range corresponding to a given level of confidence (e.g., stating a high and a low estimate such that there is a 90% chance that the correct answer falls somewhere between those numbers).

Two-choice and confidence-range judgments both have important analogues in the real world. We are often called upon to make a choice between two alternatives, and then our subjective confidence in that choice determines how much we commit to one course, how much we seek further information, and how much we hedge our bets. Confidence-range estimates may seem less natural, but they also have everyday counterparts. Confidence ranges are implied whenever one plans a "margin of error." In budgeting, for example, the goal is to allocate an amount that is sufficiently certain to cover needs, yet not so much that resources are tied up unnecessarily. Research by Yaniv and Foster (1995) also indicates that people communicate something akin to a confidence range in natural language by varying the "grain" of an estimate. For example, someone who did not know the exact price of an item might report a best guess of "a few hundred bucks" or "around $350" or "$347.50."

From the early 1970s until the early 1990s, there was a general consensus that judges in both two-choice and confidence-range tasks showed consistent and substantial overconfidence. In a review of previous two-choice studies, for example, Lichtenstein et al. (1982) report that when participants say they are about 70% sure they have the correct answer, they are right less than 60% of the time; when they say they are 90% sure, they are right about 75% of the time. Overconfidence with range questions may be even more extreme (Lichtenstein et al., 1982). Russo and Schoemaker (1992), for example, found that business managers asked to provide 90% confidence ranges had the correct answer within the stated range between 42 and 62% of the time, depending on the domain and the participant group. Fifty-percent ranges contained the correct answer about 20% of the time.

Studies also reveal a significant difference between easy and hard questions. Hard questions are typically defined as those for which many participants guess the wrong answer, and a hard *set* of questions is defined as one for which the average participant has a low percentage correct. It is widely found that overconfidence is more pronounced for harder sets of questions and for harder questions within a set. For easy questions, judges may even be underconfident.

Why Are People Generally Overconfident?

Two main categories of explanations have been offered for overconfidence. These are (a) biases in information processing and (b) effects of unbiased

[2] Approximately 200 calories.

judgmental error. (Although they have sometimes been treated as competing explanations, both may be true.) Most early investigators attributed overconfidence to information-search strategies and motivation. They hypothesized that the judge first searches memory for relevant information and arrives at a tentative answer. Then, with this answer in mind, the judge searches for more evidence. Mechanisms of associative memory facilitate retrieval of information that is consistent with initial impressions, and those impressions also color the interpretation of subsequent ambiguous evidence. Judges, however, believe their processes to be unbiased, and thus perceive more consistent support for the initial guess than is warranted (e.g., Hoch, 1985; Klayman, 1995; Koriat, Lichtenstein, & Fischhoff, 1980). In many situations, motivational factors can exacerbate the bias. People like to think that they are intelligent and knowledgeable, and they may have reasons for wanting a particular answer to be true (e.g., Babad, 1987; Kunda, 1990; Langer, 1975; Larrick, 1993).

Another class of explanations highlights the role of unbiased judgmental error in producing overconfidence. Possible sources of error include imperfections in learning the predictive validity of different sources of information (Gigerenzer et al., 1991; Soll, 1996), in evaluating the available information (Erev, Wallsten, & Budescu, 1994), and in mapping one's subjective feeling of confidence to a response scale (Erev et al., 1994; Ferrell, 1994). Both accuracy and confidence are affected by random variation. With regard to accuracy, sometimes even good-quality information can point in the wrong direction; how often that happens is partly a matter of chance. With regard to confidence, people's judgments about the quality of their information include some unsystematic error. Given an imperfect correlation between accuracy and confidence, it is inevitable that low accuracy is on average associated with not-so-low confidence, and so on. This produces the typical pattern of "miscalibration": overconfidence when confidence is high, underconfidence when confidence is low (see Fig. 1, for example). It also leads to an effect of difficulty: overconfidence for those questions that show low accuracy (hard questions), underconfidence when accuracy is high (easy questions). People may also make errors in estimating what the mean level of accuracy will be for a whole set of questions, so harder *sets* of questions are more likely to also be those that are harder than they seem. This results in more overconfidence on harder sets of questions (see Ferrell, 1994; Ferrell & McGoey, 1980; Suantak, Bolger, & Ferrell, 1996). Note that none of these explanations assumes that judgments are systematically biased, only that they are imperfect (see also Harvey, 1997; Soll, 1996).

It has also been suggested that the apparent predominance of overconfidence in research stems not from pervasive cognitive bias, but from experimenters' tendency to choose harder-than-normal questions (Gigerenzer et al., 1991; Juslin, 1993, 1994; May, 1986). To test this possibility, Gigerenzer et al. (1991) and Juslin (1993, 1994) conducted experiments using two-choice questions that were randomly sampled from a domain, thus approximating the natural level of difficulty for questions in that domain. They found that overconfidence (but not miscalibration) disappeared. These results are intriguing, but the question

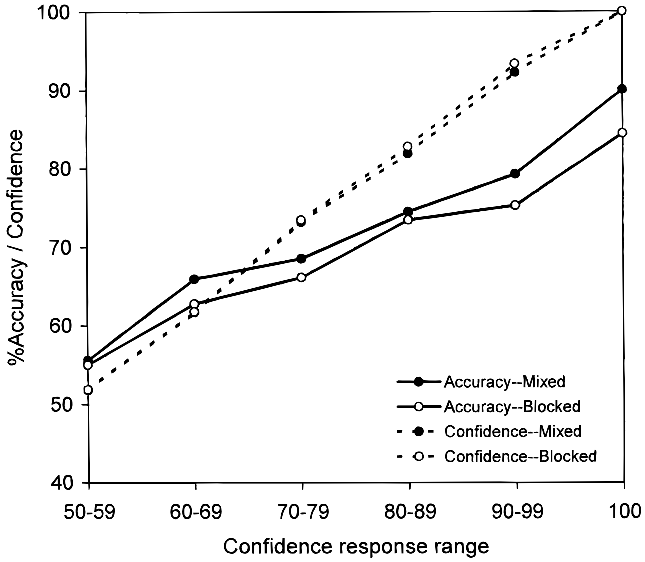

FIG. 1. Calibration curves for two-choice questions from the mixed and blocked conditions of Experiment 1.

of whether experimenter selection is the sole cause of the apparent overconfidence bias remains open. There are many examples of studies that demonstrated overconfidence despite the fact that questions were seemingly not selected for difficulty (Brenner, Koehler, Liberman & Tversky, 1996; Budescu, Wallsten, & Au, 1997; Dawes, 1980; Dunning, Griffin, Milojkovic, & Ross, 1990; Fischhoff & MacGregor, 1982; Griffin & Tversky, 1992; Lichtenstein & Fischhoff, 1977; Paese & Sneizek, 1991; Peterson & Pitz, 1988). Also, it remains controversial whether the observed correlation between difficulty and overconfidence can be fully explained by the effects of unbiased imperfections in judgment (see Juslin, Olsson, & Bjorkman, 1997). Nevertheless, the point is well taken that selective sampling has the potential to make appropriately confident participants appear overconfident.

The Experiments

We present three experiments using methods and analyses designed to distinguish systematic judgmental biases from the effects of sampling error and unsystematic imperfections in judgment. One feature of these methods is random sampling of questions from multiple domains. We concur with Gigerenzer and others that random sampling is the best practical way to produce sets of questions that are, on average, unbiased representatives of the populations of questions from which they are drawn. Using multiple domains permits examination of domain-to-domain differences and helps ensure generalizability. Another feature of our methods is the use of split-sample analyses, that is, analyses that compare performance on one set of questions to performance on another, independent set. This method avoids the problems associated with sorting questions according to performance measures (see the following section for details).

In the three studies presented here we look for (a) overconfidence and miscalibration; (b) variations in overconfidence from domain to domain, in particular as a function of difficulty; and (c) individual differences in the accuracy of confidence judgments. In addition, we test whether calibration and overconfidence are affected by whether participants make repeated judgments of the same type in the same domain or a variety of different judgments on different topics. Many recent studies have used a procedure in which participants are asked questions from only a single domain of knowledge. This contrasts with earlier experiments which usually presented a mix of questions on a wide variety of topics.[3] It is possible that larger and more homogeneous sets of questions are less prone to biased information processing. If judges have to answer hundreds of similar questions, they may settle on a consistent judgmental policy, using the same types of information and weighing evidence similarly from trial to trial. Doing so might reduce the tendency to favor information consistent with an initial impression, which has been hypothesized to be a major source of bias. Sources of unsystematic error in judgments might also be reduced with repeated practice in making the same judgments. Thus, large, homogeneous sets of questions might be expected to show less overconfidence than heterogeneous sets. The first two experiments use two-choice questions with confidence judgments expressed on a scale from 50% to 100%. The third experiment extends the use of representative question sets into another kind of judgmental task for which systematic overconfidence has also been demonstrated, namely estimates of subjective confidence ranges.

APPROPRIATE METHODS OF MEASURING (OVER)CONFIDENCE AND DIFFICULTY

Before presenting the studies, we describe a model of the role of unsystematic error in confidence judgments, integrating many of the constructs proposed by researchers over the last decade. Using this model, we explain how our split-sample analyses can be used to separate the effects of systematic and unsystematic error. We argue that the crux of the analytical problem is as follows. There are always some questions for which the answer runs counter to the judges' information; these are the questions judges get wrong. We call these *contrary* questions.[4] Hard questions are those that are contrary for many judges (or, alternatively, a set of which many are contrary for a given judge). Some questions are hard because judges have little useful information to go on. They are

[3] Gigerenzer et al. (1991) used both mixed- and single-domain question sets and found no difference. However, the two sets were not randomly selected. Rather, they were selected to be difficult, and equally so. Also, only one domain was tested in a single-domain presentation. Thus, the Gigerenzer et al. study does not afford a direct comparison of mixed- and single-domain procedures.

[4] They have also been referred to as "misleading" (May, 1986) and as "deceptive" (Fischhoff, Slovic, & Lichtenstein, 1977). We avoid these terms because they imply some special features that fool people; in fact, contrary questions merely fail to conform to one's prediction.

mostly guessing, and in nearly half the cases, the answer will run contrary to whatever slight inclination their information suggests. The better the available information, the smaller the chance of a contrary question, and we might hope that judges tune their confidence to the strength of the available information accordingly. However, there are always some contrary questions, regardless of how good the judges' information is (assuming it is less than perfect). Thus, a question may be hard not because judges lack information, but because the answer happens to run contrary to some usually good sources of information that many judges use. In such cases, reasonable judges will be both wrong and confident. For these hard questions, then, overconfidence is inevitable even if judges are right about how accurate their information usually is. The better the judges' information, the fewer contrary questions there will be, but the more confidence they will have in their wrong answers.

Researchers are interested in how well judges estimate the strength of their information, not in whether they can tell which particular questions are contrary. The latter is impossible; if judges could tell then the questions would not be contrary. It is therefore necessary to devise analytical methods that measure the ability to judge information strength independent of the inability to recognize contrary questions.

To explain the potential problems in measuring overconfidence and its relation to difficulty, we use a general model of confidence judgments in the two-alternative task. This model incorporates conceptual elements common to several more specific models that have been recently proposed (e.g., Erev et al., 1994; Ferrell, 1994; Ferrell & McGoey, 1980; Gigerenzer et al., 1991; Juslin, 1993, 1994; Soll, 1996):

(a) To answer a question, the judge retrieves and weighs some information, which produces an internal stimulus signal that favors one or the other alternative with some strength.[5]

(b) The judge chooses an answer according to the direction or sign of that signal.

(c) The stronger the internal signal, the more likely is a correct answer, on average.

(d) The judge makes a confidence judgment based on the subjective strength of the signal.

(e) The stronger the internal signal, the higher is the expressed confidence, on average.

(f) The judge attempts to match the confidence judgment to the probability of a correct answer given the signal strength, but does so with some error (systematic error, random error, or both).

Suppose you are interested in the performance of a judge or group of judges on some population of questions, Q. Q could be a domain, such as distances between cities, or it might be broader, such as almanac questions in general. In principle, you could measure the average confidence, C_{Q}, and the proportion

[5] In the models of Ferrell and colleagues, what we refer to as a signal is modeled as the separation between two signals, one for each alternative.

correct, P_Q, across the entire population of questions. These are each a function of the overall information strength of this population of questions, I_Q. In other words, $C_Q = C(I_Q)$ and $P_Q = P(I_Q)$. Overconfidence in the population is the difference,

$$O_Q = C(I_Q) - P(I_Q). \tag{1}$$

(A negative O_Q represents underconfidence.)

In practice, though, one usually tests only a subset of questions, S. The selected questions will have an average signal strength of I_S. The proportion of correct answers obtained in response to the subset of questions is $P_S = P(I_S) + dP_S$. Performance on the sample of course depends on the strength of information available: $P(I_S)$ is the proportion of correct answers among the population of all questions or all sets of questions with information strength I_S. The added term, dP_S, is there because the proportion of contrary and noncontrary questions in subset S will vary some from the expected proportions given signal strength I_S, just by luck in drawing particular questions. (The concept of "representative sample" presented by Gigerenzer et al., 1991, corresponds to a sample in which $dP_S = 0$.) Similarly, the average confidence for the subset of questions will be $C_S = C(I_S)$. (Most, but not all, models also assume some unsystematic variation in confidence judgments. This would not affect our analysis, so we omit it for simplicity.) Overconfidence in the subset, then, is

$$O_S = C_S - P_S = C(I_S) - P(I_S) - dP_S. \tag{2}$$

How will overconfidence in the subset compare to overconfidence in the overall population? Combining Eqs. 1 and 2,

$$\begin{aligned} O_S - O_Q &= [C(I_S) - P(I_S) - dP_S] - [C(I_Q) - P(I_Q)] \\ &= [C(I_S) - C(I_Q)] - [P(I_S) - P(I_Q)] - dP_S. \end{aligned} \tag{3}$$

If you generate your subsample of questions by randomly sampling from all questions in the population (as suggested by Gigerenzer et al., 1991, and Juslin, 1993), then overconfidence in subset and population should be the same, on average. Information strength, I_S, will average the same as the population strength, I_Q, and luck in drawing contrary and noncontrary questions into the sample, dP_S, should average out to about zero.

Sometimes, though, researchers have been interested in sets of questions that do not match the whole population. In particular, they have been interested in comparing hard and easy questions to test the hypothesis that, in our terms, confidence errors vary with signal strength. Signal strength per se is not an observable quantity, so harder questions are typically defined as those for which the proportion correct (P) is low. A set of questions may be harder than the overall population for two reasons: The average signal strength for these questions is lower, or there are more contrary questions in the sample than one might expect given the signal strengths, or both. Suppose, for example, that an experimenter selects a question that is correctly answered by 55% of participants in a domain in which the average percentage correct is 70. Performance may be poor because the informational signal is weak for most participants ($I_S < I_Q$). They have little information to draw upon, or the available information does not clearly distinguish the answers. Or performance may be poor because this question happens to be a contrary one for many participants ($dP_S < 0$). The answer runs counter to some usually good information that many participants use.

Whether overconfidence varies with difficulty depends on how judges respond to these two aspects of difficulty. The question of primary interest to researchers is whether judges lower their confidence appropriately when there is less valid information (lower signal strength). In our terms, does $[C(I_S) - C(I_Q)] = [P(I_S) - P(I_Q)]$? Perhaps it does, on average. Alternatively, in accord with earlier hypotheses, judges may lower their confidence insufficiently in response to lower signal strength. In that case, overconfidence will increase as signals get weaker (see Eq. 3). The other element that makes a set of questions hard is if there are more contrary questions in it than expected given the signal strengths. Do judges lower their confidence appropriately in response to that? They cannot. If there are more contrary questions than expected given the signal strengths ($dP_S < 0$), judges have no way of detecting that. So, to the extent that there are more contrary questions than expected, the proportion of correct answers will be lower, but confidence will not. This kind of difficulty effect is not a psychological phenomenon, but a statistical inevitability.

How could one test for the first, cognitive type of difficulty effect, without confounding it with the second, statistical type? One might suppose that these problems could be avoided by defining difficulty independently of observed proportion correct. For example, one could measure how close together the two answers are on the dimension in question (e.g., the absolute difference between the two city populations for a "Which city is larger . . ." question). Call this difference measure Δ. The measure Δ is presumably correlated with signal strength, because questions whose alternatives are far apart are more likely to provide strong directional signals, on average. Unfortunately, as Juslin, Olsson, and Winman (1998) demonstrate, selecting questions with high or low Δ also selects for low and high proportions of contrary questions, respectively. So, questions with lower than average Δ have lower than average signal strengths, but they also have a lower than average proportion correct given the signal strengths (in our terms, $dP_S < 0$). Intuitively, consider that when underdogs win elections, the outcome is usually close; underdogs seldom win by a landslide. The contrary questions (upset victories) are almost always low Δ (close races), regardless of how sure the outcome seemed ahead of time (signal strength). (See Juslin et al., 1998, for a computer simulation of this effect.)

A different approach is to select large, random samples of questions from different domains and then compare samples with different proportions of correct answers (e.g., see Juslin et al., 1997). Sampling from different domains should produce samples that differ widely in the strength of available information. Selecting large, random samples from each domain should minimize the effect of accidental over- or undersampling of contrary questions. Thus, differences in overconfidence from sample to sample should be influenced mostly by the relationship between confidence and signal strength, and not much by accidents of sampling. Nevertheless, if participants do lower their confidence appropriately, on average, in response to weaker signals, accidental variations in the proportion of contrary questions could still produce a statistical difficulty effect. Accuracy will be negatively correlated with confidence-minus-accuracy (i.e., overconfidence) as long as both are measured on the same sample of questions (Juslin et al., 1997, 1998). This could explain the finding of Juslin et al. (1997) that there is a strong correlation between domain difficulty and overconfidence, but the actual magnitude of the effect is small.

The Split-Sample Method

To avoid the difficulties highlighted above, we use a method in which we select two separate samples of questions from each of a number of different domains. We measure domain difficulty on one sample of questions, and we measure over- or underconfidence on the other sample from that domain. If the two samples are independent and random, then accidents of sampling in the first should not affect performance on the second. Thus, tests of differences between domains are not confounded with effects of sampling errors.

Each domain of questions, Q, has an associated strength of information, I_Q. If S1 and S2 are two independent, random samples of questions from Domain Q, their information strengths will differ some from the whole population of Q, just by luck of the draw. Thus, $I_{S1} = I_Q + eI_{S1}$ and $I_{S2} = I_Q + eI_{S2}$. The proportion of correct answers for questions in S1 depends on the signal strengths in S1, plus some variation due to accidentally sampling more or fewer contrary questions than expected given the signal strength (see Eq. 2). That is,

$$P_{S1} = P(I_Q + eI_{S1}) + dP_{S1}. \tag{4}$$

Next, consider the second sample from this domain, S2. Confidence in this sample is $C_{S2} = C(I_Q + eI_{S2})$ and the proportion correct is $P_{S2} = P(I_Q + eI_{S2}) + dP_{S2}$ (see Eqs. 2 and 3). Subtracting the latter from the former, we get the overconfidence for sample S2:

$$O_{S2} = C(I_Q + eI_{S2}) - P(I_Q + eI_{S2}) - dP_{S2}. \tag{5}$$

Now, suppose we compare one domain to another, and observe that domains with lower proportion correct in one subsample (as in Eq. 4) tend to show higher overconfidence in the other sample from the same domain (as in Eq. 5). The eI and dP elements represent random variations due to accidents of sampling; none is correlated with anything else. The only thing that can account for a correlation from domain to domain between P_{S1} and O_{S2} is a correlation between $P(I_Q)$ and $C(I_Q) - P(I_Q)$, that is, a “real” difficulty effect. The split-sample technique can similarly provide tests of individual differences and domain differences that are unconfounded with effects of sampling and selection.

EXPERIMENT 1

Method

Participants. The participants were 32 University of Chicago students solicited by posted advertisements. They were paid $6 for participating. Sessions were self-paced, and required about 30 min to complete.

Materials. A total of 480 two-choice questions were prepared for presentation on a computer monitor. The question set included 40 paired comparisons from each of 12 domains (see Table 1). Each question asked the participant to

TABLE 1

Domains of Questions Used in Experiment 1, with Mean Performance Measures

Domain of questions	Proportion correct	Confidence	Over-confidence[a]
Which of these American museums or galleries had more visitors in 1991?	.541	.670	.130
Which of these 1992-model cars gets more miles-per-gallon in real driving (mix of highway and city)?	.613	.662	.050
Which of these "tourist cities" has a warmer daily high temperature in July, on average?	.637	.762	.124
Which of these U.S. colleges or universities charged higher tuition in 1991? (For state colleges, use the in-state-resident tuition.)	.656	.689	.033
Which of these food items has more calories?	.672	.795	.123
Which brand of shampoo costs more per ounce (national average)?	.684	.715	.031
Of these two "principal mountains of the world," which is taller?	.688	.615	−.073
Which of these states had a higher percentage of its population with incomes below the federal poverty line in 1990?	.697	.712	.015
Which of these cities is farther from Los Angeles, in air miles?	.728	.801	.073
Which of these states had a higher population in 1990?	.750	.799	.049
Which of these nations has higher life expectancy, averaged across men and women?	.791	.774	−.017
Which of these U.S. presidents held office first?	.856	.869	.012
Total	.693	.739	.046

Note. Facts used are from 1990–1992 because the earliest experiment was begun in 1993.

[a] Overconfidence = confidence − proportion correct. Negative numbers indicate underconfidence.

make an ordinal comparison of two items.[6] Representative sampling of questions from a domain was approximated by random sampling from the population of all possible questions in a domain. To do that, we selected 12 domains for which comprehensive lists were available. These lists either included all possible members (e.g., poverty levels of U. S. states or inauguration dates of U.S. presidents) or provided subsets that were not likely to be biased in favor of unusual or surprising items (e.g., the heights of "principal mountains of the world" or the listing of prices for all shampoos reviewed in a consumer magazine). Once an appropriate list was obtained for each domain, items were sampled with replacement using a random number table, and were paired consecutively. Pairings were replaced only when they duplicated an earlier question.

Questions were presented one at a time on a video monitor, with the two choices labeled as (A) and (B). The order of choices was randomly determined on each presentation. The question was followed by a response line with the prompt "(A) or (B)?" When one of those two letters was typed, a second response line appeared, with the prompt "Chance correct, 50–100." After typing a number in that range, the participant could either confirm the responses, in which case the next question appeared, or request to make a change.

Design. Participants were assigned alternately to one of two conditions (16 in each). In the *blocked* condition, each participant received 40 questions in a row from each of three domains (e.g., 40 questions about the sequence of presidents, followed by 40 questions about the price of shampoo, followed by 40 questions on life expectancies in various countries). Domains and individual questions were assigned to participants so that each of the 480 questions was presented once in each group of 4 participants. The order of the questions in each block of 40 was randomized individually for each participant. In the *mixed* condition, each participant received 10 questions from each of the 12 domains, mixed together in random order. As in the blocked condition, questions were assigned to participants so that each question was shown once in each group of 4 participants.

Procedure. Following the presentation of written consent forms containing a brief summary of the procedure, the presentation of instructions and stimuli was controlled by computer. Instruction screens told participants that they would be answering 120 two-choice questions that would "vary in difficulty from obvious to obscure, with most somewhere in between." The following example was given: "Who is older, (A) Bill Clinton or (B) Madonna?" This was followed by instructions on how to enter their answers and how they could change their answers prior to committing to them. Once participants confirmed their answers to a given question, they could not return to that question. The following instructions were provided for the confidence scale:

[6] In our usage, an *item* is one of the members of the list for a given domain (e.g., the poverty level of Vermont), as distinct from a *question* posed to participants (e.g., "Which of these states . . ."). Thus, two-choice questions require a comparison of two items.

> In addition to choosing answer A or B, we ask you to also indicate your confidence by telling us what you think the chance is you are right. If you are absolutely in the dark, no better than flipping a coin, that would correspond to a 50% chance of getting it right. If you are absolutely certain you have the answer, so you think the odds are less than one in a hundred you might be wrong, then that's a 100% chance your answer is right.
>
> Please feel free to use any number between 50 and 100 to indicate what you think the chance is that your answer is right. Numbers less than 50 are not allowed, because if you think there's LESS than a 50/50 chance your answer is right, you ought to choose the other answer!

Participants in the blocked condition were told to expect "40 questions on each of three different topics"; in the mixed condition, "120 questions on a wide variety of different topics."

Results and Discussion

Overall calibration and confidence. First, we examined calibration using the traditional technique of plotting actual percent correct contingent on participants' stated level of confidence. In plotting such calibration curves, it is necessary to group participants' responses into categories, in order to obtain a reasonable sample of responses from which to calculate the actual percent correct. To facilitate comparison to past results, we used the most common method, grouping together responses in the ranges of .50–.59, .60–.69, .70–.79, .80–.89, .90–.99, and 1.0. Figure 1 shows the calibration curves obtained for the mixed and blocked conditions.

The results in Fig. 1 replicate earlier calibration findings, in that (a) the overall calibration curve is flatter than the ideal line, matching average confidence to percent correct, and (b) it lies almost completely below the ideal line, indicating a predominance of overconfident judgments. There is little difference between the mixed and blocked conditions; if anything, the participants in the blocked condition may have done worse, contrary to our prediction.

Data like these can be formally analyzed using components of the Brier score, chiefly the calibration and resolution components (Lichtenstein & Fischhoff, 1977; Murphy, 1973). The calibration component of the Brier score measures the degree to which participants' confidence judgments match the obtained percentage correct across different segments of the confidence scale. Resolution is the degree to which participants' confidence judgments differentiate between correct and incorrect answers. The mixed and blocked conditions did not differ significantly on either of these measures.

We also performed an ANOVA with condition as a between-participants variable and first half (i.e., first 60 questions) vs second half as a within-participants variable, with objective percentage correct and subjective confidence as repeated measures. (In these and subsequent statistical tests, both objective and subjective measures were converted to unbounded scales using a log-odds transform.) There was a significant difference between the objective and subjective measures, $F(1, 30) = 8.49$, $p < .007$, with a mean percentage correct of .69 and a mean confidence of .74. There was no significant main effect for condition, $F < 1$, and no significant interaction between condition and type of measure, $F < 1$. Thus, there was again no evidence that blocked and mixed administrations differed in real or perceived difficulty or in the degree of overconfidence. There were no significant effects involving first vs second half.

We also conducted a similar analysis with participants in the blocked condition, this time looking at the first, second, and third domain seen, and the first half vs second half of each domain. There was no effect of domain order, but a significant effect of domain half, $F(1, 15) = 6.38$, $p < .03$. Contrary to expectations, participants performed better in the first half of each domain than in the second (.70 correct vs .66). This is presumably an effect of boredom or fatigue. Participants were also slightly less confident in the second half of each block (.74 vs .73), producing no significant interactions between measure and block half. Thus, there is no evidence that overconfidence changed with practice in answering questions in a given domain.

A variety of analyses also show the expected difference between hard and easy questions. Accuracy on the questions from the six hardest domains averaged .63, with average overconfidence of .08, while accuracy on the six easiest domains averaged .75, with overconfidence of .01. There were differences in the same direction in overconfidence for harder and easier questions within domains as well. However, as discussed earlier, such effects could reflect the impossibility of detecting an excess of contrary questions, rather than an inability to appreciate the poverty of one's information. Thus, we will not report analyses of these measures in detail. Instead, we turn to analyses that compare independently sampled sets of questions.

Differences between hard and easy domains. We examined the effect of question difficulty on overconfidence by looking at judgments for easier and harder domains of questions. If we look at all questions and all participants combined for each domain, the traditional effect of difficulty shows up as a negative correlation between accuracy and overconfidence of −.52, $N = 12$, $p < .09$. However, as discussed earlier, accidental oversampling of contrary questions in one domain or undersampling in another might contribute to that correlation. To avoid this confound, we used split sampling, using one sample to define domain difficulty (the *definition set*, S1), and the other sample to measure overconfidence (the *measurement set*, S2). Specifically, we took the questions each participant saw from each domain, and divided them into subsample S1, consisting of the first, third, fifth, etc., question that that participant saw from each domain, and a complementary subsample S2, consisting of the second, fourth, sixth, etc. Because each of the questions in a given domain was separately sampled from the entire list of possibilities, S1 and S2 are independent. Because order of presentation was randomized, the composition of the subsamples is different for each participant.

For each participant in the blocked condition, we identified the hardest and easiest domains the participant saw, according to the accuracy of that participant on the definition set. In the mixed condition, we similarly identified the four hardest and four easiest domains. In case of ties, the number of domains in each category was adjusted to keep equally difficult domains together. For each participant, we then took two dependent variables from the measurement set: the differences in accuracy between domains designated as easy and hard, and the difference in overconfidence. The mean difference in accuracy was .10 (mean proportions correct were .73 and .63). This was significantly different from zero, $t(30) = 2.28$, $p = .01$. The mean difference in overconfidence was $-.04$ (overconfidence means were .03 and .07). This is in the direction of more overconfidence for harder domains, but is not significant, $t(30) = -1.15$, $p = .26$. We double-checked the results by reversing the roles of the definition and measurement sets. The difference between domains designated as easy and hard was .05, which was different from zero, $t(30) = 2.09$, $p < .05$. The difference for overconfidence was .03, this time in the direction of *less* overconfidence for harder domains, but was again not significant, $t(30) = 1.08$, $p = .29$. There were no significant differences between conditions on any of these measures.

The evidence regarding the effect of difficulty on overconfidence is clearly equivocal. However, the ability to analyze differences between domains is limited in the present study. In the mixed condition each participant saw only a small number of questions from each domain, and in the blocked condition, each domain was seen by only a few participants. Experiment 2 affords more powerful tests of difficulty effects, and we will return to the question of domain differences there.

Individual differences. Split-sample analyses can also be used to examine individual differences in confidence judgments. The logic of doing so parallels the rationale for using split samples to measure difficulty effects. By comparing measures taken on different subsamples, we avoid the possibility that apparent results are due only to accidents of sampling. For example, luck of the draw might give a participant a larger or smaller number of contrary questions in sample S1, and this would produce the appearance of a larger or smaller degree of overconfidence within that set of questions. Luck in subsample S1 is uncorrelated with luck in S2, however, so if overconfidence in S1 correlates with overconfidence in S2, that represents something more stable than luck.

For this analysis, we combined the data from all domains, and defined subsample S1 for each participant as those questions seen on odd-numbered trials, and subsample S2 as those seen on even-numbered trials. We then measured each participant's proportion correct, confidence, and overconfidence separately for S1 and S2. Table 2 shows the correlations between measures for the 32 participants on one subsample and measures for those participants on the other subsample. The results shown in Table 2 indicate that, indeed, some participants are consistently more overconfident than others, at least in the particular set of domains they were asked about. The correlation between overconfidence on one subsample and overconfidence on the other is .66. This corroborates findings of individual differences reported by Soll (1996) and Budescu, Wallsten, and Au (1997). Overall, 23 participants were overconfident, and 9 underconfident (binomial $p < .025$).

TABLE 2

Correlations across Participants between Measures Taken on Different Subsamples of Questions, S1 and S2, in Experiment 1

	Proportion correct: Sample S2 (M = .688, SD = .088	Confidence: Sample S2 (M = .735, SD = .072)	Overconfidence: Sample S2 (M = .048, SD = .102)
Proportion correct: Sample S1 (M = .698, SD = .085)	.619****	.269	−.343*
Confidence: Sample S1 (M = .742, SD = .072)	.190	.940****	.497***
Overconfidence: Sample S1 (M = .044, SD = .098) N = 32	−.394**	.456***	.660****

Note. For comparison, correlations between measures taken on the whole, unsplit sample were as follows: r(proportion correct, confidence) = .246, *ns*; r(proportion correct, overconfidence) = −.660, $p < .001$; r(confidence, overconfidence) = .566, $p < .001$.

* $p < .06$.

** $p < .03$.

*** $p < .01$.

**** $p < .001$.

Other correlations suggest how individual differences in overconfidence arise. Notably, there is a correlation of .94 between confidence on one subsample and confidence on the other: Clearly, some participants are consistently more confident than others. Differences in confidence could be justified: The correlation between accuracy on one set of questions and accuracy on the other is .62, indicating that some participants were consistently more accurate than others. However, there are only weak correlations between participants' confidence relative to that of other participants and their accuracy relative to that of others (.19 and .27). Even when we measure confidence and accuracy on all questions, without splitting the sample, the correlation is only .25. Individuals have an overall sense of confidence that is reliable, but not very accurate. Thus, more confident people tend to be more overconfident (correlations between confidence and overconfidence are .50 and .46), and less accurate people tend to be more overconfident (correlations between accuracy and overconfidence are −.39 and −.34).

The patterns of correlations observed in the mixed and blocked conditions were very similar, with one exception. The correlations between confidence and accuracy appear to be higher in the mixed condition (.50 and .41) than in the blocked condition (.10 and 0). The difference between conditions is marginally significant, but if real, it is likely due to the fact that participants make mistakes about how confident to be about particular domains, as well as overall. Mixed-condition participants see a greater variety of domains, so such domain-specific errors tend to average out. We examine the possibility of domain-specific errors in the following section.

Cross-domain comparisons. Another way to look at individual differences is to compare participants' overconfidence in one domain with their overconfidence in another domain. In the blocked condition, we identified the first, second, and third domains each participant saw. In the mixed condition, we identified the earliest four, middle four, and latest four domains seen, according to the average trial number of the questions in each domain. The sequential ordering is more meaningful in the blocked condition, but for these purposes, any arbitrary division of domains will suffice. We compared the overconfidence scores of the 32 participants on their first domain or set of domains with their overconfidence on the second. The correlation was .60. Between the first domains and the third the correlation was .58, and between the second and the third, .51. These are all significant at $p < .003$, $N = 32$. Similar results were obtained looking at each condition separately. All single-condition correlations were significant at $p < .06$, $N = 16$. These results confirm that some participants are generally more overconfident than others, and not just in particular domains.

Conclusions. It is clear from Fig. 1 and subsequent analyses that people's confidence judgments about two-choice questions are not well calibrated, even when tested in a large, varied, and representative sample of two-choice questions. Contrary to our expectations, the accuracy of confidence judgments is not materially affected by whether the participant receives three blocks of 40 questions from a single domain, or a mix of 120 questions from 12 domains.

The finding of miscalibrated judgments is not controversial: The same pattern is observed in Gigerenzer et al.'s (1991) study of German city sizes and in a variety of domains studied by Juslin and colleagues (see Juslin et al., 1997). People are not sure enough when they are very unsure, and they are too sure when they are very sure. We also find stable individual differences in overconfidence, such that more confident people tend to be more overconfident, as do less accurate people. These findings are consistent with models in which confidence judgments are imperfect, but unbiased. Miscalibration implies error in reading the accuracy of the information available in answering each question. The patterns of individual differences imply that errors occur not only in each individual confidence judgment, but also in setting one's overall level of confidence (see also Suantak et al., 1996). There is some evidence in this study that the errors people make are not completely unbiased, however. About 70% of individuals err in the direction of overconfidence. Overall, overconfidence is reliable but modest: between 4% and 5%. It seems likely that the tendency toward under- or overconfidence varies not only from individual to individual, but also from domain to domain. However, the present study provides little evidence that differences in difficulty account for differences in overconfidence between one domain and another.

EXPERIMENT 2

In this experiment we adopt the method used by Gigerenzer and others of asking each participant a long series of question in a single domain. Accordingly, domain is a between-participants variable. Although this design precludes some of the analyses performed in Experiment 1, the larger sample of questions from a given domain for each participant permits some more powerful comparisons of different domains. The single-domain procedure also allows a more powerful test of the effects of repeated questioning within one domain. The results of Experiment 1 suggest that the use of blocks of single-domain questions cannot explain previous findings of little or no overconfidence in representative sets of two-choice questions. Net overconfidence was statistically significant but modest, and mixed and blocked conditions did not differ. Nevertheless, we might still find a difference between mixed and blocked administrations if we used a stronger manipulation. Perhaps it takes more than 40 questions for a typical participant to settle on a consistent policy for retrieving and weighing information. Also, in most representative-sample studies, the questions are composed of different pairwise combinations of a smaller set of individual items. Repetition of the same items might also contribute to a more consistent use of information. Accordingly, we also constructed our question sets using repetitions of individual items in different pairwise combinations.

Method

Participants. The participants in this study were 54 University of Chicago students who had not participated in the earlier study. They were solicited and reimbursed as in Experiment 1. One participant was excluded and replaced because he repeated the same confidence value for nearly all of his answers.

Procedures. Each participant received a total of 150 questions presented in the same manner as in the previous study, except all from a single domain. A short break was provided at the halfway point. The 150 questions consisted of pairs composed from the 40 to 80 individual items used in each domain in the previous study.

Six domains were used, with nine participants assigned to each. The domains were shampoo prices, order of U.S. presidents, height of mountains, temperatures of cities, populations of U.S. states, and poverty levels of U.S. states. These were selected on the basis of being domains with small lists of items from which to draw (i.e., there are only 51 states/districts and 41 former presidents, and the published lists of shampoos, principal mountains, and major cities were limited in size). This was done to enhance the extent to which a sample of questions composed from a randomly selected subset of items would approximate representative sampling of the domain.

Results and Discussion

Comparing the two studies. We first checked our hypothesis that answering large numbers of questions from a single domain would improve judgments of confidence. To do this, we contrasted the present data with those of Experiment 1. We combined the two conditions of Experiment 1, and looked only at questions from the six domains that were also used in Experiment 2. (Two participants in the blocked condition were eliminated, because they had not received any questions from the eligible domains.) The resulting calibration curves are shown in Fig. 2. There seems to be no appreciable difference between the mixed or short-block procedures of Experiment 1 and the long-block procedure of the present study. Indeed, differences between experiments on Brier score components did not approach significance.

We conducted a MANOVA with experiment (1 or 2) as a between-participants variable and first half vs second half of trials as a within-participants variable, with objective accuracy and subjective confidence as repeated measures. There were no significant effects. The only effect reliable at $p < .20$ was a main effect of measure (i.e., overconfidence), $F(1, 82) = 2.05$. Overall overconfidence appears to be lower in Experiment 2 than in Experiment 1, with average overconfidence of 0.4% vs 3.0% on these domains. Also, these numbers are both lower than the average of 4.6% overconfidence obtained on the 12 domains in Experiment 1. However, neither of these differences is statistically reliable. We now turn to the data from only the 54 participants in the present study.

Domain differences. The present design allows us to make a more direct comparison between different domains. Domain can be included as a between-participants variable in a MANOVA, with accuracy and confidence as repeated measures. This analysis showed a main effect of domain, $F(5, 48) = 14.1$, $p < .001$, and an interaction between domain and measure, $F(5, 48) = 8.60$, $p < .001$. Together, these indicate that there were reliable differences between the domains in accuracy, confidence, and overconfidence. Were differences in overconfidence between domains a function of difficulty? The correlation between accuracy and overconfidence across the six domains is $-.08$, which is not

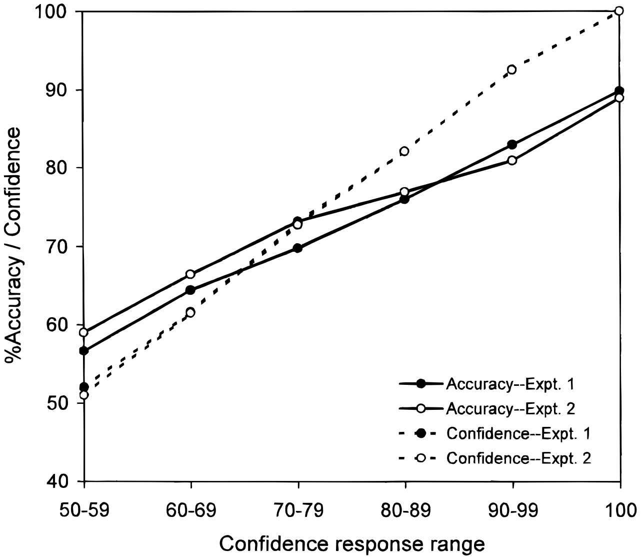

FIG. 2. Calibration curves for two-choice questions from the six domains used in both Experiment 1 and Experiment 2.

near significance. Of course, six domains is a small sample, but the differences between domains on overconfidence seem highly idiosyncratic (see Table 3).

The MANOVA indicates reliable differences among domains in proneness to overconfidence, but this could have resulted from accidents of sampling with regard to contrary questions in different domains. The magnitude of the observed differences suggests this is not the whole story, given the number of questions and participants sampled. We can check that using split samples, aggregating the data from different participants on each subset of questions in each domain. Given the design of the present experiment, it was necessary to split the questions in a different way than was done in the first study. Recall that in this study, the questions in each domain were formed by recombining the individual items (between 40 and 80 per domain) in different pairwise combinations. Therefore, simply splitting the sample into two halves by trial number would not yield completely independent samples. Accordingly, we used a simple computer algorithm (basically repeated trial and error) to identify two subsets of questions for each domain that shared no items in common. This resulted in two independent samples for each domain of roughly 50 questions each, with another 50 questions eliminated from the analysis. Although the elimination of questions reduces the power of our test, we did so in order to guarantee strictly independent samples of each domain for each participant. We used the same division of questions for all participants in a given domain. In doing this, there is a risk that idiosyncrasies of selection might produce differences in the two samples. However, it permits us to conduct analyses that combine data from multiple participants while preserving independence of samples, which is useful in examining differences between domains.

One additional modification was also necessary. Because there are different participants in each domain, we need to control for any effects of individual differences. To do this, we split not only the questions but also the participants. We separated the participants into odd and even according to the order in which they participated in the study. Thus, measures for each domain were taken on four subsamples: (a) Subsample S1 questions from odd-numbered participants, (b) Subsample S1 questions from even-numbered participants, (c) Subsample S2 questions from odd-numbered participants, and (d) Subsample S2 questions from even-numbered participants. We then compared the

TABLE 3

Performance Measures from Experiment 2

Domain	Proportion correct	Confidence	Overconfidence[a]
July temperatures	.642	.792	.150
State poverty levels	.643	.628	−.014
Mountain heights	.660	.545	−.115
State populations	.756	.779	.023
Shampoo prices	.772	.747	−.025
Presidential sequence	.862	.870	.008

[a] Overconfidence = confidence − proportion correct. Negative numbers indicate underconfidence.

TABLE 4

Correlations across Domains between Measures Taken on Different Subsamples of Questions, S1 and S2, and Presented to Different Subgroups of Participants in Experiment 2

	Proportion correct: Sample S2 (*M* = .738, .772; *SD* = .096, .088)	Confidence: Sample S2 (*M* = .742, .741; *SD* = .142, .114)	Overconfidence: Sample S2 (*M* = .004, −.031; *SD* = .126, .090)
Proportion correct: Sample S1 (*M* = .696, .690; *SD* = .074, .091)	.624, .861**	.499, .838**	.088, .216
Confidence: Sample S1 (*M* = .703, .720; *SD* = .102, .130)	.295, .600	.870,**.901**	.757,* .551
Overconfidence: Sample S1 (*M* = .008, .030; *SD* = .079, .094)	−.203, .002	.655, .442	.893,**.556
N = 6			

Note. For the first number in each pair, Subsample S1 of questions was seen by odd-numbered participants and Subsample S2 by even-numbered participants, and vice versa for the second number. For comparison, correlations between measures taken on the whole, unsplit sample were as follows: *r*(proportion correct, confidence) = .671, $p < .15$.; *r*(proportion correct, overconfidence) = −.115, *ns*; *r*(confidence, overconfidence) = .660, $p < .16$.

* $p < .10$.

** $p < .05$.

six domains to one another, comparing only scores that came from different questions *and* different participants: (a) vs (d) and (b) vs (c). That way, any correlations represent consistent differences between domains, independent of accidents of sampling, and independent of individual differences between participants. Results are shown in Table 4.

With $N = 6$ it is difficult to obtain statistical significance, plus, individual differences add noise to these correlations. Nevertheless there are clear patterns.[7] The correlations between confidence on one sample and confidence on the other are .87 and .90: Some domains consistently engender greater confidence than others. The correlations between accuracy on one subsample and accuracy on the other are .62 and .86: Some domains are harder than others. However, judgments of relative confidence and relative accuracy from domain to domain are only moderately correlated ($r = .50$, .84, .30, and .60). The result is that some domains are consistently more prone to overconfidence than others: The correlations between overconfidence on one subsample and overconfidence on the other are .89 and .56. Note though that there is again no evidence that harder domains are more prone to overconfidence, with correlations between accuracy and overconfidence of .09, .22, −.20, and 0.

[7] Recall that there are always two comparisons between samples of questions (i.e., two ways of comparing data from different questions and different participants within each domain). Each comparison uses separate data, so where both agree in sign, the results are more reliable than either alone.

Individual differences. The same split samples can be used to examine individual differences in confidence. As before, though, there is a confound between domain differences and individual differences because different participants saw different domains. In this analysis, we subtracted out domain differences by measuring accuracy, confidence, and overconfidence for each participant relative to the mean of all participants in that domain. For example, a participant's relative confidence would be the average confidence that participant expressed minus the mean confidence of all participants in that domain.

Table 5 shows the correlations obtained between measures taken on the two independent subsamples, S1 and S2, for each of the 54 participants. The correlation between accuracy on one set of questions and accuracy on the other is .39: Some participants are consistently more accurate than others in their domain, although there is a good deal of variation from one set to the other. The correlation between confidence on one set and confidence on the other is .93: Some participants are consistently more confident than others in their domain, and they tend to be very consistent in their overall sense of how well they are doing. Once again, more confident participants have only moderate cause to be so, with correlations between accuracy and confidence of .54 and .19. If we measure confidence and accuracy using all questions, without split samples, the correlation is .51. Thus, we find that some participants are consistently more overconfident than others in their domain: The correlation between overconfidence on one sample and overconfidence on the other is .50. As in the first study, the most confident participants tend to be the most overconfident,

TABLE 5

Correlations across Participants between Measures taken on Different Subsamples of Questions in Experiment 2, Relative to Other Participants in the Same Domain

	Proportion correct: Sample S2 (*SD* = .087)	Confidence: Sample S2 (*SD* = .080)	Overconfidence: Sample S2 (*SD* = .097)
Proportion correct: Sample S1 (*SD* =. 104)	.390***	.541****	.093
Confidence: Sample S1 (*SD* = .092)	.190	.931****	.591****
Overconfidence: Sample S1 (*SD* = .090)	−.253*	.331**	.497****
N = 54			

Note. By definition, the dependent measures all have a mean of 0. For comparison, correlations between measures taken on the whole, unsplit sample were as follows: *r*(proportion correct, confidence) = .502, $p < .001$.; *r*(proportion correct, overconfidence) = −.472, $p < .001$; *r*(confidence, overconfidence) = .525, $p < .001$.

* $p < .07$.

** $p < .02$.

*** $p <.01$

**** $p < .001$.

with correlations of .59 and .33. We did not replicate the earlier finding that less accurate participants also tend to be more overconfident, although the association was marginally significant in one of the two tests.

Conclusions. In most ways, the findings of the present study confirm those of Experiment 1. Answering long sets of questions from a single domain, even using the same items repeatedly in different combinations, does not seem to have any important effect on performance compared to receiving a mixed set of questions. This, then, is not a likely explanation for the finding of little or no overall overconfidence in recent studies such as those of Gigerenzer et al. (1991) and Juslin (1994), although there may be some effect in that direction. As in the first study, we also find that some participants are consistently more overconfident that others.

With the present study, we can also say that some *domains* are consistently more prone to overconfidence than others, at least for our population of participants. However, we still find no evidence that the harder domains are the ones more prone to overconfidence when we use split samples that control for the possibility of accidental missampling of contrary questions. At this point, we have no explanation for what makes some domains seem harder than they are, and others easier.

EXPERIMENT 3

Like the majority of earlier studies, our first two experiments examine only one kind of confidence judgment, namely how likely it is that one of two alternatives is the correct answer. There is no guarantee, however, that the processes behind such judgments are the same as those required in other tasks. In this study, we examine another type of confidence judgment, namely the setting of a confidence range of fixed probability for a single estimate. As noted earlier, there have been a number of studies documenting overconfidence in such tasks, but none that we know of has used questions randomly sampled from a variety of domains. As with earlier studies of two-choice tasks, it is possible that the questions used were selected for difficulty, and thus are artificially prone to overconfidence. This experiment serves two goals: (a) to test findings of overconfidence in confidence-range tasks using representative samples from multiple domains and (b) to determine if the patterns found in the first two experiments are also observed in range estimates.

Method

Participants. Thirty-two University of Chicago students were recruited as in Experiment 1. None participated in either of the other two studies.

Procedure. This study used the same procedures as in Experiment 1, except that participants were asked single-item questions for which they were to provide 90% confidence ranges. A total of 480 questions of this kind were prepared, 40 in each of the same 12 domains used in Experiment 1 (see Table

1), sampled randomly from the same sources as were used in that study. Prior to presenting the questions, the concept of a 90% range was explained as follows:

> You will be asked 150 questions for which you are to make an estimate of some number, like "How old is Madonna?" However, instead of estimating an exact number, we ask that you give a range, such that you think there is a 90% chance that the correct answer lies somewhere in the range. In other words, give a range such that you would expect to be wrong only about one out of ten times. (Answers that hit your high or low number exactly will be counted as correct.) . . . Type in your range by typing
>
> NUMBER ⟨Enter⟩
> NUMBER ⟨Enter⟩
>
> indicating that you think there is a 90% chance the right answer is between one number and the other. It doesn't matter if you give the lower number first or the higher number.

Participants were reminded of the intended criterion for the range with each question because the request for responses said "90% sure the answer is between this >[then, on the next line, following the first response] and this >." Participants were also told how to change their responses once entered, but were not permitted to return to earlier questions.

Results and Discussion

Normatively, participants who indicate their 90% confidence ranges should obtain approximately 90% correct answers.[8] In the present study, the correct answer fell inside the participants' confidence ranges 43% of the time. It is clear that the degree of overconfidence with these range estimates is much greater than for two-choice questions from the same domains. For a direct comparison, consider that when participants in Experiment 1 indicated a confidence of .90 ± .03, they were right 75% of the time. Alternatively, range judgments may be more like two separate judgments of .95 confidence, one for each end of the range. In Experiment 1, confidence of .95 ± .03 yielded 80% correct answers. So, given the level of confidence observed for two-choice questions, we might expect 60–75% of answers to be within range, compared with the observed 43%. In the general discussion we consider possible explanations for this substantial difference between two-choice and confidence-range questions.

As in Experiment 1, we performed a MANOVA to look for effects of condition (mixed questions from all 12 domains vs blocks of 40 questions from each of 3 domains) and time on the task. Condition was a between-participants variable, and first half vs second half of trials was within participants. Because reported confidence is held constant at 90%, the only dependent measure was the log-odds of the proportion of responses that were inside the range. Participants in the blocked condition seem to have done slightly better, with 49% of their answers inside their ranges compared to 37% in the mixed condition. However, the MANOVA indicated no effects that approached significance. We also looked at the blocked condition alone, separating trials into first, second, and third

[8] In the absence of complicated proper-scoring rules and incentives, participants could obtain this result by giving impossible answers on 10% of the questions and near-infinite ranges on 90%. However, neither we nor other researchers have found any evidence of this.

domain seen, and into first vs second half of each domain. There were no significant effects.

As in the two previous studies, we can use split samples to look for stable differences between individuals and between domains. To look for differences between domains, we divided the questions in each domain randomly into two subsamples, keeping the same division of questions for all participants. Because all of the items asked about in the present study were different, we did not have to exclude any, as we did in the previous study. For each of the 12 domains, we obtained the average proportion of in-range answers for each of the two subsamples. In the blocked condition, different participants see different combinations of 3 domains each, so domain differences may be confounded with individual differences. Therefore, we consider only the mixed condition, in which all participants see questions from all domains. Combining data from all participants, we find a correlation of .92 between a domain's average on one subsample and the other. Clearly, there are stable differences between domains. In the most overconfidence-prone domain (distance from Los Angeles), 22% of answers were within range; in the least overconfident domain (car mileage), 65%.

Previous research suggests that there may be a difficulty effect in confidence-range estimates (Lichtenstein et al., 1982; O'Connor & Lawrence, 1989), with more excessively narrow ranges for harder domains. There is no direct way to check for difficulty effects with the present method: Accuracy and overconfidence are one and the same. However, we can check whether the domains that were hardest or most prone to overconfidence in Experiment 1 were also more prone to overconfidence here. They were not: Across the 12 domains, overconfidence in this study correlated .22 with accuracy in Experiment 1, and $-.07$ with overconfidence. Neither is near significance.

Next, we examine individual differences. To do this, we switch to dividing the questions for each participant into those seen on even- vs odd-numbered trials, as in Experiment 1. This splitting method can be used when data are not to be combined across participants, and it has the advantage of varying which particular questions fall into which subset for each participant. We obtained measures of accuracy for each of the participants on each of the two subsamples. As before, data from the blocked condition confound individual differences with domain differences, so we examined the mixed condition separately. The correlation between a participant's accuracy on one subsample and the other was .84 ($N = 16$, $p < .001$). Clearly, there are strong, stable individual differences in overconfidence in this task. The most overconfident participant had 1% of the answers inside his or her ranges; the single underconfident participant had 92%.[9]

[9] The worst participant's answers were all off by orders of magnitude, and the single within-range answer seems to have been the result of a typing error. However, this participant did appear to take the task seriously (did not use the same number repeatedly, did not use arbitrary numbers such as 0, used different scales of numbers for the different domains, etc.) The next-worst participants had 8% and 9% of their answers within range. We reran our analyses eliminating the worst participant, in case he or she had misunderstood something. The results were substantively the same; the overall proportion of within-range answers rose to .45.

Conclusions. Confidence judgments based on setting 90% confidence ranges do not behave like confidence judgments in two-choice questions. In the two-choice task, it is clear that people make errors on each judgment, on how hard a particular domain is, and on how well they are performing overall. However, the amount of overall bias is modest: less than 5% overconfidence on average. With confidence-range questions, overconfidence is large, on the order of 45%. Differences between domains and between individuals are strong as well.

GENERAL DISCUSSION

As a number of investigators have shown, some familiar methods for studying confidence can produce misleading results. Researchers must be careful to select questions and to measure performance in ways that do not confound the effects of accidents of sampling, unsystematic judgment errors, and cognitive biases. What do we find when we do so? As in the vast majority of previous studies, the more confident participants are, the more overconfident (or less underconfident) they are. That is, in calibration curves like those in Figs. 1 and 2, the slope of the curve is less than 1. We also find evidence of systematic bias toward overconfidence. That bias does not depend much on the difficulty of the domain, nor on the opportunity to consider many questions from a single domain, as we expected it would. However, there are reliable differences between types of questions, between domains of questions, and between individuals.

How You Ask

The distinction that makes the biggest difference in our studies is the nature of the questions, two-choice vs confidence range. In our studies of two-choice questions, the overall difference between accuracy and confidence was modest (about 5% across 12 domains) and it varied considerably from person to person and from domain to domain. These findings are consistent with the conclusion that miscalibration results principally from unsystematic error rather than from biased retrieval and interpretation of information.

There are good reasons to expect that two-choice questions would not be fertile ground for biased information gathering. That is because both alternatives are explicit and available to the judge at the time of judgment. Many studies of hypothesis testing have shown that asking participants “is it A or is it B?” produces much less biased processing than asking “is it A?” (see Klayman, 1995; also Tversky & Koehler, 1994). This suggests that the main problem with two-choice questions is not that people base their answers on information that is biased toward consistency. Rather, people do not seem to appreciate how difficult it is to estimate what a collection of information implies about the chance of getting the answer right. In the absence of clear and plentiful feedback, people fail to learn that when they feel 90% sure, they are right only 77% of the time, and when they feel they are purely guessing, they are right about 56% of the time (if the findings of our studies are typical).

Instead, they maintain a kind of overconfidence *about their judgments of confidence*. It may also be that people do have a reasonable sense of the fallibility of their confidence judgments, but fail to regress feelings of extreme confidence or unconfidence toward the mean as much as they should, given that fallibility (see Griffin & Varey, 1996; Kahneman & Tversky, 1973). One can argue, as Ferrell (1994) has, that people cannot be expected to regress their estimates toward the mean if they do not know what the mean should be. Note, though, that when Gigerenzer et al. (1991) asked participants to estimate after the fact how many questions they got right, they were quite accurate, *under*estimating by a little.

The story is different for confidence-range judgments. Overconfidence averaged 47% across the 12 domains. This is consistent with earlier findings of very large overconfidence for range estimates (e.g., Russo & Schoemaker, 1992). We suspect that range judgments are more prone to overconfidence than are two-choice questions because they are more susceptible to biased information processing. With confidence-range questions there are no explicit alternatives. Rather, one can form an initial impression of a single answer and attempt to recruit information that supports or refutes that estimate. It is exactly this kind of one-sided situation that is most prone to confirmation bias (Klayman, 1995). Thus, biased retrieval and interpretation of evidence may very well explain the substantial overconfidence observed in confidence-range questions, even if their importance in two-choice judgments is in doubt.

What You Ask

With both two-choice and confidence-range questions, we found significant differences between performance on different domains of questions. It is a longstanding finding that miscalibration and overconfidence differ for different sets of questions. In particular, the conclusion has been that harder questions show greater overconfidence. Recently, though, Gigerenzer et al. (1991) and Juslin (1993, 1994) have concluded that the apparent difference between hard and easy sets of questions was an artifact of experimental methods, and would disappear with random sampling from defined populations of questions. Our results generally support that contention. We find no reliable signs of a relation between the amount of overconfidence and the difficulty of the domain of questions. (Some recent studies that have looked at larger amounts of data from a larger array of domains have found modest effects of difficulty, however; see Juslin et al., 1997). Nevertheless, we do find that domains differ significantly in the extent to which they elicit under- or overconfidence. Some domains are harder than they seem and some easier, and our split-sample analyses show that those differences are not attributable to accidents of sampling. At present, we have no hypotheses about what features of a domain predict its degree of overconfidence. This finding highlights the advice given by Brunswik (1952) that researchers who wish to make generalizations need to have good samples of the population of stimuli, as well as of the population of participants.

Whom You Ask

We also found consistent individual differences in the degree of overconfidence participants expressed (see also Soll, 1996). At its least interesting, this finding might simply mean that different judges have consistently different ideas about how to use the numerical response scale. However, the magnitude of individual differences observed in Experiment 3 is such that it seems very unlikely they result from differing interpretations of what a probability of 90% means.

A recent study (Klayman & Burt, 1998) provides further evidence that individual differences in overconfidence are psychologically substantive. That study found a correlation between MBA students' overconfidence on confidence-range questions and the social environment in which they worked. Greater overconfidence was associated with more "constrained" social networks. Those are small networks with many strong interconnections; someone in a central, coordinating position; and weak connections to outsiders (Burt, 1992, 1997). Such social structures are good for forming shared beliefs and they provide many opportunities for members to tell one another that they are right and few opportunities for prevailing ideas to be challenged.

If there are real differences among people in their proneness to overconfidence, that has interesting implications for the real world. In a world in which competence is hard to measure, confidence often wins the day. This is troubling if, as our results suggest, the most confident people are also the most overconfident. The whole issue of individual differences in judgment is worthy of further study.

Conclusions

In the 1980s, the question of bias in confidence judgments seemed settled: People are grossly overconfident on all but the easiest of questions. In the 1990s, the matter was reopened, and a new conclusion was proposed: People are imperfect but generally unbiased judges of confidence; only the choice of questions was biased. Miscalibration of confidence judgments has important implications in either case, but the distinction is significant. If people are materially overconfident, then we can aid them by developing debiasing techniques that help them attend to contradictory evidence (Fischhoff, 1982; Hoch, 1985; Koriat et al., 1980; Lord, Lepper, & Preston, 1984). If random error is the main culprit, then such "debiasing" might be misguided, merely trading one direction of error for another. Instead, the focus might be on better feedback and training to reduce noise, or on means of extracting valid information from noisy judges (see Clemen, 1986, 1989; Wallsten, Budescu, Erev, & Diederich, 1997). The finding that biased samples of questions can produce apparent over- or underconfidence has additional implications. Many judges receive a biased subset of questions in their work. For example, doctors, lawyers, and consultants get questions that are harder than those in the environment at large. Do they know enough to adjust their confidence accordingly?

Our understanding of subjective confidence is limited by a lack of evidence concerning underlying cognitive processes (as Griffin & Varey, 1996, also point out; for some recent efforts in this direction, see, e.g., Gigerenzer & Goldstein, 1996; Yates, Lee, Shinotsuka, & Sieck, 1998). For example, both two-choice judgments and range estimates are of theoretical and practical importance. These two types of judgments must share some cognitive elements, but they clearly have unique elements as well. There are also many other interesting kinds of confidence judgments. How sure are you about the meaning of "supercilious"? How certain are you that you heard the phone ring on your way out of your house (cf. Bjorkman, Juslin, & Winman, 1993; Juslin & Olsson, 1997)? It is unlikely that any single mechanism can explain all types of confidence judgments. Similarly, differences in overconfidence between domains and between individuals point to systematic effects of information content, information processing, and the relation between them. Why are some people, some domains, and some types of judgment more prone to overconfidence than others? The present studies do not provide definitive answers, but they do underscore the importance of asking these questions in seeking to understand the processes by which confidence judgments are made, and how they might be aided.

REFERENCES

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Making*, **10,** 153–285.

Babad, E. (1987). Wishful thinking and objectivity among sports fans. *Social Behavior*, **4,** 231–240.

Bjorkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, **58,** 386–405.

Bjorkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*, **54,** 75–81.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, **65,** 212–219.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: Univ. of Chicago Press.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment: Part I. New theoretical developments. *Journal of Behavioral Decision Making*, **10,** 157–171.

Budescu, D. V., Erev, I., Wallsten, T. S., & Yates, J. F. (Eds.). (1997). Stochastic and cognitive models of confidence [Special issue]. *Journal of Behavioral Decision Making*, **10**.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment: Part II. Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, **10,** 173–188.

Burt, R. S. (1992). *Structural holes*. Cambridge, MA: Harvard Univ. Press.

Burt, R. S. (1997). The contingent value of social capital. *Administrative Science Quarterly*, **42,** 339–365.

Clemen, R. T. (1986). Calibration and the aggregation of probabilities. *Management Science*, **32,** 312–314.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5,** 559–609.

Dawes, R. M. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments.

In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 327–345). Bern, Switzerland: Huber.

Dunning, D., Griffin, D. W., Milojkovic, J., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology,* **58,** 568–581.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review,* **101,** 519–528.

Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 411–451). Chichester: Wiley.

Ferrell, W. R., & McGoey, P. J. (1980). A mode of calibration for subjective probabilities. *Organizational Behavior and Human Performance,* **26,** 32–53.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge Univ. Press.

Fischhoff, B., & MacGregor, D. (1983). Judged lethality: How much people seem to know depends on how they are asked. *Risk Analysis,* **3,** 229–236.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance,* **3,** 552–564.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review,* **103,** 650–669.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review,* **98,** 506–528.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology,* **24,** 411–435.

Griffin, D. W., & Varey, C. A. (1996). Towards a consensus on overconfidence. *Organizational Behavior and Human Decision Processes,* **65,** 227–231.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science,* **1,** 78–82.

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., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review,* **104,** 344–366.

Juslin, P., Olsson, H., & Bjorkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making,* **10,** 189–209.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes,* **73,** 3–26.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review,* **80,** 237–251.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 509–520). Cambridge, UK: Cambridge Univ. Press.

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

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of learning and motivation: Vol. 32. Decision making from a cognitive perspective* (pp. 365–418). New York: Academic Press.

Klayman, J., & Burt, R. S. (1998). *Individual differences in confidence and experiences in social networks*. Working paper, Center for Decision Research, Graduate School of Business, University of Chicago.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* **6,** 107–118.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, **108,** 480–498.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology,* **32,** 311–328.

Larrick, R. P. (1993). Motivational factors in decision theories: The role of self-protection. *Psychological Bulletin*, **113,** 440–450.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?: The calibration of probability judgments. *Organizational Behavior and Human Performance,* **20,** 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge Univ. Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology,* **47,** 1231–1243.

May, R. S. (1986). Inferences, subjective probability, and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, B. Jungermann, P. Lourens, & G. Sevon (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 G. Wright & P. Ayton (Eds.), *Subjective probability* (pp.453–482). Chichester: Wiley.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology,* **12,** 595–600.

Murphy, A. H., & Winkler R. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, **79,** 489–500.

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.

Paese, P. W., & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision making. *Organizational Behavior and Human Decision Processes*, **48,** 100–130.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **14,** 85–92.

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes,* **58,** 203–213.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review,* **33,** 7–17.

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.

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

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, **10,** 243–268.

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

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F., Lee, J.-W., Shinotsuka, H., & Sieck, W. R. (1998, November). *Oppositional deliberation: Toward explaining overconfidence and its cross-cultural variations*. Paper presented at the meeting of the Psychonomics Society, Dallas, TX.

Received September 21, 1998