Journal of Experimental Psychology:
Human Perception and Performance
1977, Vol. 3, No. 4, 552-564

Knowing with Certainty: The Appropriateness of Extreme Confidence

Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein
Decision Research, A Branch of Perceptronics
Eugene, Oregon

How often are people wrong when they are certain that they know the answer to a question? The studies reported here suggest that the answer is "too often." For a variety of general-knowledge questions (e.g., absinthe is [a] a liqueur or [b] a precious stone), subjects first chose the most likely answer and then indicated their degree of certainty that the answer they had selected was, in fact, correct. Across several different question and response formats, subjects were consistently overconfident. They had sufficient faith in their confidence judgments to be willing to stake money on their validity. The psychological bases for unwarranted certainty are discussed in terms of the inferential processes whereby knowledge is constructed from perceptions and memories.

Two aspects of knowledge are what one believes to be true and how confident one is in that belief. Both are represented in a statement like, "I am 70% certain that Quito is the capital of Equador." While it is often not difficult to assess the veridicality of a belief (e.g., by looking in a definitive atlas), evaluating the validity of a degree of confidence is more difficult. For example, the 70% certainty in the above statement would seem more appropriate if Quito is the capital than if Quito isn't the capital, but that is a rather crude assessment. In a sense, only statements of certainty (0% or 100%) can be evaluated individually, according to whether the beliefs to which they are attached are true or false.

This research was supported by the Advanced Research Projects Agency (ARPA) of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-76-0074 (ARPA Order No. 3052) under a subcontract from Decisions and Designs, Inc. to Oregan Research Institute.

We would like to thank Bernard Corrigan, Robyn Dawes, Ward Edwards, and Amos Tversky for their contributions to this project.

Requests for reprints should be sent to Baruch Fischhoff, Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401.

One way to validate degrees of confidence is to look at the calibration of a set of such confidence statements. An individual is well calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true is equal to the probability assigned. For example, half of those statements assigned a probability of .50 of being true should be true, as should 60% of those assigned .60, and all of those about which the individual is 100% certain. A burgeoning literature on calibration has been surveyed by Lichtenstein, Fischhoff, and Phillips (in press). The primary conclusion of this review is that people tend to be overconfident, that is, they exaggerate the extent to which what they know is correct. A fairly typical set of calibration curves, drawn from several studies, appears in Figure 1. We see that when people should be right 70% of the time, their "hit rate" is only 60%; when they are 90% certain, they are only 75% right; and so on.

People's poor calibration may be, in part, just a question of scaling. Probabilities (or odds) are a set of numbers that people use with some internal consistency (e.g., the curves in Figure 1 are more or less monotonically increasing) but not in accordance

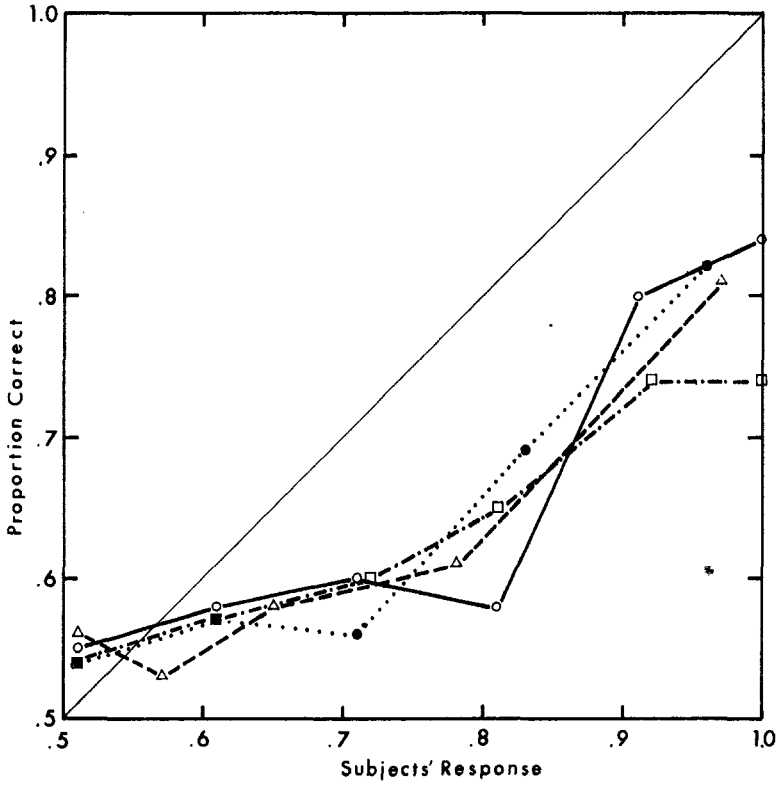


Figure 1. Some representative calibration curves. (Taken from Lichtenstein, Fischhoff, & Phillips, in press. Copyright 1977 by D. Reidel Publishing Co. Reprinted by permission.)

with the absolute criteron of calibration. Miscalibration can have serious consequences (see Lichtenstein et al., in press), yet people's inabilty to assess appropriately a probability of .80 may be no more surprising than the difficulty they might have in estimating brightness in candles or temperature in degrees Fahrenheit. Degrees of certainty are often used in everyday speech (as are references to temperature), but they are seldom expressed numerically nor is the opportunity to validate them often available (Tversky & Kahneman, 1974).

The extremes of the probability scale are, however, not such foreign concepts. Being 100% certain that a statement is true is readily understood by most people, and its appropriateness is readily evaluated. The following studies examine the calibration of people's expressions of extreme certainty. The studies ask, How often are people wrong when they are certain that they are right? In Experiment 1, the answer is sought in probability judgments elicited by questions posed in four different ways.

Experiment 1

Method

Stimuli. The questions covered a wide variety of topics, including history, music, geography, nature, and literature. The four formats used were the following:

1. Open-ended format. Subjects were presented with a question stem, which they were asked to complete; for example, "Absinthe is a _____." After writing down an answer, they estimated the probability that their answer was correct, using a number from .00 to 1.00.

2. One-alternative format. Subjects were asked to assess the probability (from .00 to 1.00) that simple statements were correct; for example, "What is the probability that absinthe is a precious stone?" The statement of fact being judged was sometimes true and sometimes false.

3. Two-alternative format (half range of responses). For each question, subjects were asked

to choose the correct answer from two that were offered. After making each choice, they judged the probability that the choice was correct; for example, "Absinthe is (a) a precious stone or (b) a liqueur." Since they chose the more likely answer, their probabilities were limited to the range from .50 to 1.00.

4. Two-alternative format (full range of responses). Instead of having subjects pick the answer most likely to be correct as in Format 3, the experimenters randomly selected one of the two alternatives (e.g., [b] a liqueur) and had subjects judge the probability that the selected alternative was correct. Here the full range [.00, 1.00] was used. As in Format 3, one answer was correct.

Subjects and procedure. The subjects were 361 paid volunteers who responded to an ad in the University of Oregon student newspaper. They were assigned to the four groups according to preference for experiment time and date. Each group received the questions in only one of the four formats. Besides the differences in question format, the specific questions used differed somewhat from group to group. Instructions were brief and straightforward, asking subjects to choose or produce an answer and assign a probability of being correct in accordance with the format used.

Results

Lichtenstein and Fischhoff (in press) and Fischhoff and Lichtenstein (Note 1) have reported on the calibration of the entire range of probability responses of subjects in Experiment 1. Here we examine only their extreme responses. Table 1 shows (a) the frequency with which subjects indicated 1.00 or .00 as the probability an alternative was correct and (b) the percentage of answers associated with these extreme probabilities that were, in fact, correct. Answers assigned a probability of 1.00 of being correct were right between 20% and 30% of the time. Answers assigned a probability of .00 were right between 20% and 30% of the time. In Formats 2 and 4, where responses of 1.00 and .00 were possible, both responses occurred with about equal frequency. Furthermore, alternatives judged certain to be correct were wrong about as often as alternatives judged certain to be wrong were correct. The percentage of false certainties ranged from about 17% (Format 1) to about 30% (Format 2), but comparisons across formats should be made with caution because the items differed. Clearly, our subjects were wrong all too often when they were certain of the correctness of their choice of answer.

Experiment 2

Experiment 1 might be faulted because of the insensitivity of the response mode. With probabilities, subjects using the stereotypic responses of .50, .55, .60, and so on, have few possible responses for indicating different degrees of high certainty. At the extreme, most subjects restricted themselves to the responses .90, .95, and 1.00, corresponding to odds of 9:1, 19:1, and ∞:1. Perhaps with a more graduated response mode, subjects would be better able to express different levels of certainty. In Experiment 2, subjects were presented with general-knowledge questions concerned with a single topic—the incidence of different causes of death in the United States—and

Table 1
Analysis of Certainty Responses in Experiment 1

Question format	No. items	No. subjects	Total no. responses	Certainty responses (p)	% certainty responses	% correct certainty responses
1. Open ended	43	30	1,290	1.00	19.7	83.1
2. One alternative	75	86	6,450	1.00	14.2	71.7
				.00	13.8	29.5
3. Two alternative (half range)	75	120	9,000	1.00	21.8	81.8
4. Two alternative (full range)	50	131	6,500	1.00	17.3	80.7
				.00	19.1	20.5

asked to express their confidence in their answers in odds. The odds scale is open ended at the extremes, easily allowing the expression of many different levels of great certainty (e.g., 20:1, 50:1, 100:1, 500:1, etc.).

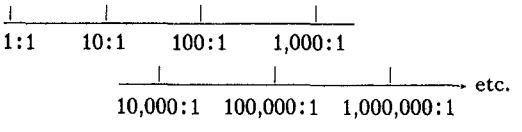
Method

Stimuli. All items involved the relative frequencies of the 41 lethal events shown in Table 2. They were chosen because they were easily understood and had fairly stable death rates over the last 5 years for which statistics were available. The event frequencies appearing in Table 2 were estimated from vital statistics reports prepared by the National Center for Health Statistics and the "Statistical Bulletin" of the Metropolitan Life Insurance Company. These frequencies provided the correct answers for the questions posed to our subjects.

From among these 41 causes of death, 106 pairs were constructed according to the following criteria: (a) Each cause appeared in approximately six pairs and (b) the ratios of the statistical rates of the more-frequent event to the less-frequent event varied systematically from 1.25:1 (e.g., accidental falls vs. emphysema) to about 100,000:1 (e.g., stroke vs. botulism).

Procedure. Subjects' instructions read as follows:

> Each item consists of two possible causes of death. The question you are to answer is: Which cause of death is more frequent, in general, in the United States?
>
> For each pair of possible causes of death, (a) and (b), we want you to mark on your answer sheet which cause you think is more frequent.
>
> Next, we want to decide how confident you are that you have, in fact, chosen the more frequent cause of death. Indicate your confidence by the odds that your answer is correct. Odds of 2:1 mean that you are twice as likely to be right as wrong. Odds of 1,000:1 mean that you are a thousand times more likely to be right than wrong. Odds of 1:1 mean that you are equally likely to be right or wrong. That is, your answer is completely a guess.
>
> At the top of the answer sheet we have drawn a scale that looks like this:



> This scale is used to give you an idea of the kinds of numbers you might want to use. You don't have to use exactly these numbers. You could write 75:1 if you think that it is 75 times more likely that you are right than you are wrong, or 1.2:1 if you think that it is only 20% more likely that you are right than wrong.
>
> Do not use odds less than 1:1. That would mean that it is less likely that you are right than that you are wrong, in which case you should indicate the other cause of death as more frequent.

Table 2
Lethal Events Whose Relative Frequencies Were Judged by Subjects in Experiments 2 and 3

Lethal event	Actual deaths per 100 million[a]
Smallpox	0
Poisoning by vitamins	0.5
Botulism	1
Measles	2.4
Fireworks	3
Smallpox vaccination	4
Whooping cough	7.2
Polio	8.3
Venomous bite or sting	23.5
Tornado	44
Lightning	52
Nonvenomous animal	63
Flood	100
Excess cold	163
Syphilis	200
Pregnancy, childbirth, and abortion	220
Infectious hepatitis	330
Appendicitis	440
Electrocution	500
Motor vehicle – train collision	740
Asthma	920
Firearm accident	1,100
Poisoning by solid or liquid	1,250
Tuberculosis	1,800
Fire and flames	3,600
Drowning	3,600
Leukemia	7,100
Accidental falls	8,500
Homicide	9,200
Emphysema	10,600
Suicide	12,000
Breast cancer	15,200
Diabetes	19,000
Motor vehicle (car, truck, or bus) accident	27,000
Lung cancer	37,000
Cancer of the digestive system	46,400
All accidents	55,000
Stroke	102,000
All cancers	160,000
Heart disease	360,000
All diseases	849,000

[a] Per-year death rates are based on 100 million United States residents.

Table 3
Percentage of Correct Answers for Major Odds Categories

Odds	Appropriate % correct[a]	Lethal events: Experiment 2 N	% N	% correct	Lethal events: Experiment 3 N	% N	% correct	General-knowledge questions: Experiment 4 N	% N	% correct
1:1	50	644	9	53	339	8	54	861	19	53
1.5:1	60	68	1	57	108	2.5	59	210	5	56
2:1	67	575	8	64	434	10	65	455	1	63
3:1	75	189	2	71	252	6	65	157	3.5	76
5:1	83	250	4	70	322	8	71	194	4	76
10:1	91	1,167	17	66	390	9	76	376	8	74
20:1	95	126	2	72	163	4	81	66	1.5	85
50:1	98	258	4	68	227	5	74	69	1.5	83
100:1	99	1,180	17	73	319	8	87	376	8	80
1,000:1	99.9	862	13	81	219	5	84	334	7	88
10,000:1	100	459	7	87	138	3	92	263	6	89
100,000:1	100	163	2	85	23	.5	96	134	3	92
1,000,000:1	100	157	2	90	47	1	96	360	8	94
Total		6,098	88		2,981	70		3,855	75	
Overall % correct				71.0			72.5			73.1

Note. % *N* refers to the percentage of odds judgments that fell in each of the major categories. There were 66 subjects in Experiment 2, 40 in Experiment 3, and 42 in Experiment 4.
[a] For well-calibrated subjects.

In case some of the causes of death are ambiguous or not well defined by the brief phrase that describes them, we have included a glossary for several of these items. Read this glossary before starting.

Subjects. The subjects were 66 paid volunteers who answered an ad in the University of Oregon student newspaper.

Results[1]

Table 3 shows the percentages of correct answers, grouped across subjects, for each of the most frequently used (major) odds categories. At odds of 1:1, 1.5:1, 2:1, and 3:1, subjects were reasonably well calibrated. However, as odds increased from 3:1 to 100:1, there was little or no increase in accuracy. Only 73% of the answers assigned odds of 100:1 were correct. Accuracy jumped to 81% at 1,000:1 and to 87% at 10,000:1. For the answers assigned odds of 1,000,000:1 or greater, accuracy was 90%. For the latter responses, the appropriate degree of confidence would have been odds of 9:1. The 12% of responses that are not listed in Table 3 because they fell between the major odds categories showed similar calibration.

As in Experiment 1, subjects in Experiment 2 exhibited great overconfidence. They were frequently wrong at even the highest odds levels. Moreover, they gave many extreme odds responses. Of 6,996 odds judgments, 3,560 (51%) were greater than 50:1. Almost one fourth of the responses were greater than 1,000:1.

Experiment 3

Although the tasks and instructions for Experiments 1 and 2 seemed reasonably straightforward, we were concerned that subjects' extreme overconfidence might be

[1] A more detailed description of subjects' performances on this task and several related ones can be found in Lichtenstein, Slovic, Fischhoff, Combs, and Layman (Note 2).

due to lack of motivation or misunderstanding of the response scale. Experiment 3 replicated Experiment 2, giving more care and attention to instructing and motivating the subjects.

Method

Experiment 3 used the 106 causes-of-death questions and odds response format of Experiment 2. The experimenter started the session with a 20-minute lecture to the subjects. In this lecture, the concepts of probability and odds were carefully explained. The subtleties of expressing one's feelings of uncertainty as numerical odds judgments were discussed, with special emphasis on how to use small odds (between 1:1 and 2:1) when one is quite uncertain about the correct answer. A chart was provided showing the relationship between various odds estimates and the corresponding probabilities. Finally, subjects were taught the concept of calibration and were urged to make odds judgments in a way that would lead them to be well calibrated. (The complete text of the instructions is available from the authors.)

The subjects for Experiment 3 were 40 persons who responded to an ad in the University of Oregon student newspaper. As in previous experiments, they were paid for participating. Group size was held to about 20 to increase the likelihood that subjects would ask questions about any facet of the task that was unclear.

Results

The proportion of correct answers for each of the most frequent odds categories is shown in the center portion of Table 3. The detailed instructions had several effects. First, subjects were much more prone to use atypical odds such as 1.4:1, 2.5:1, and so on. Only 70% of their judgments fell within the major odds categories of Table 3, as compared to 88% for Experiment 2. Second, their odds estimates tended to be smaller. About 43% of their estimates were 5:1 or less, compared to 27% for this category in Experiment 1. Third, subjects in this experiment were more often correct at odds above 10:1 and thus were better calibrated.

Nevertheless, subjects again exhibited unwarranted certainty. They assigned odds greater than or equal to 50:1 to approximately one third of the items. Only 83% of the answers associated with these odds were correct. When subjects estimated odds of 50:1, they were correct 74% of the time and thus should have been giving odds of about 3:1. At 1,000:1, they should have been saying about 5:1.

Although only 70% of the responses fell in the major odds categories of Table 3, inclusion of the remaining 32% would not have changed the picture. Odds estimates falling between major categories were calibrated similarly to estimates within those categories. Elaborate instruction tempered subjects' extreme overconfidence, but only to a limited extent.

Experiment 4

Is there something peculiar to the causes-of-death items that induces such overconfidence? Experiment 4 replicated Experiment 3 using general-knowledge questions (of the type used in Experiment 1) matched in difficulty with the 106 causes-of-death items. In addition, subjects' faith in their odds judgments was tested by their willingness to participate in a gambling game based on those judgments.

Method

The questionnaire consisted of 106 two-alternative items covering a wide variety of topics; for example, "Which magazine had the largest circulation in 1970? (a) *Playboy* or (b) *Time*"; "Aden was occupied in 1839 by the (a) British or (b) French"; "Bile pigments accumulate as a result of a condition known as (a) gangrene or (b) jaundice." These items were taken from a large item pool with known characteristics. Availability of this pool allowed us to select items matched in difficulty, question by question, with the 106 items about lethal events studied in Experiments 2 and 3.

The subjects were 42 paid volunteers, recruited by an ad in the University of Oregon student newspaper. The instructions paralleled those of Experiment 3. Subjects first received the detailed lecture describing the concepts of probability, odds, and calibration. They then responded to the 106 general-knowledge items, marking the answer they thought to be correct and expressing their certainty about that answer with an odds judgment.

After responding to the 106 items, they were asked whether they would be willing to accept gambles contingent on the correctness of their answers and the appropriateness of their odds estimates. If subjects really believe in their extreme (extremely overconfident) odds responses, it

should be possible to construct gambles that they are eager to accept but which, in fact, are quite disadvantageous to them. The game was described by the following instructions:

> The experiment is over. You have just earned $2.50, which you will be able to collect soon. But before you take the money and leave, I'd like you to consider whether you would be willing to play a certain game in order to possibly increase your earnings. The rules of the game are as follows.
>
> 1. Look at your answer sheet. Find the questions where you estimated the odds of your being correct as 50:1 or greater than 50:1. How many such questions were there? _____ (write number)
>
> 2. I'll give you the correct answers to these "50:1 or greater" questions. We'll count how many times your answers to these questions were wrong. Since a wrong answer in the face of such high certainty would be surprising, we'll call these wrong answers "your surprises."
>
> 3. I have a bag of poker chips in front of me. There are 100 white chips and 2 red chips in the bag. If I reach in and randomly select a chip, the odds that I will select a white chip are 100:2 or 50:1, just like the odds that your "50:1" answers are correct.
>
> 4. For every "50:1 or greater" answer you gave, I'll draw a chip out of the bag. (If you wish, you can draw the chips for me.) I'll put the chip back in the bag before I draw again, so the odds won't change. The probability of my drawing a red chip is 1/51. Since drawing a red chip is unlikely, every red chip I draw can be considered "my surprise."
>
> 5. Every time you are surprised by a wrong answer to a "50:1 or greater" question, you pay me $1. Every time I am surprised by drawing a red chip, I'll pay you $1.
>
> 6. If you are well calibrated, this game is advantageous to you. This is because I expect to lose $1 about once out of every 51 times I draw a chip, on the average. But since your odds are sometimes higher than 50:1, you expect to lose less often than that.
>
> 7. Would you play this game? Circle one. Yes No

Subjects who declined were then asked if they would play if the experimenter raised the amount he would pay them to $1.50 whenever he drew a red chip, while they still had to pay only $1 in the event of a wrong answer. Those who still refused were offered $2 and then a final offer of $2.50 for every red chip. Since the experimenters expected the game to be unfair to subjects (by capitalizing on a "known" judgmental bias), it was not actually played for money.

Results

The proportion of correct answers associated with each of the most common odds responses is shown in the right-hand column of Table 3. Compared with the previous studies, subjects in Experiment 4 gave a higher proportion of 1:1 odds (19% of the total responses). A few difficult items led almost all of the subjects to give answers close to 1:1, indicating that they were trying to use small odds when they felt it was appropriate to do so. However, this bit of restraint was coupled with as high a percentage of large odds estimates as was given by the untutored subjects in Experiment 2. About one quarter of all answers were assigned odds equal to or greater than 1,000:1.

Once again, answers to which extremely high odds had been assigned were frequently wrong. At odds of 10:1, subjects were correct on about three out of every four questions, appropriate to odds of 3:1. At 100:1, they should have been saying 4:1. At 1,000:1 and at 100,000:1, estimates of about 7:1 and 9:1 would have been more in keeping with subjects' actual abilities. Over the large number of questions for which people gave odds of 1,000,000:1 or higher, they were wrong an average of about 1 time out of every 16.

The gambling game. Of the 42 subjects, 27 agreed to play the gambling game described above for $1. Six more agreed when the stakes were raised to $1.50 every time the experimenter drew a red chip. Of the holdouts, 3 subjects agreed to play at $2 for every red chip and 2 more agreed when the final offer of $2.50 was made. Only 3 subjects refused to participate at any level of payment per red chip.

After subjects had made their decisions about playing the game, they were asked whether they would change their minds if the game were to be played, on the spot, for real money. No subject indicated a desire to change his or her decision. Two subjects approached the experimenter after the experiment requesting that they be given a chance to play the game for cash. Their request was refused.

Of course, this game is strongly biased in favor of the experimenter. Since subjects were wrong about once for every eight answers assigned odds of 50:1 or greater, the game would have been approximately fair had the experimenter removed 86 of the white chips from the bag, leaving its contents at 14 white and 2 red chips.

The expected outcome of playing the game with each subject was simulated. Every wrong answer on a "50:1 or greater" question was assumed to cost the subject $1. The experimenter was assumed to have drawn 1/51 of a red chip for every answer given at odds greater than or equal to 50:1; his expected loss was then calculated in accordance with the bet the subject had accepted. For example, if a subject accepted the experimenter's first offer ($1 per red chip) and gave 17 "50:1 or greater" answers, the experimenter's simulated loss was 17/51 dollars (33¢).

The subjects who agreed to play averaged 38.3 questions with odds greater than or equal to 50:1. Thirty-six persons had expected monetary losses, and three had expected wins. Individual expected outcomes ranged between a loss of $25.63 and a gain of $1.84. The mean expected outcome was a loss of $3.64 per person and the median outcome was a loss of $2.35. Ten persons would have lost more than $5. The 39 subjects would have lost a total of $142.13 across 1,495 answers at odds greater than or equal to 50:1, an average loss of 9.5¢ for every such answer. The two persons who earnestly requested special permission to play the game had expected losses totaling $33.38 between them.

Experiment 5: Playing for Keeps

Subjects in Experiment 4 viewed their overconfident odds judgments as faithful enough reflections of their state of knowledge that they were willing to accept hypothetical bets more disadvantageous than many that can be found in a Las Vegas casino. Before concluding that there is money to be made in "trivia hustling," we decided to replicate Experiment 4 with real gambling at the end.

Method

Nineteen subjects participated in Experiment 5. It differed from Experiment 4 only in that the gambling game was presented as a real game. After responding to the 106 items, subjects heard the gambling game instructions and decided whether or not they would play. They were told that they could lose all the money they had earned in the experiment and possibly even more than that. After they made their decisions about playing the game, subjects were told that any earnings from the game would be added to their pay for the experiment, but that if they lost money, none of the money initially promised them for participating would be confiscated. The game was then played on those terms.

Six of the 19 subjects agreed to play the game as first specified (with a $1 payment for each "experimenter's surprise"). Three more agreed to play when the experimenter offered to increase the payment to $1.50 per red chip. Increasing the payment to $2 brought in one additional player, and three more agreed to play at $2.50. Six subjects consistently refused to participate; some because they felt they were not well calibrated, others because they did not like to gamble.

Results

When the game was actually played, the 13 participating subjects missed 46 of the 387 answers (11.9%) to which they had assigned odds greater than or equal to 50:1. All 13 subjects would have lost money, ranging from $1 to $11 (in part because, by chance, no red chips were drawn). When the experimenter's part of the game was simulated as in Experiment 4, four subjects would have lost more than $6, and the average participating subject would have lost $2.64. Thus, the hypothetical nature of the gamble in Experiment 4 apparently had minimal influence on subjects' willingness to bet.

General Subject and Item Analyses

Is undue confidence found only in a few subjects or only for a few special items? If cases of extreme overconfidence are concentrated in only a few subjects, then the generality of our conclusions would be limited. Pathological overconfidence on the part of a small sector of the public would be worth

Table 4
Frequency of Extreme Overconfidence (Odds Greater Than or Equal to 50:1 That Were Assigned to Wrong Answers)

No. cases of extreme overconfidence	Number of subjects		No. extremely overconfident subjects	Number of items	
	Experiment 3	Experiment 4		Experiment 3	Experiment 4
0	5[a]	3	0	33[b]	41
1	3	7	1	25	20
2	6	5	2	19	16
3	6	5	3	13	10
4	4	9	4	3	8
5	1	1	5	2	2
6	4	1	6	4	1
7	2	2	7	1	3
8	1	0	8	1	0
9	0	3	9	0	0
10	2	2	10	0	3
11	0	1	11	2	0
12	1	0	12	1	0
13	1	0	13	0	1
14	0	0	14	1	0
15	0	2	15	1	1
16	2	0			
17	1	0			
More than 17[c]	1 (32)	1 (27)			

[a] There were five subjects in Experiment 3 who never showed extreme overconfidence.
[b] There were 33 items in Experiment 3 for which no subject showed extreme overconfidence.
[c] Actual number of cases is in parentheses.

exploring further but would not tell us much about cognitive functioning in general. The results of the gambling games reported above show that this was not the case. Most subjects were willing to play and most would have lost money because they were too often wrong when using extreme odds. The left columns of Table 4 show the distribution of cases of extreme overconfidence (defined as giving odds of 50:1 or greater and being wrong) over subjects for Experiments 3 and 4. The great majority of subjects had one or more cases of extreme overconfidence. The median number was 4 in Experiment 4 and between 3 and 4 in Experiment 3, well over what would be expected with well-calibrated subjects. In each experiment, one subject appeared to be an outlier (those subjects having 32 and 27 cases). Reanalyzing the data after removing those two subjects had no effect on our conclusions.

The right columns of Table 4 show the distribution of cases of extreme overconfidence over items. If most cases were concentrated in only a few items, the situation would be rather different than if a broad section of items fooled some of the people some of the time. It would not necessarily be less interesting, for it would remain to be explained why people went astray on those few items. As the results in Table 4

Table 5
Percentage Wrong with Deceptive and Nondeceptive Items

Experiment and item	Percentage wrong associated with odds of		
	≥50:1	≥100:1	≥1000:1
Experiment 3			
All items (106)	16.6	14.1	12.9
Deceptive items (18)	73.9	75.5	72.3
Nondeceptive items (88)	8.9	6.7	6.9
Experiment 4			
All items (106)	13.8	13.1	10.8
Deceptive items (17)	73.2	76.7	70.6
Nondeceptive items (89)	7.6	6.8	5.2
Expected with perfect calibration	≤1.96	≤.99	≤.10

indicate, both situations seem to have been true. There are some items on which many people gave high odds to the wrong answer, but most items did show a few such cases.

The items on which six or more subjects showed extreme overconfidence were all items that might be described as "deceptive," ones which less than 50% of the subjects answered correctly. Some correlation between deceptiveness and extreme overconfidence is inevitable; many subjects must get an answer wrong before many can get it wrong *and* be certain that they are right. There were 18 items in Experiment 3 and 17 items in Experiment 4 answered correctly by less than 50% of our subjects.

Table 5 shows the incidence of cases of extreme overconfidence with deceptive and nondeceptive items. Although extreme overconfidence is disproportionately prevalent with the deceptive items, it is still abundant with the nondeceptive ones. If the deceptive items are removed from the sample, then the remaining distribution of cases of extreme overconfidence over items closely resembles a Poisson distribution, which is what one would expect if such cases were distributed at random over items. One third of the easiest items, those answered correctly by 90% or more of our subjects, had at least one case of a subject answering wrongly and giving odds of being correct of 1,000:1 or greater. Deleting the one extreme subject from each of Experiments 3 and 4 had little effect on this result. Clearly, a few subjects or items are not responsible for the extreme overconfidence effect.

General Discussion

These five experiments have shown people to be wrong too often when they are certain that they are right. This result was obtained with both probability and odds responses, with minimal and extensive instructions and with two rather different types of questions. Subjects were sufficiently comfortable with their expressions of certainty that they were willing to risk money on them in both hypothetical and real gambles. Finally, cases of extreme overconfidence were widely distributed over subjects and items.

Although these studies have shown the effect to be a robust one, they have certainly not closed the topic. Further research with different subjects, different items, and different instructions would be most useful. Some moderately informed guesses at the results of such additional studies are possible. Lichtenstein and Fischhoff (in press) have found that the calibration of probability responses associated with general-knowledge questions is relatively invariant with regard to several factors not considered here, including subjects' intelligence, subjects' expertise in the subject-matter area of the questions and subjects' reliance on the stereotypic responses of .50 and 1.00. They did, however, find that calibration varies with item difficulty.

A crucial question for generality is how well the level of item difficulty found in these experiments represents the level found in the world. Although no simple answer to this question is possible, it is worth noting that the items in Experiments 2 and 3 were not constructed with the intention of eliciting extreme overconfidence. Rather, they were constructed to vary in difficulty from very hard to very easy, as defined by the ratio of the statistical frequencies of death from each of the two causes. Items in Experiments 4 and 5 were matched to these items in difficulty.

To explain these results, we must understand both how people answer questions and how they assess the validity of their answering process. Collins (Collins, Warnock, Aiello, & Miller, 1975; Collins, Note 3) has shown that people use many different strategies in answering questions. We suspect, therefore, that extreme overconfidence can come from a variety of sources. Every answering procedure may have its own ways of leading people astray and its own ways of hiding that misguidance when people try to assess answer validity. Some possible pathways to overconfidence are described below.

Many of the items we presented to our subjects are on topics for which they do not

Table 6
Deceptive Items in Experiment 3

Causes of death compared[a]	Percent correct	No. cases of extreme overconfidence[b]
Pregnancy, abortion, childbirth versus appendicitis	15	15
All accidents versus stroke	17.5	14
Homicide versus suicide	25	12
Measles versus fireworks	25	5
Suicide versus diabetes	27.5	8
Breast cancer versus diabetes	30	1

[a] Subjects judged the first cause of death listed to be less frequent than the second.
[b] Data are the number of subjects (out of 40) who gave odds greater than or equal to 50:1 to the wrong alternative.

have a ready answer stored in memory. They must infer the answer from other information known to them. But people may be insufficiently critical of their inference processes. They may fail to ask "What were my assumptions in deriving that inference?" or "How good am I at making such inferences?" For example, when people draw a few instances of a category from memory to get an idea of the properties of the category, they may not realize that readily available examples need not be representative of the category (Tversky & Kahneman, 1973). Wason and Johnson-Laird (1972) have shown that people have considerable confidence in their own erroneous syllogistic reasoning. Collins et al. (1975) have described a variety of inferential strategies that people use in producing answers without realizing their limitations. Summarizing her studies on the inference process in perception, Johnson-Abercrombie (1960) concluded, "The[se erroneous] inferences were not arrived at as a series of logical steps but swiftly and almost unconsciously. The validity of the inferences was usually not inquired into; indeed, the process was usually accompanied by a feeling of certainty of being right" (p. 89). Pitz (1974), who also observed overconfidence in probability estimates, elaborated a similar hypothesis. He proposed that people tend to treat the results of inferential processes as though there was no uncertainty associated with the early stages of the inference. Such a strategy is similar to the "best-guess" heuristic that has been found to describe the behavior of subjects in cascaded inference tasks (e.g., Gettys, Kelly, & Peterson, 1973).

For other questions, people believe that they are answering directly from memory without making any inferences. People commonly view their memories as exact (al-

Table 7
Deceptive Items in Experiment 4

General-knowledge question[a]	Answers[b]	Percent correct	No. cases of extreme overconfidence[c]
1. Three fourths of the world's cacao comes from	Africa* or South America	4.8	15
2. Which causes more deaths in the U.S.?	Appendicitis* or pregnancy, abortion, and childbirth	19.0	13
3. When was the first air raid?	1849* or 1937	26.2	10
4. Adonis was the god of	Love or vegetation*	31.0	10
5. Kahlil Gibran was most inspired by which religion?	Buddhist or Christian*	33.3	7
6. *Dido and Aeneas* is an opera written by	Berlioz or Purcell*	33.3	2
7. Potatoes are native to	Ireland or Peru*	35.7	10

[a] Some questions have been abbreviated slightly.
[b] Correct answer carries an asterisk.
[c] Data are the number of subjects (out of 42) who gave odds greater than or equal to 50:1 to the wrong alternative.

though perhaps faded) copies of their original experiences. However, considerable evidence has demonstrated that memory is more than just a copying process (e.g., Neisser, 1967). According to this view, people reach conclusions about what they have seen or what they remember by reconstructing their knowledge from fragments of information, much as a paleontologist infers the appearance of a dinosaur from fragments of bone. During reconstruction, a variety of cognitive, social, and motivational factors can introduce error and distortion into the output of the process. Examples of this are the foibles of eyewitness testimony documented by Buckhout (1974), Loftus (1974), Münsterberg (1908), and others.

If people are unaware of the reconstructive nature of memory and perception and cannot distinguish between assertions and inferences (Harris & Monaco, in press), they will not critically evaluate their inferred knowledge. In general, any process that changes the contents of memory unbeknownst to people will keep them from asking relevant validity questions and may lead to overconfidence. In his classic studies of reconstructive processes in memory, Bartlett (1932) found that subjects not only created new material but were often highly certain about that which they had invented.[2]

We present these ideas more as a framework for future research and conceptualization than as an explanation for our results. Nonetheless, if these speculations have some validity, it should be possible to find apparent examples in our data. Tables 6 and 7 present the most deceptive items from Experiments 3 and 4, respectively. Although cases of extreme overconfidence were distributed over most items, these "deceptive" items produced a disproportionate share. In the absence of detailed protocols from subjects, these cases where many people went astray may provide better clues to our intuitions than situations where just one or two subjects had trouble with an item.

Looking at the deceptive items in Experiment 3 (see Table 6), we find that in many cases the cause of death incorrectly judged to be more frequent (the first one listed in each pair) is a dramatic, well-publicized event, whereas the underestimated cause is a more "quiet" killer. Considering the first three examples, (a) pregnancy, abortion, and childbirth, (b) accidents, and (c) homicide seem disproportionately more newsworthy and better reported than their comparison cause of death.[3] In these cases, people may be relying on the greater availability in memory of examples of the "flashier" causes of death without realizing that availability is an imperfect inferential rule (Tversky & Kahneman, 1973).[4] Other items suggest other answering processes. Subjects' confident—but erroneous—beliefs (not shown in Table 6) that there were fewer deaths from smallpox vaccine than from the disease itself may have been based on the generally valid assumption that vaccines are safer than the diseases they are meant to prevent. With smallpox, however, the vaccine has been so successful that no one has died of the disease in the U.S. since 1949, while from 6 to 10 people have died

---

[2] An example of the subtle role of assumptions in the reconstruction of knowledge comes from the experience of one of the authors who became embroiled in a friendly debate with a colleague about the dates of a forthcoming conference. Both parties agreed that the conference was to last about 4 to 5 days. But the dispute centered about whether these dates were March 30 to April 3 or April 30 to May 3. The author was certain of the former dates because he specifically recalled the date March 30 in the organizer's letter. His colleague was certain of the latter period because he specifically recalled the date May 3 in the letter. Bets were placed, and the letter was consulted to resolve the dispute. To the surprise of both parties, the letter stated the dates as March 30 to May 3, an obvious mistake. Thus, both parties were correct regarding the fragment of information they recalled, but one fragment led to the wrong conclusion.

[3] This speculation has been empirically affirmed by Lichtenstein, Slovic, Fischhoff, Combs, and Layman (Note 2).

[4] Subjects in Experiment 3 were asked to select one answer about which they were certain and to write a short statement indicating why they were so confident. One subject explained odds of 2,000:1 that death from pregnancy was more frequent than deaths by appendicitis by writing "I've never heard of a person dying of appendicitis, but I have many times heard of persons dying during childbirth and abortion."

annually from complications arising from vaccination.

For the general-knowledge questions of Experiment 4 in Table 7, we will give a few interpretations of the varied ways that unrecognized or inadequately questioned assumptions can obscure the tenuousness of erroneous beliefs; the reader can surely provide others. Regarding Item 1, cacao is native to South America. Subjects who knew this fact (or guessed it from the Spanish-sounding name) may have been misled by assuming that the continent of origin is also the continent of greatest production. Similar reasoning may have been involved with Item 7. The potato's prominence in Irish history does not mean that it originated there. Regarding Item 3, it may not have occurred to subjects that an air raid could be conducted by balloons, which were used by Austria to bomb Venice in 1849. The fact that Adonis was a handsome youth who had an affair with Venus, the Goddess of Love, may have suggested that he, too, was a diety of love (Item 4). And so on.

Finally, let us add a warning that extreme overconfidence cuts both ways. Our sources for the answers to general-knowledge questions were a variety of encyclopedias and dictionaries. We viewed the answers they provided with great confidence. Much to our chagrin, we discovered on several occasions that these authorative sources disagreed, a possibility we had never considered. Fortunately, our own overconfidence was discovered before conducting these experiments; the offending items were deleted and the remaining ones double- and triple-checked until we were *certain* of their accuracy.

Reference Notes

1. Fischhoff, B., & Lichtenstein, S. *The effect of response mode and question format on calibration* (Rep. No. 77-1). Eugene, Oregon: Decision Research, 1977.
2. Lichtenstein, S., Slovic, P., Fischhoff, B., Combs, B., & Layman, M. *Perceived frequency of low probability, lethal events* (Rep. No. 76-2). Eugene, Oregon: Decision Research, 1976.
3. Collins, A. *Processes in acquiring knowledge* (Rep. No. 3231). Cambridge, Mass.: Bolt Beranek and Newman, Inc., 1976.

References

Bartlett, F. C. *Remembering.* Cambridge, England: Cambridge University Press, 1932.

Buckhout, R. Eyewitness testimony. *Scientific American,* 1974, *231,* 23–31.

Collins, A., Warnock, E. H., Aiello, N., & Miller, M. Reasoning from incomplete knowledge. In D. Bobrow & A. Collins (Eds.), *Representation and understanding.* New York: Academic Press, 1975.

Gettys, C. F., Kelly, C., III, & Peterson, C. R. The best guess hypothesis in multistage inference. *Organizational Behavior and Human Performance,* 1973, *10,* 364–373.

Harris, R. J., & Monaco, G. E. Psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology: General,* in press.

Johnson-Abercrombie, M. L. *The anatomy of judgment.* New York: Basic Books, 1960.

Lichtenstein, S., & Fischhoff, B. Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance,* in press.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungerman & G. de Zeeuw (Eds.), *Decision making and change in human affairs.* Amsterdam, The Netherlands: Reidel, in press.

Loftus, E. The incredible eyewitness. *Psychology Today,* December 1974, pp. 116–119.

Münsterberg, H. *On the witness stand.* New York: Doubleday, Page, 1908.

Neisser, U. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

Pitz, G. Subjective probability distributions for imperfectly known quantities. In G. W. Gregg (Ed.), *Knowledge and cognition.* New York: Wiley, 1974.

Tversky, A., & Kahneman, D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology,* 1973, *5,* 207–232.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science,* 1974, *185,* 1124–1131.

Wason, P. C., & Johnson-Laird, P. N. *Psychology of reasoning: Structure and content.* Cambridge, Mass.: Harvard University Press, 1972.

Received December 14, 1976 ■