Journal of Experimental Psychology:
Human Learning and Memory

VOL. 6, NO. 2 MARCH 1980

Reasons for Confidence

Asher Koriat
University of Haifa, Haifa, Israel

Sarah Lichtenstein and Baruch Fischhoff
Decision Research, a Branch of Perceptronics
Eugene, Oregon

People are often overconfident in evaluating the correctness of their knowledge. The present studies investigated the possibility that assessment of confidence is biased by attempts to justify one's chosen answer. These attempts include selectively focusing on evidence supporting the chosen answer and disregarding evidence contradicting it. Experiment 1 presented subjects with two-alternative questions and required them to list reasons for and against each of the alternatives prior to choosing an answer and assessing the probability of its being correct. This procedure produced a marked improvement in the appropriateness of confidence judgments. Experiment 2 simplified the manipulation by asking subjects first to choose an answer and then to list (a) one reason supporting that choice, (b) one reason contradicting it, or (c) one reason supporting and one reason contradicting. Only the listing of contradicting reasons improved the appropriateness of confidence. Correlational analyses of the data of Experiment 1 strongly suggested that the confidence depends on the amount and strength of the evidence supporting the answer chosen.

One remarkable characteristic of human memory is its knowledge of its own content. Judgments of confidence in the correctness of recall and recognition performance are moderately valid predictors of that performance (Murdock, 1966; Tulving & Thomson, 1971). Even when unable to recall an item from memory, people can estimate with some accuracy whether the item will be retrieved or recognized in subsequent tests (Blake, 1973; Gruneberg & Monks, 1974; Hart, 1965). Although some systematic distortions in evaluating one's own knowledge have been noted (Koriat & Lieblich, 1977), on the whole, people seem able to monitor successfully the contents of their memories. This ability has been regarded as important, if not indispensable, to the effective operation of a memory system (Hart, 1967).

This research was supported by the Advanced Research Projects Agency of the Department of Defense, and was monitored by Office of Naval Research under Contract N00014-79-C-0029 (ARPA Order No. 3668) to Perceptronics, Inc.

Our thanks to Mark Layman and Barbara Combs for their help, and to Helmut Jungermann, Henry Montgomery, and Paul Slovic for comments on an earlier draft.

Requests for reprints should be sent to Sarah Lichtenstein, Decision Research, a branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401.

Somewhat different conclusions have been recently emphasized by investigators in the area of decision making. In these studies (reviewed by Lichtenstein, Fischhoff, & Phillips, 1977), confidence judgments were elicited as assessments of the probability that a statement is true. Appropriateness of confidence was measured by comparing these assessed probabilities with the observed relative frequencies of being correct (*hit rates*). An individual is *well calibrated* if, over the long run, for all answers assigned a given probability, the proportion correct equals the probability assigned. The general conclusion from these studies is that people are rather poorly calibrated. Although higher probabilities are typically associated with larger hit rates, in an absolute sense the probabilities and hit rates diverge considerably. The major systematic deviation from perfect calibration is overconfidence, an unwarranted belief in the correctness of one's answers (Lichtenstein et al., 1977). Typical results have shown, for example, an observed hit rate of .6 associated with assessed probabilities of .70, whereas for answers assigned probabilities of .90, only about 75% are correct. Recent studies using a variety of question and response formats found that answers endorsed with absolute certainty (i.e., $p = 1.00$) were wrong about 20% of the time (Fischhoff, Slovic, & Lichtenstein, 1977). Subjects had enough faith in these expressions of certainty that they were willing to risk money on them.

Although overconfidence has been a pervasive bias in a variety of tasks, existing experimental data leave the underlying psychological process(es) quite unclear. Building on some speculations by Fischhoff et al. (1977) and Koriat and Lieblich (1977), the present work looks at the origins of unwarranted certainty. To assess one's confidence in the truth of a statement, one first arrives at a confidence judgment based on internal cues or "feelings of doubt" (Adams & Adams, 1961). The judgment is then transformed into a quantitative expression, such as the probability that the statement is correct. Unwarranted certainty might be linked either to the confidence judgment or to the translation of that judgment into a number.

The monotonic relationship between assessed probability and percentage correct indicates that people can monitor the correctness of their answers with some success. Thus, miscalibration may be due primarily to inappropriate translation of those feelings into probabilistic terms. Such translation difficulties might be correctable by proper training. In fact, a calibration training program that provided calibration feedback after each of 11 sessions of 200 items was found to be effective by Lichtenstein and Fischhoff (in press). Their finding that almost all the improvement was accomplished after the first round of feedback suggests that what subjects learned was to make simple adjustments in the magnitude of the subjective probabilities they reported.

These results, however, do not preclude the possibility that miscalibration arises from the way in which information is used and evaluated when making confidence judgments rather than from the translation of these judgments into probabilities. The fact that the improvement in Lichtenstein and Fischhoff's (in press) subjects failed to generalize to probability assessment tasks with the same response mode but different content suggests that translation problems are not the whole story.

One information-processing mechanism that would produce overconfidence is to rely more heavily on considerations consistent with a chosen answer than on considerations contradicting it. Such a predisposition could express itself either during memory search and retrieval, by gradually biasing the search toward evidence supporting a tentatively preferred answer, or in a postdecisional stage in which the evidence is reviewed and confidence is assessed. Whichever is the case, forcing subjects to write down as much pertinent evidence as possible before choosing an answer or evaluating its validity should reduce the selective bias in the utilization of evidence and result in an improved calibration.

In the first experiment to be reported, subjects were given two-alternative general-knowledge questions. For each question they were asked to choose the correct alternative and state the probability that their choice was correct. The debiasing manipulation presented an additional task: to write down all possible reasons that argue for or against both alternatives and to rate the strength of each reason. We anticipated that a balanced survey of the pertinent evidence would reduce overconfidence and improve calibration. To the extent that confidence judgments are biased in favor of the selected alternative, the amount of evidence supporting that answer should be a better predictor of confidence than the amount of evidence inconsistent with it.

Experiment 1

Method

Stimuli. Six sets of 10 questions each were selected from the 150 general-knowledge questions used in Experiment 3 of Lichtenstein and Fischhoff (1977). The sets were closely matched in terms of item difficulty, as measured by the percentage of subjects answering each correctly in that experiment. The range of item difficulties used was from 32% to 97%, with the mean for each of the six sets varying from 64.2% to 64.9%. Five additional questions from the same pool were used for warm-up. The questions covered a wide variety of topics including history, literature, geography, and nature. All had a two-alternative format. For example, "the Sabines were part of (a) ancient India or (b) ancient Rome."

Subjects. The subjects were 73 paid volunteers who responded to an ad in the University of Oregon student newspaper. Each subject attended one of two identical group sessions.

Procedure. Each subject answered two of the six sets of questions, the first under *control* instructions and the second under *reasons* instructions. The control instructions directed subjects to choose the correct alternative for each question and to judge the probability that the chosen alternative was correct by writing down a probability between .5 and 1.0. Fifteen questions then followed, the first 5 of which were treated as warm-up and discarded in analyses.

Instructions for the reasons condition directed the subject to

> spell out all the possible reasons that you can find favoring and opposing each of the answers. Such reasons may include facts that you know, things that you vaguely remember, assumptions that make you believe that one answer is likely to be correct or incorrect, gut feelings, associations and the like.

Subjects were required to write each reason in the appropriate cell of a 2 × 2 table depending on whether it spoke for Answer a, for Answer b, against Answer a, or against Answer b. They were urged to provide reasons for all four cells of the table and to formulate each statement in a manner that conveyed their degree of certainty in it (e.g., "I know for sure. . . ," "I vaguely remember . . ."). Finally, Subjects were asked to rate each reason on a 7-point scale according to how strongly it spoke for or against the corresponding answer (1 was labeled "weakest possible"; 7 was labeled "strongest possible").

Each of the 10 questions in the reasons condition appeared on a separate page. The two possible answers were placed above the columns of a large 2 × 2 table; the words "reasons for" and "reasons against" labeled the rows. A space was provided at the bottom of each page to choose an answer and provide a probability assessment. The 10 pages of questions were shuffled before the compilation of each booklet.

For the first group of subjects, 3 min were allowed for each question. Since some subjects found this procedure annoying, in the second session only the first question was timed. Subjects were instructed to continue working on the remaining items at their own pace, spending an average of 3 min on each item.

Results

Five subjects did not provide probability assessments for all the questions or used the probability scale inappropriately. Their data were excluded from the following analyses. Each of the six sets of questions was answered by 9–13 of the remaining 68 subjects in the control condition and 10–12 subjects in the reasons condition. Since preliminary analyses revealed no systematic differences between the data of the two sessions, they were combined in the following analyses.

The appropriateness of confidence judgments can be represented by a calibration curve showing the percentage of correct responses associated with each subjective probability value. With perfect calibration, probability equals percentage correct and the calibration curve becomes the identity line. With overconfidence, the curve lies under the identity line; with underconfidence, it lies over the identity line.

Figure 1 presents the calibration curves for all subjects and items for the control and reasons conditions.[1] The interval around each point represents the 50% credible interval using uniform priors (Phillips, 1973). The calibration curve for the reasons condition is clearly superior to that of the control condition. The control condition displays the overconfidence typically observed in previous studies. Some overconfidence is also observed in the reasons condition, but it is mostly confined to probability assessments of .9 and 1.0. Table 1 presents the proportions of correct responses and mean probability assessments for the two conditions as well as several measures of the quality of the probability responses.

The over/underconfidence measure reported in Table 1 is the signed difference between the mean assessed probability and the overall proportion correct. The next entry in Table 1, the Brier score (Brier, 1950), is a strictly proper quadratic scoring rule which serves as a general measure of the quality of probability assessments. The lower the score, the better the performance; a perfect score of 0 can be earned only by always choosing the correct answer and assigning to it the probability of 1.0. The score can go as high as 1; a score of .25 will be earned by always choosing an alternative at random and assigning it a probability of .5. The three measures following the Brier score in Table 1 are the three additive partitions of the Brier score developed by Murphy (1973). The knowledge component is a function of proportion correct and is ideally 0. The calibration score is the weighted mean of the squared differences between the data points in Figure 1 and the identity line. Resolution reflects the slope of calibration curves. This term is subtracted from knowledge and calibration to get the Brier score; thus, the larger it is, the better. (For further details on these measures, see Lichtenstein et al., 1977.)

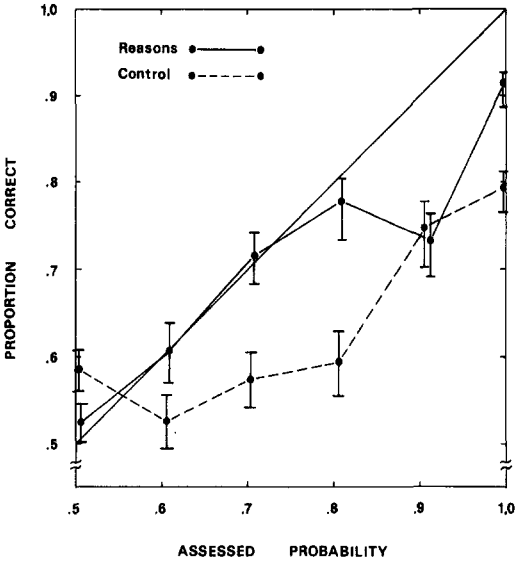

Figure 1. Calibration curves for the control and reasons conditions (Experiment 1) with 50% credible intervals.

Table 1
Summary Indices for the Reasons and Control Conditions

Indices	Control	Reasons	t
Proportion correct	.629	.669	1.65
Mean probability assessments	.720	.697	2.34*
Over/under confidence	+.091	+.028	2.36*
Brier total	.246	.209	2.81**
Knowledge	.233	.221	1.63
Calibration	.022	.005	2.69**
Resolution	.009	.018	1.64

* $p < .05$. ** $p < .01$.

These measures are extremely unstable when based on small samples, such as the 10 responses given by each subject in each condition. Thus each measure in Table 1 was calculated over the data from all subjects. A modified jackknifing analysis (Mosteller & Tukey, 1968) was used to compare the two conditions. The indices listed in Table 1 were calculated for the control and reasons data 68 times, each time leaving out a different subject. The results of t-test comparisons of the control and reasons conditions over the 68 reduced samples are presented in the last column of Table 1.

[1] The assessed probabilities were grouped in categories .5 to .59, .6 to .69, . . . , .9 to .99, and 1.0.

As can be seen, the calibration and Brier scores for the reasons condition are sigficantly superior to those obtained in the control condition. In fact, although the calibration observed for the control condition is fairly typical for items with the present level of difficulty, the calibration for the reasons condition is one of the best we have ever observed. This improvement in calibration came about through both a decrease in confidence (the subjects used .5 responses more frequently and 1.0 responses less frequently, with little change in the use of intermediate values) and an increase in the proportion of correct choices. The net result was a significant reduction in overconfidence from .091 to .028 and an improvement in calibration.

Discussion

Calibration improved in the reasons condition in the absence of any feedback regarding the adequacy of the probability judgments. This strongly suggests that cognitive biases in the assessment of uncertainty (and not merely inappropriate translation) are involved in the overconfidence observed in our control condition and in previous research.

The degree of improvement in calibration achieved by listing reasons is roughly comparable to that achieved in Lichtenstein and Fischhoff's (in press) training study. With items of comparable difficulty, the initial calibration of their 12 subjects improved from .015 to .005,[2] overconfidence was reduced from .063 to .020.

Examination of the reasons supplied and their rated strength will be postponed until the results of Experiment 2 have been discussed.

Experiment 2

The results of Experiment 1 are consistent with the hypothesis that overconfidence is the result of neglecting evidence that contradicts the chosen answer. Under this hypothesis, writing down both supporting and contradicting reasons should have led subjects to a more balanced weighing of the evidence. If supporting reasons are recruited naturally, then the requirement to spell out contradicting reasons is the key to the improved calibration in the reasons condition.

Experiment 2 was designed to evaluate this interpretation by contrasting the effectiveness of instructions to write down contradicting reasons with that of instructions to write down supporting reasons. The task was further simplified by requiring subjects to write only one or two reasons. Three conditions were used, each of which had subjects first choose the correct answer, then produce one or two reasons and finally assess the probability of being correct. In the *supporting* condition, subjects were instructed to write down one reason speaking for the selected alternative; in the *contradicting* condition, they were asked to write one reason speaking against it; in the *both* condition, they were to write one reason of each type. If our hypothesis is correct, then writing reasons in response to the supporting instructions should produce no improvement in calibration, since those instructions roughly simulate what people normally do. By drawing attention to contrary reasons, both the contradicting and both conditions should improve calibration.

Experiment 2 was also designed to provide some information regarding two alternative explanations of the results of Experiment 1. The first involves order effects. Since the reasons condition followed the control condition for all subjects in Experiment 1, it might be argued that the improved calibration is due to some kind of learning or practice effect. Experiment 2 examined this possibility by systematically manipulating the order of the items within all conditions. A second alternative interpretation is that requiring subjects to work harder induced a more serious attitude toward the task and increased the subjects'

[2] Calibration scores are generally larger when based on smaller amounts of data. Since these are the means of calibrations computed on just 200 responses, whereas the calibrations in Table 1 were computed on 680 responses, they are not strictly comparable.

motivation. If so, in Experiment 2 there should be little difference in calibration between the supporting and contradicting conditions and somewhat better calibration in the both condition.

Method

Stimuli. The same 6 sets of 10 questions used in Experiment 1 were employed in Experiment 2. All subjects received three sets under regular (control) instructions and three sets under one of the three experimental instructions.

All the questions were compiled in a booklet that also contained all instructions. The control instructions appeared first, followed by 35 questions, the first 5 of which were discarded in analysis. The remainder of the booklet contained the supporting, contradicting, or both instructions followed by the remaining 30 items. The order of the sets was systematically varied so that within each condition all six sets appeared approximately equally often in each ordinal position.

Subjects. Subjects were 200 paid volunteers who responded to an ad in the University of Oregon student newspaper. Each subject attended one of five identical group sessions. There were 66 subjects in the supporting condition, 66 in the contradicting condition, and 68 in the both condition.

Procedure. The instructions for the control conditions were the same as those used in Experiment 1. The instructions for the three experimental conditions all required the subject first to choose the correct alternative, then to write down one or two reasons, and finally to assess the probability that the chosen alternative was correct. The conditions differed only in the type of reasons called for.

In the supporting condition, subjects were instructed to

> write down in the space provided one reason that supports your decision. Please write the *best* reason you can think of that either speaks for or provides evidence for the alternative you have chosen, or speaks against or points against the alternative you rejected.

The contradicting instructions were similarly phrased except that they called for the best contradicting reason. The both instructions required one reason of each type. The rest of the instructions were similar to those used in the reasons condition of Experiment 1, specifying what might constitute a reason, and encouraging the use of statements conveying degrees of certainty in the reason provided and indicating the source of the evidence. Unlike the reasons condition, however, no strength ratings were required. The task was self-paced.

Results

Order effects. Since each of the six sets of 10 questions appeared with approximately equal frequency in each of the six ordinal positions, the possibility of order effects could be evaluated by comparing the calibration curves for the first, second, and third successive sets of 10 questions over all subjects, and again for the fourth, fifth, and sixth. These comparisons revealed no systematic trends within either of the two blocks of three sets. The calibration indices for the first, second, and third sets (the control condition) were .014, .017, and .014, respectively. Pooling over all three experimental conditions, calibration

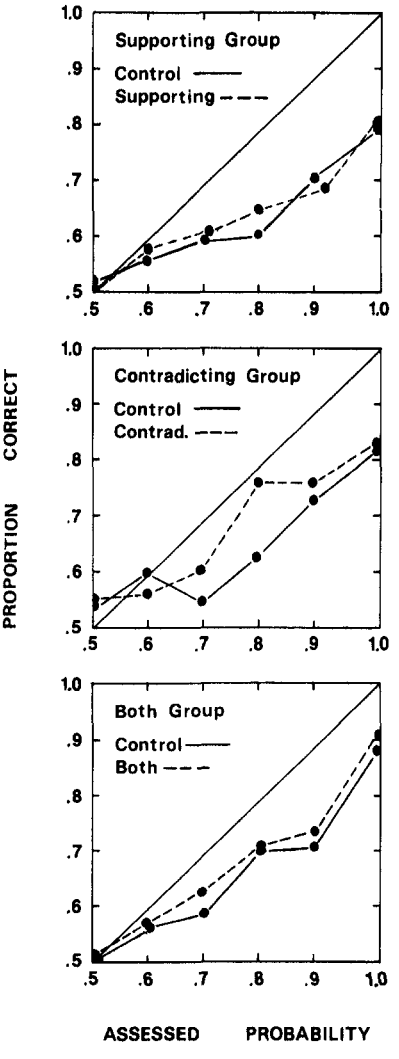

Figure 2. Calibration curves for the control and experimental conditions for the supporting, contradicting, and both groups (Experiment 2).

Table 2
Summary Indices for Experiment

Indices	Supporting (n = 66)			Contradicting (n = 55)			Both (n = 68)		
	Control	Reasons	t	Control	Reasons	t	Control	Reasons	t
Proportion correct	.624	.622	.14	.638	.653	.76	.645	.646	.06
Mean probability assessments	.722	.716	.87	.713	.704	.88	.712	.703	1.22
Over/under confidence	+.098	+.094	.22	+.074	+.052	1.19	+.067	+.057	.61
Calibration	.020	.018	.36	.015	.009	2.10*	.012	.006	1.24

*$p < .05$.

indices for the fourth, fifth, and sixth sets of items were .011, .010, and .012, respectively. It appears safe to conclude that the improvement found in Experiment 1 cannot be attributed to the order in which the control and reasons conditions were administered.

Manipulation checks. In evaluating the effects of the three experimental manipulations, 11 subjects had to be eliminated from the contradicting group because they wrote only or mostly supporting reasons. In fact, judging both from the behavior of subjects during the experiment and from the nature of the reasons offered, subjects apparently found it harder to produce a single contradicting reason than either a single supporting reason or both a supporting and a contradicting reason. Even some of the subjects retained in the contradicting group occasionally used supporting reasons. These errors seemed to reflect either momentary lapses or changes of mind as to which alternative was chosen, indicated by erasing or crossing out of a previous answer (without a search for a reason contradicting that new answer).

Calibration. The calibration curves in Figure 2 clearly show that the strongest improvement was achieved by the contradicting instructions. Producing both kinds of reasons resulted in a slight but systematic improvement, whereas the supporting instructions had no effect.

Jackknifing analyses similar to those employed in Experiment 1 were used to evaluate the differences in calibration between the experimental conditions and their respective controls. A summary of the results appears in Table 2. For the supporting group, the means of the supporting and control conditions were almost identical for each of the indices investigated. Although statistically insignificant, the effects of the both instructions were in the direction of improved calibration.

The contradicting instructions, on the other hand, resulted in a significant improvement in calibration.[3] Although the changes in proportion correct and mean probability were slight, their pattern is similar to that observed in the reasons condition of Experiment 1: increased proportion correct combined with reduced confidence.

Discussion

The results of the contradicting group are consistent with the idea that overconfidence derives in part from the tendency to neglect contradicting evidence and that calibration may be improved by making such evidence more salient.

Although significant, the improvement in calibration achieved by the contradicting instructions was not as marked as that obtained by the reasons instructions of Experiment 1. This may have been because the reasons instructions required a more

[3] This difference remained significant even when the contradicting subjects who failed to follow the instructions were included in the analysis. Thus the different results found for the supporting and contradicting groups could not be attributed to the exclusion of some of the subjects from the contradicting group.

thorough and detailed analysis of the pertinent evidence than the contradicting instructions with its minimal requirement of specifying a single reason.

The fact that supporting instructions had no effect on performance suggests that producing a supporting reason is approximately what people normally do when asked to assess the likelihood that an answer is correct.

The failure of the both condition to produce a significant effect was unexpected. If supporting reasons have no effect and contradicting reasons improve calibration, then the net effect of both should be better calibration, assuming that the two effects are additive. The substance of the reasons the subjects supplied, however, casts doubt on this assumption. Contradicting reasons provided in the both condition were often of a somewhat different character than those in the contradicting condition. Specifically, the former were seldom independent of the supporting reason provided and were sometimes secondary to it (e.g., by qualifying it or doubting its validity). Manipulating the order in which the supporting and contradicting reasons were solicited would probably not have helped, as subjects could easily follow their natural predilection and write down the supporting reason first anyway. Another possible explanation is that in Experiment 1 subjects were asked to list reasons before choosing an answer, whereas in Experiment 2 they were asked to choose an answer before listing reasons. To see if this change in order could explain the difference between the results of Experiment 1 in which calibration did improve and the both condition of Experiment 2 in which it did not improve, we ran a new group of 44 subjects in a modified both condition with the order reversed. After responding to 35 control items, subjects were given 30 more items and asked to list one reason for each of the two alternatives:

In the space for "a reason supporting answer *a*," please write the *best* reason you can think of that either

1. speaks for or provides evidence for alternative *a*, or
2. speaks against alternative *b*.

In the space for "a reason supporting answer *b*," please write the *best* reason you can think of that either

1. speaks for or provides evidence for alternative *b*, or
2. speaks against alternative *a*.

After writing the two reasons, the subjects selected one alternative and gave a probability. The instructions were otherwise the same as in the both condition of Experiment 2. This new both/reversed condition produced no improvement in calibration over its own control, suggesting that the order in which subjects gave reasons and chose an answer made no appreciable difference. Of course, there was no way to prevent the subject from covertly choosing an alternative before listing reasons, even when the instructions and format suggested otherwise.

The failure of the both condition to produce stronger improvement than the contradicting condition would also argue against interpreting the reasons effect in Experiment 1 as the result of increased effort. An effort hypothesis would have also predicted improved calibration with the supporting instructions, an effect that was totally absent.

Analysis of Reasons

The subjects in the reasons condition of Experiment 1 were asked to list all pertinent reasons arguing for or against each of the two alternative answers and to rate the strength of each reason on a 7-point scale. They then chose an answer and assessed the likelihood of its being right. The reasons were analyzed to gain further insight into the role of evidence in assessing confidence.

Six subjects who had failed to provide strength ratings or had used the rating scale inappropriately (e.g., used a .5–1.0 scale or provided strength ratings for the question as a whole) were eliminated from the following analyses. The remaining 62 subjects supplied an average of 3.17 reasons for each question. No subject provided more than four reasons for any single cell.

The reasons data were analyzed in terms of the type of evidence provided, the relationship between this evidence and the chosen alternative, and the relationship between the reasons and confidence.

The interrogation of memory for pertinent evidence. A content analysis of the reasons may offer valuable insight into the type of evidence people seek when choosing alternatives and assessing confidence (Collins, Warnock, Aiello, & Miller, 1975). This task was not attempted with the present data because of the great variety of substantive questions used. We did, however, examine the distribution and strength of *reasons for* and *reasons against.* This distinction is independent of the alternative chosen, unlike the distinction between supporting and contradicting reasons, which is defined in relationship to the chosen alternative. The former contrast may be most relevant to the process of memory interrogation, whereas the latter seems more pertinent to the evidence review stage.

Two aspects of the data suggest that when interrogating their memories for pertinent evidence, subjects are tuned more toward arguments *for* than toward arguments *against.* First, subjects gave more reasons *for* than reasons *against.* The average subject provided at least one reason *for* for 14.7 of the 20 alternatives while providing at least one reason *against* for only 12.3 alternatives, t (61) = 6.04, $p < .001$. Overall, subjects produced an average of 18.3 reasons *for* and 13.4 reasons *against*, t (61) = 7.99, $p < .001$. Second, reasons *for* were assigned higher strength ratings than reasons *against.* The mean rating given the first reason *for* was 3.8; the mean rating of the first reason *against* was 3.5, t (61) = 2.92, $p < .005$. The same pattern was found for ratings averaged across all *for* reasons and all *against* reasons given.

Choice of answer. The choice of an alternative was more highly related to *for* than to *against* reasons. Pooling over all subjects and items, there was a .61 point-biserial correlation between choice of alternative and the difference between the number of reasons *for* given to the two alternatives (i.e., many more reasons *for* were associated with chosen than with rejected alternatives). The comparable correlation between choice of alternative and the difference between number of reasons *against* was only .42.

A similar trend was found when predicting the choice of an alternative from the differences in the strengths of the first *for* and *against* reasons supplied to each of the alternatives. This analysis used only those instances in which *for* or *against* reasons were supplied to both alternatives. The choice of a particular alternative correlated .64 (n = 313) with the differences in the strength of the first *for* reasons (i.e., people were much more likely to choose the alternative with the higher strength assigned to the first reason *for*) and .29 (n = 241) with the differences in the strength of the first *against* reasons.

Assessment of confidence. In the following analyses, the reasons that the subjects supplied were classified according to their relationship to the alternative chosen as (a) *for chosen*, (b) *against rejected*, (c) *for rejected*, or (d) *against chosen.* The first two categories constitute supporting reasons and the last two categories constitute contradicting reasons.

Consider first the number of reasons produced. Pooling over all subjects and items (n = 620), the correlations between confidence and number of reasons were as follows: .28 with reasons *for chosen*, .17 with reasons *against rejected*, −.06 with reasons *for rejected*, and −.07 with reasons *against chosen.* Thus confidence was more strongly related to the number of supporting reasons than to the number of contradicting reasons. Notice that within the supporting category, the *for* reasons appear to have slightly more weight than the *against* reasons, a trend that is more pronounced in the strength data presented next.

Figure 3 shows the mean assessed probability as a function of the strength of the first reason, for each of the four categories of reasons (a strength of zero means that no reason of that type was given). The assessed probabilities were most strongly

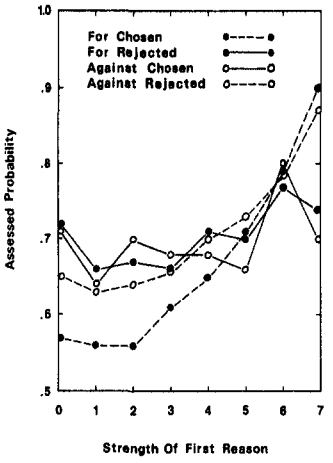

Figure 3. Mean assessed probability as a function of the strength of the first reason for four types of reasons (Experiment 1).

related to the strength of *for chosen* reasons; with strengths less than 3, the mean probability was below .6, whereas for strength ratings of 7, the mean probability was .9. The correlations between assessed probability and the strength of the first reason for the data included in Figure 3 were as follows: .66 for reasons *for chosen*, .39 for reasons *against rejected*, .00 for reasons *for rejected*, and −.02 for reasons *against chosen*. Here, too, the greater relevance of supporting reasons (and to a lesser extent reasons *for*) is apparent.

Figure 3 ignores the number of reasons supplied. The sum of the strengths of all the reasons produced seems to capture better the impact of all evidence for or against an answer. Confidence correlated .56 with the sum of strengths of reasons *for chosen*, .37 for reasons *against rejected*, .03 for reasons *for rejected*, and .00 for reasons *against chosen*. Thus, the amount of contradicting evidence (*for rejected*, *against chosen*) appears to have no bearing on confidence. Once again, positive evidence (favoring the selected alternative) seems to count more than negative evidence (against the rejected alternative).

General Discussion

The present studies seem to shed some light on the process by which confidence is determined and on the origins of the most pervasive bias in calibration, overconfidence. In understanding the results, it is helpful to conceptualize the confidence assessment task as having two cognitive stages. The first stage involves searching one's knowledge; this stage ends when an answer is chosen. During the second stage, the evidence is reviewed and confidence in the chosen alternative is assessed.

We have shown that calibration can be significantly improved by requiring people to explicate all considerations that seem pertinent to their decision. The results suggest two biases in how people elicit and use their own knowledge, one bias corresponding to each cognitive stage.

The first bias involves favoring positive rather than negative evidence (i.e., reasons *for* over reasons *against*). Evidence of this bias was found in Experiment 1: Subjects produced more reasons *for* than reasons *against*, reasons *for* were given higher strength ratings, and both the number and strength ratings of reasons *for* were better predictors of the chosen answer.

The procedures of Experiment 1 may have induced or augmented this bias. Forcing subjects to choose the correct alternative may have focused their attention on reasons *for*. Conceivably, a bias toward reasons *against* might be exhibited with instructions to choose the incorrect alternative. Although the two types of instructions are logically equivalent (for two-alternative forced-choice items), they may not be psychologically equivalent. Another way of studying the possibility that format affects the tendency to produce positive evidence would be to present to four different groups the four variations of naturally dichotomous true/false items (e.g., gazpacho soup is served hot/ . . . is served cold/ . . . is not served hot/ . . . is not served cold). A general preference for positive evidence would produce a different pattern of reasons across the four variations than would a bias induced by the positive or negative wording of the items.

The bias against negative evidence found here is similar to the difficulties people have in accepting the relevance of negative

evidence in logical inference tasks (Johnson-Laird & Wason, 1977) as well as the neglect of negative examples in judgments of correlation (e.g., Smedslund, 1963).

The second bias in confidence assessment is a tendency to disregard evidence inconsistent with (contradictory to) the chosen answer. Particularly striking evidence for this bias came from Experiment 2. Asking subjects to write a supporting reason did not affect their calibration (presumably because they were already thinking of these reasons), whereas asking them to write a contradicting reason did. Although writing a contradicting reason did improve the realism of subjects' confidence assessments, the subjects found the task difficult; far more instances of producing no reason, or a wrong reason, were observed under the contradictory instructions than under the supporting instructions.

The correlational analyses of Experiment 1 showed that whatever measure was used (the number of reasons produced, the strength of the first reason, or the sum of the strengths of all the reasons), the assessment of confidence was heavily based on the evidence supporting the answer chosen, and not on the evidence supporting the rejected reason. These results give rise to a seeming paradox: How can it be that forcing people to give contradicting reasons leads to improved calibration (as shown in Experiment 2), yet correlational analysis reveals no relationship between the strength or number of contradicting reasons and the probabilistic responses? This paradox is more apparent than real. The results may be understood by supposing that the requirement to give contradictory reasons reduces one's overall confidence, so a difference in calibration is found *between* conditions. The correlational data, however, reflect relationships *within* the reasons condition. Apparently, once in that condition, even while feeling less confident in general, one still relies on supporting rather than contradicting evidence to select one's relative degree of confidence for specific items.

It is unclear whether this biased approach to evidence affects only the confidence assessment stage or also the earlier recruitment of information in choosing an answer. One could imagine an unbiased search for information that quickly generated a predilection for one of the choices. Subsequent search is then directed toward considerations supporting the favored alternative. In this light, the reasons manipulation succeeded by forcing a more systematic exploration of pertinent considerations before choosing an answer, as well as by lowering confidence during the postdecisional review process.

Again, we must be aware that the format used might have contributed to the bias. Asking subjects to give the probability that their chosen answer was incorrect could conceivably lead them to rely on contradicting (*for rejected* and *against chosen*) reasons when assessing confidence.

We do not fully understand why the both condition in Experiment 2 did not improve calibration. It may be that the bias toward supporting evidence is so strong that explicitly eliciting even one supporting reason establishes the bias unless, as in Experiment 1, all possible contradicting reasons are sought. This would explain why the contradicting reasons supplied by the both group were often subsidiary to the supporting reasons.

Although further research is clearly needed, we can derive some practical advice from the present results. People who are interested in properly assessing how much they know should work harder in recruiting and weighing evidence. However, that extra effort is likely to be of little avail unless it is directed toward recruiting contradicting reasons.

This message echoes a recurrent conclusion of research designed to reduce cognitive biases. Working harder will have little effect unless combined with a task restructuring that facilitates more optimal cognitive functioning. For example, Fischhoff (1977) found that hindsight bias was not reduced by mere exhortation to work harder or even explicit warning about its possible presence. It was, however, greatly reduced by having subjects write a short statement regarding how they would have

explained the occurrence of an event that did not happen (Slovic & Fischhoff, 1977), a manipulation not unlike the present enforced search for contradictory reasons.

References

Adams, J. K., & Adams, P. A. Realism of confidence judgments. *Psychological Review*, 1961, *68*, 33–45.

Blake, M. Prediction of recognition when recall fails: Exploring the feeling of knowing phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 311–319.

Brier, G. S. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950, *78*, 1–3.

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

Fischhoff, B. Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 349–358.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 522–564.

Gruneberg, M. M., & Monks, J. "Feeling of knowing" and cued recall. *Acta Psychologica*, 1974, *38*, 257–265.

Hart, J. T. Memory and the feeling of knowing experience. *Journal of Educational Psychology*, 1965, *56*, 208–216.

Hart, J. T. Memory and the memory monitoring process. *Journal of Verbal Learning and Verbal Behavior*, 1967, *6*, 685–691.

Johnson-Laird, P. N., & Wason, P. C. A theoretical analysis of insight into a reasoning task. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking*. Cambridge, England: Cambridge University Press, 1977.

Koriat, A., & Lieblich, I. A study of memory pointers. *Acta Psychologica*, 1977, *41*, 151–164.

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*, 1977, *3*, 552–564.

Lichtenstein, S., & Fischhoff, B. Training for calibration. *Organizational Behavior and Human Performance*, in press.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. deZeeuw (Eds.), *Decision making and change in human affairs*. Amsterdam: D. Reidel, 1977.

Mosteller, F., & Tukey, J. W. Data analysis including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*. Reading, Mass.: Addison-Wesley, 1968.

Murdock, B. B. The criterion problem in short-term memory. *Journal of Experimental Psychology*, 1966, *72*, 317–324.

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

Phillips, L. D. *Bayesian statistics for social scientists*. London: Nelson, 1973.

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

Smedslund, J. The concept of correlation in adults. *Scandinavian Journal of Psychology*, 1963, *4*, 165–173.

Tulving, E., & Thomson, D. M. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology*, 1971, *87*, 116–124.

Received April 13, 1979
Revision received August 6, 1979 ■