

Subjective Confidence in Forecasts

BARUCH FISCHHOFF AND DON MACGREGOR
Decision Research, A Branch of Perceptronics, Oregon, U S A

ABSTRACT

Forecasts have little value to decision makers unless it is known how much confidence to place in them. Those expressions of confidence have, in turn, little value unless forecasters are able to assess the limits of their own knowledge accurately.

Previous research has shown very robust patterns in the judgements of individuals who have not received special training in confidence assessment: Knowledge generally increases as confidence increases. However, it increases too swiftly, with a doubling of confidence being associated with perhaps a 50 per cent increase in knowledge. With all but the easiest of tasks, people tend to be overconfident regarding how much they know.

These results have typically been derived from studies of judgements of general knowledge. The present study found that they also pertained to confidence in forecasts. Indeed, the confidence-knowledge curves observed here were strikingly similar to those observed previously. The only deviation was the discovery that a substantial minority of judges never expressed complete confidence in any of their forecasts. These individuals also proved to be better assessors of the extent of their own knowledge.

Apparently confidence in forecasts is determined by processes similar to those that determine confidence in general knowledge. Decision makers can use forecasters' assessments in a relative sense, in order to predict when they are more and less likely to be correct. However, they should be hesitant to take confidence assessments literally. Someone is more likely to be right when he or she is 'certain' than when he or she is 'fairly confident'; but there is no guarantee that the certain forecast will come true.

KEY WORDS Probability Confidence Forecasting Subjective judgements
 Calibration Judgemental bias

Since the destruction of the Second Temple, prophecy has become the lot of fools.
 Hebrew expression

What constitutes a wise forecaster? It is not just someone who is usually correct; that definition would give undue deference to those who make forecasts about predictable events. It is not just

This research was supported by the Office of Naval Research under contract N00014-80-C-0150 to Perceptronics, Inc. We would like to thank Gerry Hanson and Mark Layman for their help in various parts of this enterprise.

0277-6693/82/020155-18\$01 80

Received November 1981

© 1982 by John Wiley & Sons, Ltd.

someone who is seldom proven wrong; that definition would reward the makers of vague and unverifiable forecasts. It is not just someone who provides a confident message with clear implications for action; that definition would promote arrogance over thoughtfulness.

If one is to take action on the basis of a forecast, perhaps the most desirable property is that it be appropriately qualified. That is, one wants to know how much faith to put in it. One measure of the appropriateness of expressions of faith in forecasts is their degree of *calibration*.¹ For the sake of calibration, all statements of fact are considered to carry with them an implicit or explicit expression of confidence in their truth. When that expression is given quantitative form, the archetypal statement of fact has the form 'the probability that statement *A* is true is *X*.' Statement *A* may refer to a discrete event (my bank account is overdrawn), or a continuous one (the balance in my bank account is between -\$100 and \$150). It could refer to the past (George Washington died because of poor medical treatment), present (the capital of Saudi Arabia is Mecca), or future (Quebec will be a part of Canada on 1 January 2000). Only statements about the future represent forecasts, but the evaluation of all such expressions of confidence is similar. Except for situations in which an individual is 100 per cent confident and wrong, it is hard to validate single expressions. However, one can take a set of statements and see if *X* per cent of those assigned an *X* per cent chance of being correct prove to be correct, once the truth of the statement can be ascertained. The truth of forecasts can be checked by seeing whether the predicted events occur.

The Bayesian, or subjectivist, view of probability underlying calibration studies assumes that probabilities represent an individual's state of knowledge. Hence, it makes sense to aggregate probabilities over a diverse set of statements and see how well, in general, an individual assesses the extent of his or her knowledge.

Crude retrospective assessments of calibration may be derived from looking at the confidence expressions accompanying the performance of real tasks. Thus, one might find evidence of overconfidence in professions that make confident judgements with no demonstrated validity (e.g. predictions of stock price movements [Dreman, 1979; Slovic, 1972], psychiatric diagnoses of dangerousness [Cocozza and Steadman, 1978]). Unfortunately, such evidence is not only imprecise, but also ambiguous whenever 'experts' are consulted (and paid) as a function of the confidence they inspire, suggesting that they may be tempted to misrepresent how much they know (Armstrong, 1978).

Among real-world studies, the greatest efforts to ensure candour and explicitness have been with weather forecasters, who are rewarded for good calibration. Their performance is superb (e.g. Murphy and Winkler, 1974, 1977). Whether this success is due to training in calibration or a byproduct of their general professional education is unclear. A review of other studies with experts who have not had calibration training suggests that such training, and not just education, is the effective element. Experiments, using problems drawn from their respective areas of expertise but isolated from real-world pressures, have found overconfidence with psychology graduate students (Lichtenstein and Fischhoff, 1977), bankers (Staël von Holstein, 1972), clinical psychologists (Oskamp, 1962), executives (Moore, 1977), civil engineers (Hynes and Vanmarcke, 1976), and untrained professional weather forecasters (Root, 1962; Staël von Holstein, 1971).

Overconfidence is also the predominant result of experiments using non-experts responding to general-knowledge questions (Lichtenstein, Fischhoff and Phillips, in press). Exhibit 1 provides a summary of studies that have attempted to eradicate overconfidence by a variety of manipulations

¹ An alternative use of 'calibration' found in the forecasting literature is 'to estimate the relationships (and constant terms) in a forecasting model' (Armstrong, 1978, p. 477). In addition, several other terms are at times used to describe the calibration of probability assessments (see Lichtenstein, Fischhoff and Phillips, in press).

Strategies	Studied by
Faulty tasks	
Unfair tasks	
Raise stakes	1, 28
Clarify instructions/stimuli	3, 10, 12, 13, 20
Discourage second guessing	12, 20
Use better response modes	12, 13, 19, 21, 22, 30, 32, 33?, <u>34</u> , 37?
Ask fewer questions	15
Misunderstood tasks	
Demonstrate alternative goal	13
Demonstrate semantic disagreement	3, 13, 18, 28?
Demonstrate impossibility of task	12
Demonstrate overlooked distinction	14?
Faulty judges	
Perfectible individuals	
Warn of problem	12
Describe problem	3
Provide personalized feedback	<u>20</u>
Train extensively	<u>1</u> , <u>2</u> , <u>4</u> , <u>16</u> , <u>20</u> , <u>24</u> , <u>25</u> , <u>29</u> , <u>32</u>
Incorrigible individuals	
Replace them	—
Recalibrate their responses	2, 5, 23
Plan on error	—
Mismatch between judges and task	
Restructuring	
Make knowledge explicit	17
Search for discrepant information	<u>17</u>
Decompose problem	—
Consider alternative situations	—
Offer alternative formulations	33?
Education	
Rely on substantive experts	<u>11</u> , 15, 19, 23, 27, 31, 35, 36
Use easier questions	<u>8</u> , <u>9</u> , <u>22</u> , <u>26</u> , <u>29</u> , <u>30</u>
Educate from childhood	<u>6</u> , <u>7</u>

Note: Each number represents a separate article. Manipulations that have proven at least partially successful are underlined. Those that have yet to be subjected to empirical test or for which the evidence is unclear are marked by a question mark. Details in Fischhoff (in press).

Key

- | | |
|---|--|
| 1 Adams and Adams (1958) | 20 Lichtenstein and Fischhoff (1980) |
| 2 Adams and Adams (1961) | 21 Lichtenstein, Fischhoff and Phillips (in press) |
| 3 Alpert and Raiffa (in press) | 22 Ludke, Stauss and Gustafson (1977) |
| 4 Armelius (1979) | 23 Moore (1977) |
| 5 Becker and Greenberg (1978) | 24 Murphy and Winkler (1974) |
| 6 Beyth-Marom and Dekel (in press) | 25 Murphy and Winkler (1977) |
| 7 Cavanaugh and Borkowski (1980) | 26 Nickerson and McGoldrick (1965) |
| 8 Clarke (1960) | 27 Oskamp (1962) |
| 9 Cocozza and Steadman (1978) | 28 Phillips and Wright (1977) |
| 10 Dawes (in press) | 29 Pickhardt and Wallace (1974) |
| 11 Dowie (1976) | 30 Pitz (1974) |
| 12 Fischhoff and Slovic (1980) | 31 Root (1962) |
| 13 Fischhoff, Slovic and Lichtenstein (1977) | 32 Schaefer and Borcharding (1973) |
| 14 Howell and Burnett (1978) | 33 Seaver, von Winterfeldt and Edwards (1978) |
| 15 Hynes and Vanmarcke (1976) | 34 Selvidge (1980) |
| 16 King, Zechmeister and Shaughnessy (in press) | 35 Staël von Holstein (1971) |
| 17 Koriat, Lichtenstein and Fischhoff (1980) | 36 Staël von Holstein (1972) |
| 18 Larson and Reenan (1979) | 37 Tversky and Kahneman (1974) |
| 19 Lichtenstein and Fischhoff (1977) | |

Exhibit 1. Attempts to explain or reduce overconfidence

including changing the response mode, offering detailed instructions, raising the stakes hinging on good calibration, and varying the heterogeneity of the items being judged. Each paper is represented by a number which is underlined if the manipulation seemed to improve calibration. From this large set of studies, only three procedures seem to be effective. One is extensive training with personalized feedback. The second is forcing respondents to list reasons why the statement or answer they believe in might be wrong (Koriat, Lichtenstein and Fischhoff, 1980; Study No. 18 in Exhibit 1). The third, and least interesting, is to provide easier tasks. One reflection of people's insensitivity to how much they know is the fact that their mean confidence changes relatively slowly in response to changes in the difficulty of the tasks they face. Thus, when tasks become easier, people's confidence does not rise commensurately, leaving them underconfident for the easiest of tasks. In this light, the preponderance of overconfidence in the literature reflects, in part, the (perhaps natural) tendency not to present people with very easy questions.

The subjective interpretation of probability makes no distinction between confidence in statements about the future and confidence in any other kind of statement. Hence, from a formal perspective, one would expect that the results summarized in Exhibit 1 could be generalized to the calibration of forecast probabilities. That is, one could expect to find overconfidence that is impervious to most of the various manipulations described there. Formal equivalence is not, however, the same as psychological equivalence. One might speculate, for example, that all other things being equal, people are less confident in their knowledge about the future because no one

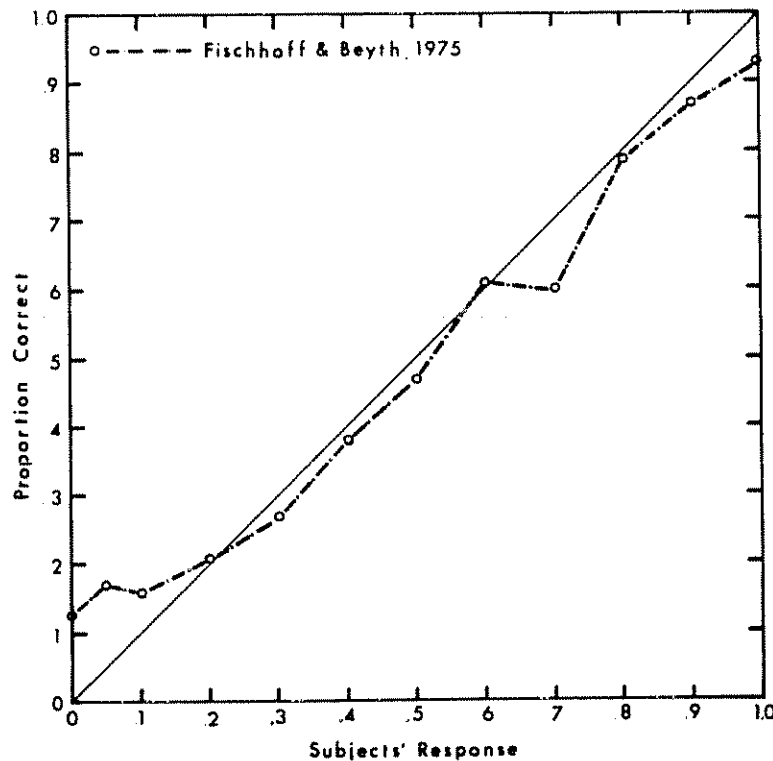


Exhibit 2 Calibration of confidence judgements for forecasts regarding possible outcomes of President Nixon's trips to China and the U.S.S.R. Source: Fischhoff and Beyth (1975)

knows about the future. Or one might speculate that they are more confident because no one can prove them wrong at the moment of prediction.

A study by Fischhoff (1976) found no difference in judgements of the likelihood of hypothetical events set in the future, present, or past. However, the hypotheticality of those events may have weakened some pertinent psychological processes. The studies involving predictions cited above (e.g. Murphy and Winkler, 1977; Root, 1962) also follow the general patterns observed in non-prediction studies (i.e. overconfidence except with easy tasks or extensive, personalized training). One intriguing possible exception to these patterns is seen in Exhibit 2, showing a study by Fischhoff and Beyth (1975) in which participants assessed the probability of various possible outcomes of President Nixon's trips to China and the U.S.S.R. (e.g. he will meet with Chairperson Mao). At the extremes here, one sees the usual overconfidence. About 10 per cent of the events that respondents were 100 per cent certain would happen, failed to happen; about 10 per cent of those that had 0 per cent chance of happening did happen. Nonetheless, over most of the range, subjects were quite well calibrated. An unpublished study by Wright and Wisudha (1979) showed less overconfidence with forecasts than with assessment of general-knowledge questions; unfortunately, the forecast questions were also less difficult, suggesting that ease might have been responsible for the difference in calibration.

Reviewing this evidence, anyone interested in eliciting and interpreting expressions of confidence in forecasts or in training forecasters to make such assessments is probably best off assuming that probability assessments for forecasts are no different than those for other problems. The present study attempts to increase or decrease the confidence with which that assumption can be made by studying confidence in forecasts using tasks that are as similar as possible to those used in studies of calibration with general-knowledge questions.

STUDY 1

The most widely used task in calibration studies is the half-range two-alternative question. Given an item with two alternative answers, one of which is guaranteed to be true (e.g. absinthe is (a) a precious stone; (b) a liqueur), the respondent must first select the answer that seems more likely to be correct and then assess the probability of that choice being the correct one. Because the more likely answer was to have been chosen, that probability should come from the upper half of the probability range: [0.5, 1.0]. Exhibit 3 shows some typical results observed in studies using such tasks. With all but the easiest tasks, one finds overconfidence, represented by calibration curves resting predominantly under the identity line that would reflect perfect calibration. Being under the identity line means that the percentage of correct answers associated with a particular expressed probability of being correct is smaller than that probability. In such figures, responses are grouped into the intervals [0.50, 0.59], [0.60, 0.69], [0.70, 0.79], [0.80, 0.89], [0.90, 0.99], and [1.00].

The one notable exception to this pattern was the study by Koriat, Lichtenstein and Fischhoff (1980) in which overconfidence was reduced (although not altogether eliminated) by having respondents provide a reason why each of their chosen answers might be incorrect. Exhibit 4 shows the effect of this contradicting reason manipulation along with the non-effect of two related manipulations. (In the exhibit, each group's performance on the experimental task is contrasted with its performance on a set of control items for which no reasons were given.) The supporting-reason group provided one reason why their chosen answer might be correct; the both-reasons groups gave one supporting and one contradicting reason. The absence of an effect with those groups indicated that the contradicting reason groups' calibration had not improved simply as a result of the additional labour involved in writing a reason.

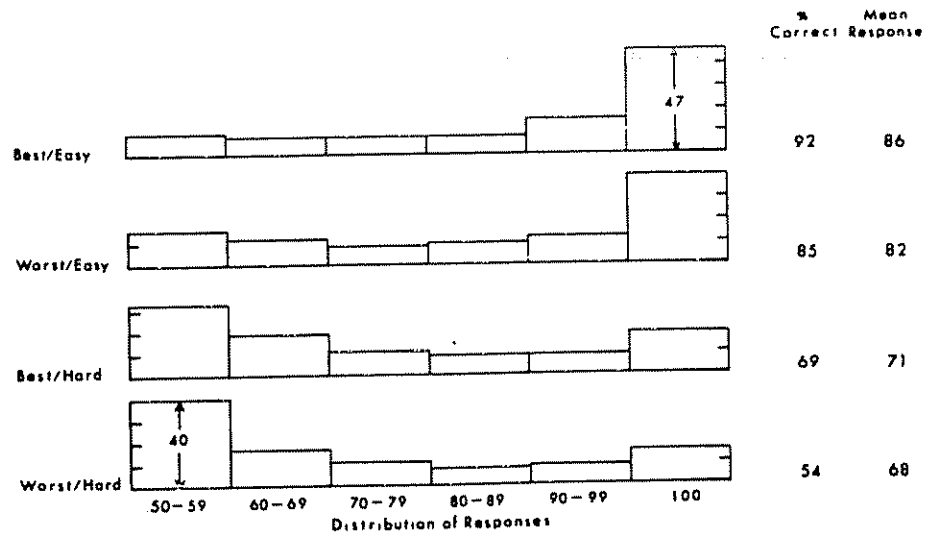
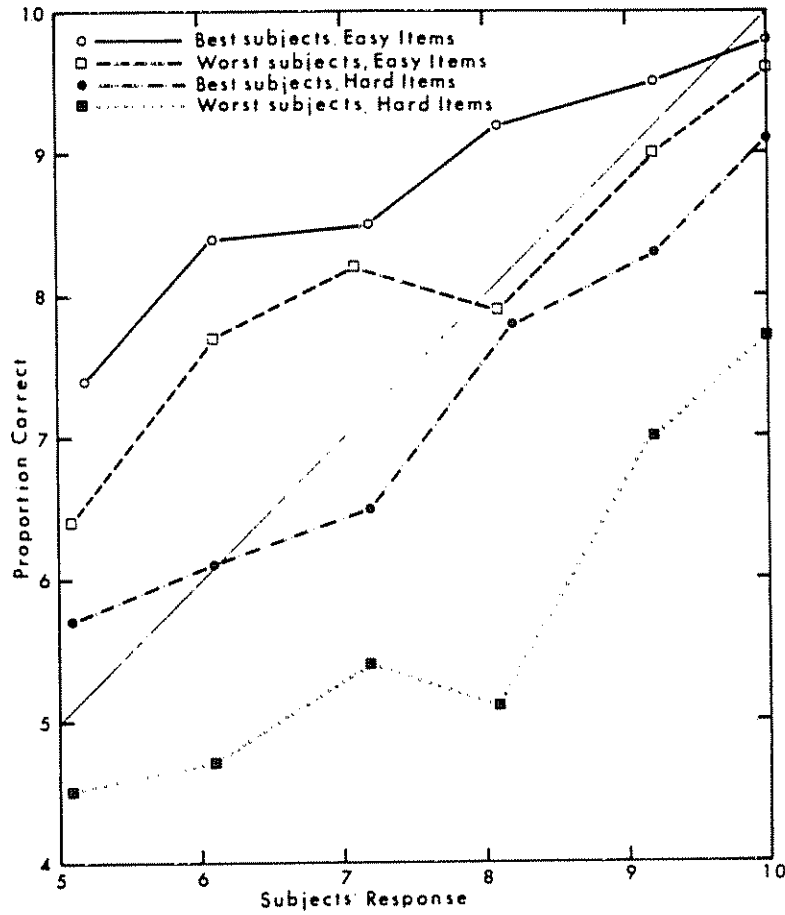


Exhibit 3. Representative calibration curves derived from studies using two-alternative, half-range tasks. Source: Lichtenstein and Fischhoff (1977)

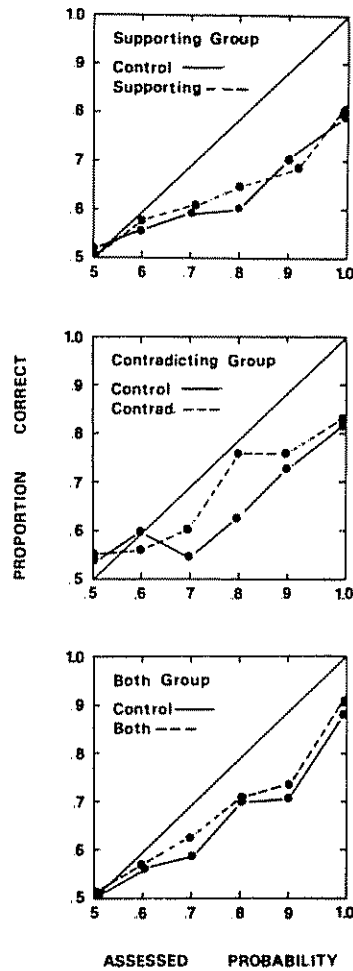


Exhibit 4 Calibration curves for individuals providing supporting, contradicting, or supporting and contradicting reasons. Each group's calibration is compared with their own performance on a set of control items Source: Koriat, Lichtenstein and Fischhoff (1980)

Study 1 replicates the three conditions of Koriat, Lichtenstein and Fischhoff (1980), using items involving future events.

Method

Design

Each participant responded to 50 two-alternative half-range questions, picking the answer most likely to be correct in each and then assigning it a probability (from 0.5 to 1.0) of being correct. The first 25 items were done using standard assessment techniques. For each of the last 25 items, after respondents had selected an answer, and prior to providing a probability, they were required to provide a reason supporting their answer, a reason contradicting it, or a reason of each type. Details may be found in Koriat, Lichtenstein and Fischhoff (1980). A no-reasons group responded to all 50 items without providing reasons.

Stimuli

Fifty items were created concerning events that would be consummated within 30 days of the time of the experiments. Some dealt with upcoming local elections (e.g. the mayor of Eugene will be (a) Gus Keller; (b) Catherine Lauris); others dealt with sporting events (e.g. who will win the following baseball game: (a) Detroit Tigers (b) California Angels (home team)); others dealt with a variety of topics. These items were separated into two sets so that items dealing with topics that seemed at all related would not appear consecutively or, to the extent possible, in the same set. Each set was used in the control condition for half of one group and in the experimental condition for the other half.

Subjects

One hundred and twelve individuals were recruited through an advertisement in the University of Oregon student paper. They were paid \$7 for completing this task as one part of a 1½-hour session. Subjects recruited in this manner typically are about half male and half female with an average age of 23. Most are involved with the university community; about 2/3 are students. They treat the tasks in a diligent manner, perhaps akin to a proctored exam. We had hoped to have a larger number of subjects, however, good weather and the proximity of final exams seem to have kept numbers down. In all, there were 32 people in the supporting reasons group, 28 in the contradicting reasons group, 26 in the both-reasons group and 26 in the group that never gave any reasons.

Results

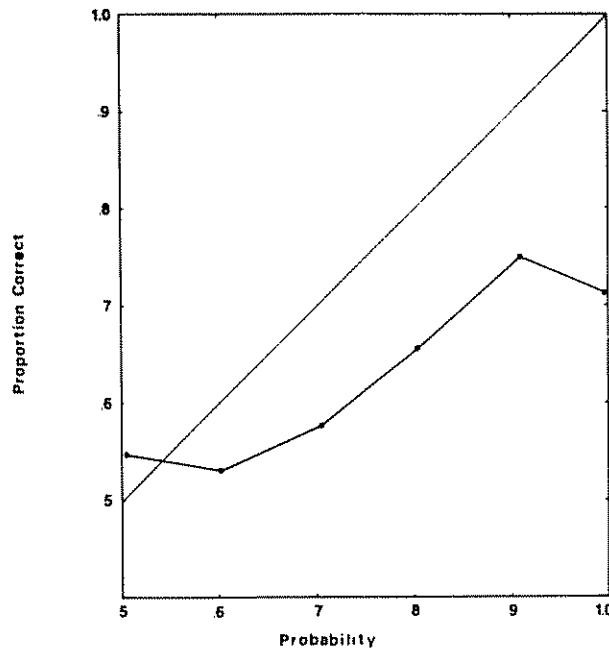
Main effect

The items we constructed proved to be fairly difficult for subjects, with the proportion of correct forecasts over all 3477 responses in the control conditions equal to only 0.618. Associated with these items was a mean confidence of 0.722. The usual measure of over- or underconfidence is the difference between these two statistics. Here it equals +0.104, indicating that subjects' percentage of correct predictions (in the control condition) should have been higher by 10.4 per cent if their level of confidence was to be justified. The calibration curve corresponding to these responses appears in Exhibit 5. Respondents' overconfidence is reflected by the fact that most of the curve falls below the identity line. The generally positive slope of the curve indicates that subjects tended to be more knowledgeable when they were more confident. Its flatness, relative to the identity line, indicates that their knowledge did not rise as quickly as did their confidence. This curve pertaining to forecasts looks strikingly like that observed with general knowledge questions of the same difficulty level (e.g. the bottom curves in Exhibit 3).

Reasons manipulation

Exhibit 6 contrasts each group of subjects' calibration on the experimental condition with their own performance on the control condition. Thus, it is comparable to Exhibit 4 from Koriat, Lichtenstein and Fischhoff (1980). As a rough guide to the stability of these curves, in each, there are approximately 100 (± 30) responses involved in determining the proportions correctly associated with probabilities of 0.6, 0.7, 0.8, and 0.9. If these were all independent responses, that would mean a standard error of estimate of approximately 0.05; however, subjects typically contributed several responses to each point. Approximately 175 (± 30) responses were associated with 0.5 and about 60 (± 30) with 1.0.

The clearest conclusion to be drawn from Exhibit 6 is that there are few, if any, systematic differences between the control and experimental conditions for any group. The supporting reasons group, which showed no change at all in the Koriat *et al.* study, seems to have improved somewhat, however, even these differences seem small relative to statistical variability. The



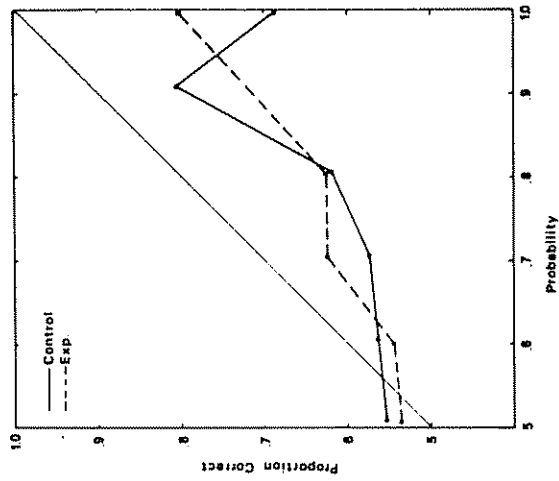
ALL CONTROL STUDY 1

Exhibit 5 Calibration of all responses to control items in Study 1. Curve includes 3447 responses produced by 112 individuals

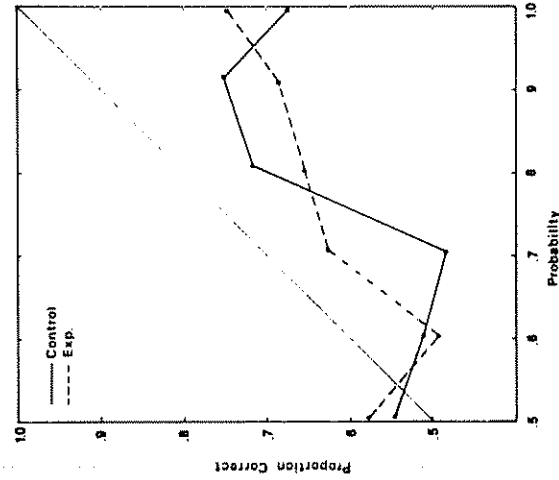
performance of these groups in the control and experimental conditions is summarized several ways in Exhibit 7. Here, we find that the experimental manipulation had little effect on the confidence of supporting or contradicting reasons subjects (slightly increasing it for the former, slightly decreasing it for the latter), but it reduced the mean confidence of both-reasons subjects from 0.724 to 0.663. This last change would have cut the overconfidence that those subjects showed in the control condition by 2/3 were there not a concomitant drop in their proportion of correct responses (from 0.626 to 0.599). All in all, each of the three groups was somewhat less overconfident in their respective experimental conditions. This modest improvement is also reflected in the group calibration scores shown in Exhibit 7. This score, derived from the partition of the Brier proper scoring rule (see Lichtenstein, Fischhoff and Phillips, in press), reflects the squared distance between the calibration curve and identity line, weighted by the number of responses at each point. It decreases as calibration improves, becoming zero with perfect calibration.

In Exhibit 6, each group's performance on the reasons task was compared to their own performance on the control (no reasons) task. Although such within-subject comparisons allow greater sensitivity of analysis, they greatly reduce the number of responses involved in each comparison. If one pools all responses to control questions (as in Exhibit 5), there appears to be slight improvement in each experimental condition, particularly with the both and contradicting reasons group.

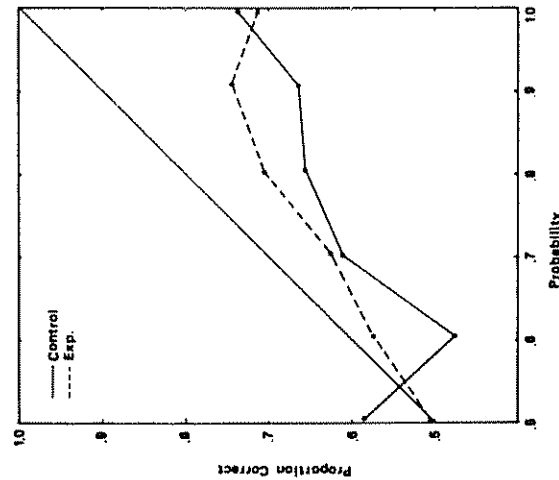
The preceding analyses assume that the experimental manipulations were uniformly effective. Koriat *et al.* discovered a moderate percentage of items for which reasons were either missing or inappropriate. This was particularly true with the contradicting reasons group, who often gave



BOTH REASONS - STUDY 1



CONTRADICTION - STUDY 1



SUPPORTING - STUDY 1

Exhibit 6. Calibration curves for individuals providing supporting, contradicting or both kinds of reasons in Study 1. Corresponding summary statistics appear in Exhibit 7

Group	Control						Experiment				
	<i>N</i>	<i>n</i>	Prop. cor	Mean conf.	Over-conf.	Calib.	<i>n</i>	Prop. cor	Mean conf.	Over-conf.	Calib.
No reasons	26	1299	0.625	0.724	0.099	0.0227	—	—	—	—	—
Supporting	32	800	0.608	0.726	0.118	0.0264	798	0.639	0.735	0.096	0.0166
Contradicting	28	698	0.607	0.711	0.104	0.0239	685	0.616	0.702	0.086	0.0211
Both	26	650	0.626	0.724	0.098	0.0225	643	0.599	0.663	0.064	0.0123
All	112	3447	0.618	0.722	0.104	0.0271	2126	0.619	0.703	0.083	0.0160

Exhibit 7 Summary statistics for Study 1

supporting reasons. Each of the present groups omitted reasons for approximately 10 per cent of all items. When supporting reasons subjects gave reasons, they were almost always appropriate to the task (99 per cent of the time). On the other hand, 11 per cent of the contradicting reasons subjects' reasons were inappropriate, constituting either supporting reasons or vague statements such as 'Maybe I'm wrong'. For both-reasons subjects, 5 per cent of their supporting reasons were inappropriate, compared with 9 per cent of their contradicting reasons. As in Koriati *et al.*, providing contradicting reasons appears to be a difficult or unnatural task. The total number of these missing and inappropriate responses was not large enough that their elimination changes the calibration curves of Exhibit 6 appreciably.

Distribution of responses

As mentioned earlier, on the control questions, subjects made roughly equal use of all the responses 0.6, 0.7, 0.8, and 0.9: they used 0.5 somewhat more, 1.0 somewhat less. Distributions for the experimental conditions were quite similar except for a slight increase in 0.5's and decrease in 1.0's. This tendency was particularly marked in the both-reasons group, whose proportion of 0.5's increased from 0.234 to 0.364 and whose proportions of 1.0's dropped from 0.112 to 0.048, thus accounting for its reduced overall confidence.

Exhibit 8 shows another aspect of response usage, the percentage of subjects who expressed confidence of 1.0 in at least one of their 25 forecasts. In previous studies with general knowledge questions, typically all or almost all subjects have used 1.0. The fact that only 77 per cent of all subjects did so on the control task suggests some tendency not to express extreme certainty in forecasts. This tendency was highlighted in the experimental tasks, where even fewer subjects used 1.0, particularly for the contradicting and both-reasons groups. The responses of subjects who did and did not use 1.0 were pooled separately over all experimental groups. For the control conditions, non-users were appreciably better calibrated all along the calibration curve (not shown).

Study Group	1		2		3		All	
	Control	Exp.	Control	Exp.	Control	Exp.	Control	Exp.
No reasons	80.8	—	—	—	—	—	80.8	—
Supporting	75.0	65.6	73.1	63.5	81.3	74.7	77.7	69.7
Contradicting	71.4	50.0	67.4	44.2	76.3	60.5	71.6	51.4
Both	76.9	50.0	75.0	50.0	79.5	61.4	77.1	54.2
All	76.8	55.8	72.0	52.4	79.8	68.2	76.2	60.2

Exhibit 8 Usage and non-usage of 1.0 (percentage of users)

Subjects who never expressed extreme confidence were not only less confident, but also more in tune with the extent of their knowledge. With the reasons conditions, the same change was observed, but its size was smaller. Users and non-users of 1.0 had highly similar percentages of correct responses, hence differences in calibration cannot be attributed to differences in difficulty level. As can be seen from the remainder of Exhibit 8, similar patterns were observed in the following studies. Calibration curves for these studies will be reported and discussed later.

Discussion

Lost in this morass of mild and inconclusive effects is the striking main effect shown in Exhibit 5. Calibration for confidence in forecasts looks just like calibration for confidence in general knowledge, when difficulty level is controlled. These forecasters' accuracy increased with their confidence, however, it did not increase as fast. As confidence rose from 0.5 to 1.0, the corresponding proportion of correct predictions only increased from 0.5 to 0.75. Respondents also tended to be overconfident in the extent of their knowledge, getting 62 per cent of their predictions right, but having a mean probability of 0.72. The one difference that does emerge is a modest reduction in usage of 1.0. The superior calibration of subjects who never used 1.0 was a promising predictor of individual differences in calibration. Despite this overall similarity, confidence in forecasts did not, however, show the same responsiveness to the reasons manipulations observed in Koriat *et al.* There was some suggestion of improved calibration with the supporting and contradicting groups. However, the relatively small sample rendered these results somewhat ambiguous. Before reaching any firm conclusion, it seemed appropriate to increase the sample size. Because the events had already occurred by the time these analyses were completed, it was not possible to add subjects to the existing groups. Instead, a second study was run, replicating the first, but with a new set of events.

STUDY 2

Method

The design of Study 2 followed that of Study 1 except for the elimination of the no-reasons group (which completed 50 forecasts without giving any reasons) and an increase in the number of forecasts from 50 to 60. As the study was completed in late October 1980, a number of the forecasts considered the elections of the following month. A total of 143 subjects, recruited as in Study 1, participated. This number, too, was somewhat less than we had hoped for, but did allow for groups roughly 2/3 larger than in Study 1.

Results

Main effect

The difficulty of the present items in the control tasks proved to be remarkably similar to that of Study 1 (62.9 per cent correct vs. 61.8 per cent), as was subjects' means confidence (0.732 vs. 0.722). subjects' overconfidence was correspondingly almost identical (0.103 vs. 0.102).

Reasons manipulation

Exhibit 9 summarizes results for the control and experimental conditions of each of the three groups. Briefly, the only apparent effect on overconfidence was the improvement of the both-reasons group. The other two groups were essentially unchanged. The calibration curves for these groups were so similar to those from Study 1 (Exhibit 6) that they will not be shown. There were again quite a few missing and inappropriate reasons, particularly for contradictory reasons. Elimination of these responses does not, however, appreciably change the patterns shown in Exhibit 9.

Group	Control						Experiment				
	N	n	Prop cor	Mean conf	Over-conf	Calib	n	Prop cor	Mean conf	Over-conf	Calib
Supporting	52	1552	0.635	0.735	0.100	0.0136	1549	0.636	0.736	0.100	0.0180
Contradicting	43	1286	0.639	0.728	0.089	0.0134	1282	0.605	0.705	0.101	0.0152
Both	48	1439	0.614	0.733	0.119	0.0226	1424	0.637	0.706	0.070	0.0186
All	143	4277	0.629	0.732	0.103	0.0159	4255	0.627	0.717	0.090	0.0159

Exhibit 9. Summary of statistics—Study 2

Distributions of responses

Presumably reflecting the increased sample size, the distributions of the three groups' probability assessments on the control tasks were quite similar. In the reasons conditions, usage of 0.5 tended to increase for all groups, whereas usage of 1.0 decreased somewhat. As in Study 1, a substantial group of subjects never used 1.0 (see Exhibit 8). Some 28 per cent followed this pattern in the control condition and 46.9 per cent in the reasons conditions. This increase was much greater for the contradicting and both-reasons groups, over half of whose subjects never used 1.0 in the experimental conditions. The calibration curves for all subjects who did not use 1.0 showed them to be less overconfident and generally better calibrated than the remaining subjects, both for reasons and controls. As in Study 1, the task was equally difficult for users and non-users.

Discussion

The major results of Study 1 have been replicated: Calibration curves for confidence in forecasts resemble those for confidence in general knowledge questions. The reasons manipulations had at best weak effects on overall calibration. The contradicting and both-reasons manipulations did, however, again reduce usage of 1.0. In general, subjects who never used 1.0 were better calibrated than their counterparts.

The overall similarity of the present confidence judgements to those observed elsewhere is encouraging for anyone who would like to exploit that literature for the elicitation and interpretation of forecasts. For example, we would expect the training techniques that have proven effective or ineffective with general knowledge items to have similar effects on the calibration of forecasters. The difference observed here between users and non-users of 1.0 may offer an additional tool for determining how much faith to place in others' confidence assessments. The weakness of the reasons manipulations is, however, disappointing, because it suggests that a simple mechanism that has proven effective in improving calibration is not as robust as one would hope. Before writing off this procedure and discussing some possible implications of this research for forecasting, we will offer one further replication designed to strengthen the reasons manipulation.

STUDY 3**Method**

Although this study was essentially a replication of the previous two, a number of changes were introduced in order to strengthen the reasons manipulation: (a) the number of items per page was reduced from 4 to 3 in order to present a less cramped format. (b) Subjects were asked to produce

not one, but two reasons of the type required by each condition. (c) The instructions were changed to emphasize that the task involved making predictions about future events, and that descriptions of things heard or read, beliefs and associations could all be used as reasons for the predictions made. (d) Subjects were asked to make a special effort to be as complete in describing their reasons as possible. (e) Subjects were assured that sufficient time had been allotted in the experiment for them to devote thought to the task. All stimuli dealt with events whose outcome would be known during the first week of June 1981. Technical aspects of subject recruitment caused responses to be elicited on two separate dates, 15 May and 29 May, two weeks before the event period and immediately before. On 15 May, half of the participants were in each of the supporting and contradicting reasons groups. On 29 May, half were in each of the supporting and both-reasons groups. Comparisons between the corresponding supporting groups at the two times will reveal whether proximity to events has any effect on calibration beyond its effect on difficulty. One hundred and seventy-three individuals participated, with roughly equal numbers on the two dates.

Results

Timing

The proportion of correct responses to control questions was higher by 0.03 for the supporting group from 29 May than for the 15 May supporting group, perhaps due to the former's closer proximity to the events in question. The 29 May group's confidence (and overconfidence) was correspondingly higher, leaving their calibration curves quite similar. Corresponding changes were seen in the two groups' responses to the experimental condition items, except that the 29 May group was a bit less overconfident (0.074 vs 0.101). As there is no apparent reason for this anomaly, the two groups' data from the two dates will be pooled in the following analyses.

Main effect

Exhibit 10 shows the same patterns in responses to the control questions as were seen in Studies 1 and 2. Each group is somewhat overconfident. The poor calibration statistic for the contradicting reasons group, despite its relatively low overconfidence, reflects a very flat calibration curve, with only a 0.12 difference between the proportions correct associated with responses of 0.5 and 1.0.

Reasons manipulation

As indicated by Exhibit 10, the reasons manipulations slightly reduced overconfidence and slightly improved calibration for all three groups. As shown in Exhibit 8, they also reduced the usage of 1.0. All these effects were somewhat larger for the contradicting and both-reasons groups. As before, non-users of 1.0 were considerably better calibrated than users.

Group	Control						Experiment				
	<i>N</i>	<i>n</i>	Prop cor.	Mean conf.	Over-conf.	Calib.	<i>n</i>	Prop cor.	Mean conf.	Over-conf.	Calib.
Supporting	91	2625	0.650	0.746	0.096	0.0198	2610	0.657	0.745	0.088	0.0151
Contradicting	38	1098	0.655	0.724	0.069	0.0275	1086	0.656	0.706	0.051	0.0231
Both	44	1264	0.654	0.737	0.083	0.0186	1258	0.652	0.723	0.071	0.0113
All	173	4987	0.652	0.739	0.087	0.0201	4954	0.655	0.731	0.076	0.0172

Exhibit 10. Summary of statistics—Study 3

DISCUSSION

Three clear patterns have emerged from these three studies, each with some possible implications for forecasting practitioners:

1. Calibration for confidence assessments regarding forecasts is largely indistinguishable from that pertaining to general knowledge questions. The overconfidence scores and calibration curves observed with the control items here were very similar to those observed with general knowledge items of similar difficulty. On the basis of these results, one should have considerably increased confidence in extrapolating the results of earlier calibration research to confidence in forecasts. Thus, one might expect calibration for forecasts to be relatively unaffected by changes in response mode, incentive payments for correct answers, or familiarity with subject matter (unless accompanied by a change in difficulty), to generalize a few results from Exhibit 1.

2. The only apparent difference between these responses and those observed previously was the appearance of a subsample of subjects who never used 1.0. Over all three studies, such subjects constituted 23.8 per cent of the control groups and 39.8 per cent of the experimental groups. As shown in Exhibit 11, non-users of 1.0 were consistently much better calibrated than users, in all three studies, for both control and experimental items. Exhibit 12 pools responses of users and non-users across the three studies. Each curve includes 5000–15,000 responses made by 100–300 subjects. Although non-users are somewhat better calibrated for most probability values, the major difference between the groups is at 1.0. Non-users simply do not produce the point that represents the greatest overconfidence, that is, the greatest discrepancy between how often one should be correct and how often one is. On the basis of these results, one might tentatively extend greater credence to the confidence assessments of forecasters who never express complete certitude.

3. The reasons manipulations had consistent but weak effects. In each study, responses in the experimental condition were better calibrated and less overconfident than those in the corresponding control conditions. Over all three studies, overconfidence decreased by 0.008 for supporting reasons subjects (from 0.101 to 0.093), by 0.007 for contradicting reasons subjects (from 0.086 to 0.079), and by 0.032 for both-reasons subjects (from 0.101 to 0.069). In an applied situation, one might wonder if such modest improvements were worth the additional time and effort the provision of reasons requires. Of course, one might also feel that the provision of explicit reasons has desirable features independent of its impact on calibration. These could include: (a)

Study Group	1		2		3		All	
	Control	Exp	Control	Exp	Control	Exp	Control	Exp.
Prop correct								
Users	0.617	0.625	0.635	0.625	0.656	0.653	0.638	0.639
Non-users	0.620	0.613	0.614	0.629	0.636	0.660	0.623	0.636
Mean prob.								
Users	0.736	0.736	0.751	0.757	0.750	0.748	0.747	0.748
Non-users	0.675	0.660	0.686	0.673	0.685	0.685	0.683	0.674
Overconfidence								
Users	0.119	0.111	0.115	0.132	0.094	0.095	0.108	0.109
Non-users	0.055	0.047	0.072	0.047	0.049	0.025	0.060	0.038
Calibration								
Users	0.0249	0.0212	0.0197	0.0257	0.0218	0.0182	0.0215	0.0208
Non-users	0.0065	0.0081	0.0065	0.0065	0.0121	0.0079	0.0078	0.0069

Exhibit 11. Summary statistics for users and non-users of 1.0 group means

providing a record of the reasons motivating one's forecasts in order to avoid the prejudicial effects of hindsight bias when the time comes to evaluate them, once the event has or has not happened (Fischhoff, 1975); (b) allowing for external review of one's reasoning, perhaps leading to the correction of misconceptions or improved communications (Hogarth and Makridakis, 1981); or (c) helping raise one's alertness to new evidence that should prompt revisions of a forecast (Armstrong, 1978).

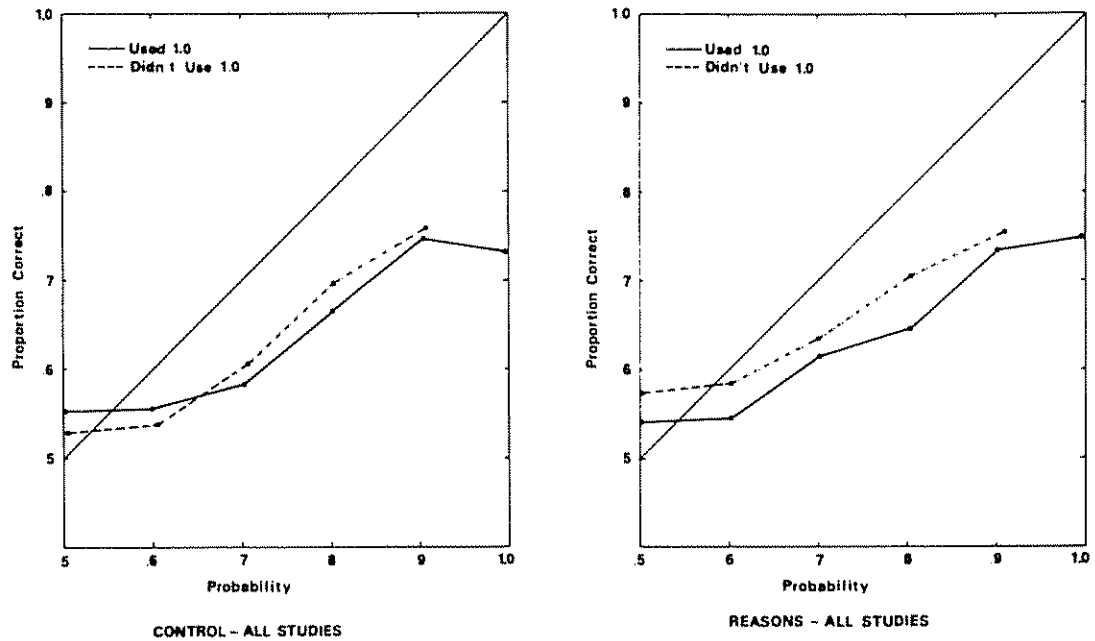


Exhibit 12 Calibration curves for users and non-users of 1.0, pooled across Studies 1-3. Corresponding summary statistics are given in Exhibit 11. Curves involve approximately 5000 to 16,000 responses produced by 100 to 300 individuals

It is worth noting in this context that the most dramatic effect demonstrated by Koriat, Lichtenstein and Fischhoff (1980) was found with a much more involved procedure than that depicted in Exhibit 4 and repeated in the present studies. In a separate experiment, they required subjects to complete a 2×2 matrix giving reasons for and against each of the two possible answers. Ten, rather than thirty, items were used in that study. That more ambitious and focused manipulation reduced confidence by 0.023, while increasing the percentage of correct answers by 0.040, thereby reducing overconfidence by 0.063. Perhaps one must conclude that provision of one or two reasons for each of a fairly large number of items cannot hurt, but it cannot be counted to help very much.

The most consistent effect of the reasons manipulations, in particular the provision of contradicting or both reasons, was to increase the proportion of subjects who never used 1.0. As mentioned, these non-users were better calibrated than users in both the control and experimental conditions. Indeed, one might speculate that the primary effect of the reasons manipulations is to convince indirectly some people never to be entirely certain.

REFERENCES

- Adams, J. K., and Adams, P. A., 'Realism of confidence judgments', *Psychological Review*, **68** (1961), 33-45.
- Adams, P. A., and Adams, J. K., 'Training in confidence judgments', *American Journal of Psychology*, **71** (1958), 747-751.
- Alpert, W., and Raiffa, H., 'A progress report on the training of probability assessors', in D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, in press.
- Armeliuss, K., 'Task predictability and performance as determinants of confidence in multiple-cue judgments', *Scandinavian Journal of Psychology*, **20** (1979), 19-25.
- Armstrong, J. S., *Long-Range Forecasting From Crystal Ball to Computer*. New York: Wiley, 1978.
- Becker, B. W., and Greenberg, M. G., 'Probability estimates by respondents: does weighting improve accuracy?', *Journal of Marketing Research*, **15** (1978), 482-486.
- Beyth-Marom, R., and Dekel, S., *Thinking under Uncertainty: A Textbook for Junior High School Students*, in press (in Hebrew).
- Cavanaugh, J. C., and Borkowski, J. G., 'Searching for meta-memory connections', *Developmental Psychology*, **16** (1980), 441-453.
- Clarke, F. R., 'Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests', *Journal of the Acoustical Society of America*, **32** (1960), 35-46.
- Cocozza, J. J., and Steadman, H. J., 'Prediction in psychiatry: an example of misplaced confidence in experts', *Social Problems*, **25** (1978), 265-276.
- Dawes, R. M., 'Confidence in intellectual judgments vs confidence in perceptual judgments', in *Essays in Honor of Clyde Coombs*, in press.
- Dowie, J., 'On the efficiency and equity of betting markets', *Economica*, **43** (1976), 139-150.
- Dreman, D., *Contrarian Investment Strategy*, New York: Random House, 1979.
- Fischhoff, B., 'Hindsight \neq foresight: the effect of outcome knowledge on judgment under uncertainty', *Journal of Experimental Psychology: Human Perception and Performance*, **1** (1975), 288-299.
- Fischhoff, B., 'The effect of temporal setting on likelihood estimates', *Organizational Behavior and Human Performance*, **15** (1976), 180-194.
- Fischhoff, B., 'Debiasing', in D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, in press.
- Fischhoff, B., and Beyth, R., "'I knew it would happen":—remembered probabilities of once-future things', *Organizational Behavior and Human Performance*, **13** (1975), 1-16.
- Fischhoff, B., and Slovic, P., 'A little learning...: confidence in multi-cue judgment', in R. Nickerson (Eds.), *Attention and Performance, VIII*, Hillsdale, N. J.: Lawrence Erlbaum, 1980.
- Fischhoff, B., Slovic, P., and Lichtenstein, S., 'Knowing with certainty: the appropriateness of extreme confidence', *Journal of Experimental Psychology: Human Perception and Performance*, **3** (1977), 552-564.
- Hogarth, R. M., and Makridakis, S., 'Forecasting and planning: an appraisal', *Management Science*, **27** (1981), 115-138.
- Howell, W. C., and Burnett, S. A., 'Uncertainty measurement: a cognitive taxonomy', *Organizational Behavior and Human Performance*, **22** (1978), 45-68.
- Hynes, M., and Vanmarcke, E., 'Reliability of embankment performance predictions', *Proceedings of the ASCE Engineering Mechanics Division Specialty Conference*, Waterloo, Ontario, Canada: University of Waterloo Press, 1976.
- King, J. F., Zechmeister, E. B., and Shaughnessy, J. J., 'Judgment of knowing: the influence of retrieval practice', *American Journal of Psychology*, in press.
- Koriat, A., Lichtenstein, S., and Fischhoff, B., 'Reasons for confidence', *Journal of Experimental Psychology: Human Learning and Memory*, **6** (1980), 107-118.
- Larson, J. R., and Reenan, A. M., 'The equivalence interval as a measure of uncertainty', *Organizational Behavior and Human Performance*, **23** (1979), 49-55.
- Lichtenstein, S., and 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*, **20** (1977), 159-183.
- Lichtenstein, S., and Fischhoff, B., 'Training for calibration', *Organizational Behavior and Human Performance*, **26** (1980), 149-171.

- Lichtenstein, S., Fischhoff, B., and Phillips, L. D., 'Calibration of probabilities: The state of the art to 1980', in D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, in press
- Ludke, R. L., Stauss, F. F., and Gustafson, D. H., 'Comparison of five methods for estimating subjective probability distributions', *Organizational Behavior and Human Performance*, **19** (1977), 162-179
- Moore, P. G., 'The manager's struggle with uncertainty', *Journal of The Royal Statistical Society, Series A*, **140** (1977), 129-165
- Morris, P. A., 'Decision analysis expert use', *Management Science*, **20** (1974), 1233-1241.
- Murphy, A. H., and Winkler, R. L., 'Subjective probability forecasting experiments in meteorology: some preliminary results', *Bulletin of the American Meteorological Society*, **55** (1974), 1206-1216
- Murphy, A. H., & Winkler, R., 'Can weather forecasters formulate reliable probability forecasts of precipitation and temperature?', *National Weather Digest*, **2** (1977), 2-9
- Nickerson, R. S., and McGoldrick, C. C., Jr., 'Confidence ratings and level of performance on a judgmental task', *Perceptual and Motor Skills*, **20** (1965), 311-316.
- Oskamp, S., 'The relationship of clinical experience and training methods to several criteria of clinical prediction', *Psychological Monographs*, **76** (1962), (28, Whole No 547)
- Phillips, L. D., and Wright, G. N., 'Cultural differences in viewing uncertainty and assessing probabilities', in H. Jungermann and G. deZeeuw (Eds.), *Decision Making and Change in Human Affairs*, Amsterdam, Reidel, 1977.
- Pickhardt, R. C., and Wallace, J. B., 'A study of the performance of subjective probability assessors', *Decision Sciences*, **5** (1974), 347-363
- Pitz, G. F., 'Subjective probability distributions for imperfectly known quantities', in L. W. Gregg (Ed.), *Knowledge and Cognition*, New York: Wiley, 1974
- Root, H. E., 'Probability statements in weather forecasting', *Journal of Applied Meteorology*, **1** (1962), 163-168.
- Schaefer, R. E., and Borcharding, K., 'The assessment of subjective probability distribution: a training experiment', *Acta Psychologica*, **37** (1973), 117-129.
- Seaver, D. A., von Winterfeldt, D., and Edwards, W., 'Eliciting subjective probability distributions on continuous variables', *Organizational Behavior and Human Performance*, **21** (1978), 379-391
- Selvidge, J., 'Assessing the extremes of probability distributions by the fractile method', *Decision Sciences*, **11** (1980), 493-502.
- Slovic, P., 'Psychological study of human judgment: implications for investment decision making', *Journal of Finance*, **27** (1972), 779-799
- Staël von Holstein, C.-A. S., 'An experiment in probabilistic weather forecasting', *Journal of Applied Meteorology*, **10** (1971), 635-645
- Staël von Holstein, C.-A. S., 'Probabilistic forecasting: an experiment related to the stock market', *Organizational Behavior and Human Performance*, **8** (1972), 139-158.
- Tversky, A., and Kahneman, D., 'Judgment under uncertainty: heuristics and biases', *Science*, **185** (1974), 1124-1131.
- Wright, G., and Wisudha, A., 'Differences in calibration for past and future events', Paper presented at the *Seventh Research Conference on Subjective Probability, Utility and Decision Making*, Göteborg, Sweden, 1979.

Authors' biographies

Baruch Fischhoff and **Don MacGregor** are with Decision Research, a branch of Perceptronics in Eugene, Oregon. Both have continuing interests in probability assessment, judgemental bias, and risk perception. Baruch, who is currently on leave at the Medical Research Council's Applied Psychology Unit at Cambridge, England, has ongoing projects dealing with value assessment, risk analysis and historical judgement. Don's projects include studies of the acceptability of decision-making methods and the perception of coincidence.

Authors' addresses:

Baruch Fischhoff, MRC/APU, 15 Chaucer Road, Cambridge, CB2 2EF, Great Britain.

Don MacGregor, Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401, U.S.A.