ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 20, 159–183 (1977)

Do Those Who Know More Also Know More about How Much They Know?

SARAH LICHTENSTEIN AND BARUCH FISCHHOFF

Decision Research, A Branch of Perceptronics

The validity of a set of subjective probability judgments can be assessed by examining two components of performance, calibration and resolution. The perfectly calibrated judge assigns probabilities so that, for all propositions assigned the same probability, the proportion true is equal to the probability assigned. For example, half of the propositions given a .50 chance of being true should in fact be true. Resolution reflects the degree to which assessors can successfully discriminate among different degrees of certainty, independent of the numerical labels assigned. A series of experiments revealed that: (1) Although people are moderately well calibrated, their probability judgments are prone to systematic biases. The most common bias is overconfidence. (2) People are calibrated differently when dealing with items of varying degrees of difficulty. (3) Calibration is unaffected by differences in intelligence, expertise, subjects' reliance on extreme probability responses, and at least some aspects of the context in which items are presented. (4) Resolution did not change as a function of difficulty, except for tasks about which subjects knew nothing. The implications of these results for decision makers are discussed.

Dealing with uncertainty is a central challenge in our day-to-day lives. In order to manage our affairs effectively, we must make predictions about the future behavior of individuals, groups, social systems, economies, and international engagements. Reflecting this situation, subjective probabilities, the numerical expression of our predictions, have found their way into psychological theories of such diverse phenomena as motivation (Feather, 1959; Weiner, 1974), attitudes (Fishbein, 1967), personality attributions (Jones & Davis, 1965), decision making (Edwards & Tversky, 1967), choice behavior (Krantz, Luce, Suppes, & Tversky, 1974), and gambling (Cohen, 1960).

We would like to thank Ward Edwards, Lewis Goldberg, Cam Peterson, Gordon Pitz, and Paul Slovic for their perceptive comments on various stages of this project; Barbara Combs for conducting the experiments and creating the large item pool used in Experiments 3, 4, and 5; and Bernie Corrigan and Mark Layman for their programming. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contracts NO0014-76-C-0074 (ARPA Order No. 2449) to Oregon Research Institute and Contract NO0014-76-C-0074 (ARPA Order No. 3052) under subcontract 76-030-0714 from Decisions and Designs, Inc., to Decision Research. The order of the authors is arbitrary. Reprints may be obtained from either author at Decision Research, 1201 Oak Street, Eugene, Oregon 97401.

Copyright © 1977 by Academic Press, Inc.
All rights of reproduction in any form reserved.

ISSN 0030-5073

Subjective probabilities are also an integral part of sophisticated techniques like cost–benefit analysis and decision analysis that are used heavily in both business and social contexts (e.g., Atomic Energy Commission, 1975; Raiffa, 1968; Slovic, Kunreuther, & White, 1974).

The quality of people's probability assessments sets an upper limit on the quality of their functioning in uncertain environments. Knowing how good people are at assessing probabilities clearly has both theoretical and applied importance.

One approach to validating probability assessments is to restrict one's attention to situations in which a "correct" probability can be consensually defined, for example, situations where extensive frequentistic data are available and probabilities are essentially estimates of relative frequencies. Peterson and Beach (1967) reviewed a number of studies that adopted this approach and found that people can estimate frequencies quite well. More recently, however, Tversky and Kahneman (1973) have suggested systematic biases that may be present in such judgments.

For many tasks, however, a consensually defined "correct" probability is unavailable. This is particularly true for probabilities reflecting judges' degrees of belief in propositions concerning "unique" events (e.g., what is the probability that Portugal will withdraw from NATO within 6 months?) or judges' knowledge about specific items of information (e.g., what is the probability that absinthe is a precious stone?). Such judgments reflect a degree of confidence entirely internal to the judge. Even if we know that Portugal did not withdraw from NATO during the period specified, or that absinthe is a liqueur, we can say nothing about how adequately the judge assessed and reported his or her own uncertainty. There is no way to evaluate an isolated judgment of this type.

Often, however, the judge makes many such responses, assessing the probability of different unique events occurring or different propositions being true. Over such a set of judgments, validity can be sought. One method of evaluating the quality of a set of probability judgments is to look at the internal consistency or coherence of the set. To be valid, subjective probability judgments must follow the axioms of the probability calculus. For example, since the two propositions given above are independent, the probability of both being true ("Portugal will withdraw from NATO" *and* "absinthe is a precious stone") should be equal to the product of the probabilities of each being true. Wyer (1974) adopted this approach in a large number of studies and found evidence of both consistency and inconsistency in different contexts. Perhaps the most interesting aspect of the latter was a tendency to overestimate the likelihood of compound events. Internal consistency is necessary for probability estimates to be valid, but it is not sufficient. Large systematic biases may exist in entirely consistent judgments.

This paper explores three more direct measures of the validity of a set of probability assessments: over- or underconfidence, calibration and resolution. We limit ourselves to situations in which a judge is always given questions with two alternative answers, one of which is true, the other false. The judge is asked to indicate which alternative is true, and to state the probability that the chosen alternative is, in fact, true. This probabilistic response is necessarily limited to the range $.5 \leq p \leq 1.0$.

Over/underconfidence. The overall tendency for a judge to be overconfident or underconfident is measured by:

$$\text{Over/underconfidence} = \frac{1}{N} \sum_{t=1}^{T} n_t \, (r_t - c_t) \ , \qquad (1)$$

where N is the total number of responses, n_t is the number of times the response r_t was used, c_t is the proportion correct for all items assigned probability r_t, and T is the total number of different response categories used (e.g., $T = 6$ for subjects who limit their responses to the single digits .5, .6, .7, .8, .9, and 1.0). A simple rearrangement of terms shows that Eq. (1) is equal to the difference between the mean of the probability responses and the overall proportion correct. Overconfidence is shown by a positive difference, underconfidence by a negative difference.

Overconfidence has been found in a variety of tasks (Lichtenstein, Fischhoff, & Phillips, in press; Fischhoff, Slovic, & Lichtenstein, in press).

Calibration. A judge is perfectly calibrated if, over the long run, for all propositions assigned the same probability, the proportion true is equal to the probability assigned. Thus, of those answers to which the perfectly calibrated assessor assigns a probability of being correct of .7, 70% will be correct; for all propositions to which .8 is assigned, 80% will be correct. For an assessor producing a large number of responses, one may group like responses and observe the hit rate for each subgroup. A graph showing the hit rate (percentage correct) for each probability response is called a "calibration curve." Calibration curve A in Fig. 1 reflects underconfidence: Whenever such a person says .7, 88% of the answers are correct; such people know more than their responses indicate. Curve B, the diagonal, represents perfect calibration. Curve C represents overconfidence; for example, only 47% of all the events to which the judge responds .7 are indeed correct.

A measure of the adequacy of calibration proposed by Oskamp (1962) replaces the parentheses in Eq. (1) by an absolute value sign, thereby measuring the mean weighted distance between the calibration curve and the identity line. An alternative measure is to take squared deviations,

$$\text{Calibration} = \frac{1}{N} \sum_{t=1}^{T} n_t \, (r_t - c_t)^2 \ , \qquad (2)$$

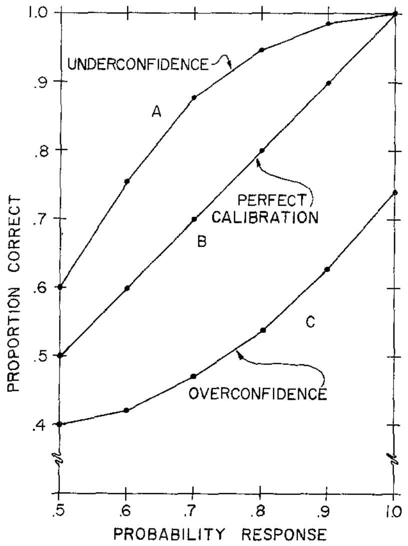


FIG. 1. Exemplar calibration curves.

as proposed by Murphy (1973). A perfectly calibrated person would score 0 on this measure. The worst possible score, 1.0, can be obtained only by a diabolical judge who always responds $r_t = 1.0$ when wrong and $r_t = 0.0$ when right.

Resolution. Murphy (1973) proposed an additional measure which is independent of calibration:

$$\text{Resolution} = \frac{1}{N} \sum_{t=1}^{T} n_t \, (c_t - c)^2 \;, \tag{3}$$

where c is the overall percentage correct. Resolution measures the ability of the responder to discriminate different degrees of subjective uncertainty by sorting the items into categories whose respective percentages correct are maximally different from the overall percentage correct (i.e., maximum variance across category hit rates). A flat (horizontal) calibration curve shows no resolution; a steep curve shows great resolution. The higher the resolution score, the better.

Murphy (1973, 1974) introduced yet one more measure of probabilistic performance, knowledge:

$$\text{Knowledge} = c(1-c) \;, \tag{4}$$

and showed that calibration, resolution, and knowledge (Eq. 2, 3, and 4) form a partition of the Brier score (Brier, 1950), a quadratic proper scoring rule (Shuford, Albert, & Massengill, 1966) such that the smaller the score, the better:

$$\begin{aligned}\text{Brier} &= \text{Knowledge} + \text{Calibration} - \text{Resolution}\\ &= c(1-c) + \frac{1}{N}\sum_{t=1}^{T} n_t\,(r_t-c_t)^2 - \frac{1}{N}\sum_{t=1}^{T} n_t\,(c_t-c)^2\ . \qquad (5)\end{aligned}$$

The Brier score is an overall measure of adequacy of performance in a probabilistic task. It has a minimum of .00 and a maximum of 1.00, obtainable only when all answers are wrong and all are assigned 1.00 (or all right and assigned .00). A more reasonable upper limit to the Brier score might be .25, earned by a judge who, having no knowledge, always responds .5 to a set of propositions half of which are correct. Such a subject would have a .25 knowledge score and .00 scores on calibration and resolution.

The distributional properties of the Brier score and its partition are not known. However, tests of differences between groups may be made by calculating such scores for each subject separately and relying on the law of large numbers.[1]

While a number of investigators have studied calibration, the only consistent finding has been that judges tend to be overconfident (for a review of this literature, see Lichtenstein, Fischhoff, & Phillips, in press). The present studies constitute a systematic look at how well people are calibrated and what affects their degree of calibration. In particular, we want to know whether the amount of knowledge a judge possesses about the content of the propositions being assessed affects her or his calibration. Earlier studies (Adams & Adams, 1961; Clarke, 1960; Nickerson & McGoldrick, 1963, 1965; Pitz, 1974; and Pollack & Decker, 1958) have reported some evidence that people who know more are better calibrated. The work reported here provides replication, clarification, and extension of these findings.

ALL EXPERIMENTS

Certain features shared by all experiments are reported here to avoid repetition.

Subjects. Except for Experiment 4, all subjects were paid volunteers who responded to advertisements in the University of Oregon student newspaper. Except for Experiment 4, the reported task was performed as part of a 2-hr group session along with several other judgmental tasks. Group size varied from 25 to 48 persons.

[1] Partition scores will depend to some degree on the number of responses used to calculate each score. The reader should be cautious in comparing scores calculated on differing amounts of data.

Tasks. All test items were dichotomous questions with the general form "Absinthe is (a) a precious stone, (b) a liqueur." In all experiments, subjects made two responses to each item. First, they chose one of the two alternatives as their best guess at the correct alternative. Second, they indicated with a number from .5 to 1.0 the probability that their choice was correct.

Measures. For each experiment, we report (1) the mean probability response across all subjects and items; (2) the percentage of items, across all subjects, for which the correct alternative was selected; (3) the difference between (1) and (2), which measures the group's over- or underconfidence; (4) the Brier score for the group, and the group scores on its partitions, knowledge, calibration, and resolution; and (5) a group calibration curve. For selected experiments, individuals' scores for the Brier and its partitions have been calculated, and mean scores across subjects are reported.

Calibration curves were constructed by grouping (over subjects and items) all the responses in the ranges .50–.59, .60–.69, .70–.79, .80–89, and .90–.99 and 1.00. The mean response for each grouping is plotted against the percentage correct (hit rate) associated with those responses.

NO KNOWLEDGE

Experiments 1a and 1b investigated the calibration of subjects with severely limited knowlege.

Experiment 1a

Method. Each of 92 subjects was asked to decide, for each of 12 small drawings, whether the artist was a European child or an Asian child and to estimate the probability that her or his selection was correct. Each set contained six drawings made by children from European countries and six drawings from Asian countries, all taken from Kellogg (1970), who had selected them to illustrate her thesis that children's drawings are the same the world over. This suggested that discrimination according to national origin would be very difficult. The test session was preceded by a brief study period in which the subjects were informed of Kellogg's thesis.

Results. As expected, the subjects had difficulty with this task. Only 53.2% of their 1104 answers were correct. Their probabilistic responses, however, indicated undue confidence, with a mean response of .677. Figure 12 shows the frequency distribution of probability responses. The calibration curve shown in Fig. 2 strongly suggests that these subjects were unaware of how little they knew. There is no relationship between their probability responses and the associated hit rates. Their abysmal knowledge, calibration, and resolution scores are presented in Table 1. Note that by virtue of giving a fair number of high probability responses

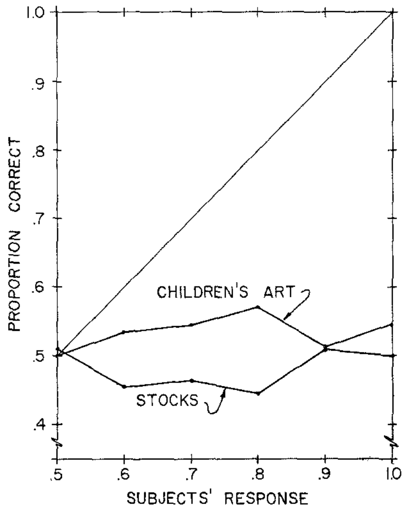


FIG. 2. Calibration curves for Experiment 1: no knowledge.

while knowing very little, their overall Brier score was worse (higher) than the .25 that they would have earned by guessing randomly and responding .5 every time.

Experiment 1b

Method. Sixty-three subjects were taught how to read the stock market charts for individual companies provided by the weekly Standard and Poor report, *Trendline*. After the instruction period, they were given charts of 12 stocks with data for the period from July 9, 1974, to February 14, 1975. For each stock, they were asked to indicate whether its March 22 closing price was higher or lower than that of February 14. Each of four test sets included six stocks that had increased and six that had decreased over the period, chosen at random from all stocks appearing in *Trendline* for February 14, 1975. Global market indexes (e.g., Dow–Jones) were similar for February 14, the last day shown on the charts, and for March 22, the target date, indicating that the market as a whole neither increased or decreased during this period.

Results. Again, the task was too difficult for subjects to perform adequately. Only 47.2% of their choices were correct. Again, they overestimated their knowledge, providing a mean probability of .654. The calibration curve shown in Fig. 2 and the Brier partition scores shown in Table 1 indicate the same insensitivity of probability judgments to level of knowledge found in Experiment 1a.

Comment. The lack of calibration evinced by the subjects in these two studies does not necessarily follow from their lack of knowledge.

TABLE 1

MEASURES OF PROBABILITY RESPONSES FOR GROUPS OF SUBJECTS

		Number	Mean	Proportion	Over-/under	Brier partition			
Experiment	Group	responses	response	correct	confidence	Knowledge	Calibration	Resolution	Total
1a	Children's drawings	1104	.68	.53	+.15	.249	.045	.001	.293
1b	Stock charts	756	.65	.47	+.18	.249	.043	.001	.291
2	Training	520	.78	.71	+.07	.206	.009	.022	.193
	No training	570	.65	.51	+.14	.250	.033	.002	.281
3	Best subjects	3000	.76	.71	+.05	.204	.011	.011	.204
	Middle subjects	2925	.71	.64	+.07	.229	.012	.013	.228
	Worst subjects	3075	.71	.56	+.15	.246	.030	.009	.267
	Best subjects								
	Easy items	1532	.80	.85	−.05	.130	.014	.008	.136
	Hard items	1468	.72	.58	+.14	.244	.032	.007	.269
	Middle subjects								
	Easy items	1472	.75	.80	−.05	.160	.012	.013	.159
	Hard items	1453	.67	.48	+.19	.250	.055	.004	.301
	Worst subjects								
	Easy items	1516	.73	.70	+.03	.212	.008	.012	.208
	Hard items	1559	.68	.43	+.25	.245	.086	.003	.328
4	Best subjects								
	Easy items	1450	.86	.92	−.06	.071	.013	.008	.076
	Hard items	1050	.71	.66	+.05	.226	.007	.017	.216
	Worst subjects								
	Easy items	1450	.82	.85	−.03	.130	.008	.013	.125
	Hard items	1050	.68	.51	+.17	.250	.037	.012	.275
5	Easy test	2250	.78	.80	−.02	.161	.005	.016	.150
	Hard test	2400	.74	.62	+.12	.236	.024	.013	.247

Subjects would have been quite well calibrated had they always given a probabilistic response of .5. This would have resulted in but one data point on the calibration curve for each experiment, but that point would have fallen reasonably close to the perfect calibration line. Only 7 of the 155 subjects in Experiments 1a and 1b acknowledged the limits of their own knowledge by following this strategy.

A LITTLE KNOWLEDGE

Will a small amount of knowledge improve calibration? Experiment 2 was designed to investigate this possibility by partially training subjects to make the requisite discrimination.

Experiment 2

Method. The stimuli were examples of the Latin phrase, "Mensa mea bona est," handwritten by either European or American adults. Twenty specimens were chosen on the basis of a pretest of 20 American subjects asked to sort 100 such specimens into two piles, American and European. The 20 specimens chosen for the experiment were correctly identified by 40–60% of these 20 subjects.[2] These specimens were randomly divided into two sets of 10, each of which included 5 European and 5 American specimens. One set was used as training stimuli; the other was used as test stimuli. This random division was performed four times, producing four paired sets of training and test stimuli.

Two of four groups of subjects ($N = 52$) received training on this task. In the training phase, they were asked to study for 5 min the 10 training stimuli, each correctly labeled. Immediately following this rudimentary training, the 10 test stimuli were presented. For each, the subjects were asked to indicate whether the specimen was European or American and to assess the probability that their answer was correct. They were not told how many of the 10 test stimuli were American.

The procedure for the two groups of untrained subjects ($N = 57$) was identical except that the specimens they studied in the first phase were not labeled as to country of origin.

Results. Training was moderately successful; the trained subjects correctly identified 71.4% of the specimens, compared with 51.2% for untrained subjects. The mean responses were .779 for the trained subjects, .653 for the untrained subjects. As can be seen in Fig. 3 and Table 1, trained subjects not only knew more, but showed better calibration and resolution. As in Experiment 1, untrained subjects showed no evidence of calibration or resolution.

[2] We are grateful to Lewis Goldberg, out of whose files we stole, without his knowledge, the handwriting specimens and pretest results.

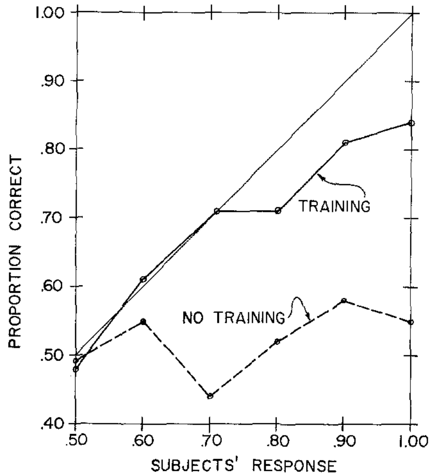


FIG. 3. Calibration curves for Experiment 2: a little knowledge.

DIFFERENT LEVELS OF KNOWLEDGE

The possibility, raised in Experiment 2, that greater knowledge improves calibration was further explored in Experiment 3.

Experiment 3

Method. The stimuli were 150 general knowledge items with highly varied content [e.g., Aden was occupied in 1839 by the (a) British, (b) French; bile pigments accumulate as a result of a condition known as (a) gangrene, (b) jaundice]. Each of 120 subjects responded to 75 items drawn from a pool of 150 items; 25 of the items received 80 responses, 100 items received 60 responses, and 25 items received 40 responses.

Results. Figure 4 presents the calibration curve over all 9000 responses. It is substantially flatter than it should be. The hit rate associated with the responses .50 and .60, and with .70 and .80, were virtually identical. Subjects generally overestimated the extent of their knowledge, getting 63.8% of the answers correct, but assigning a mean probability of .724.

The subjects were divided into three subgroups according to how knowledgeable they had been: the best subjects (40 subjects with 51 or more correct answers out of 75), the middle subjects (39 subjects with 46–50 correct answers), and the worst subjects (41 subjects with fewer than 46 correct answers). Separate analyses were performed for each group. Calibration curves appear in Fig. 5, with the corresponding measures in Table 1. These data strongly suggest that the more you know, the better is your calibration. All groups tended to overconfidence, but the most knowledgeable subjects showed the least overconfidence

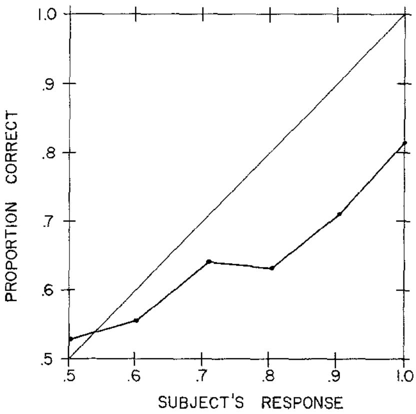


FIG. 4. Overall calibration curve for Experiment 3.

and had the calibration curve closest to the identity line (representing perfect calibration).

Dividing responses according to item difficulty rather than subject proficiency produced much the same result (not shown). The calibration curve for the easiest items was considerably closer to the identity diagonal than that for the most difficult questions.

Pushing this idea one step further, one might ask, how well calibrated were the best subjects on the easiest items? Here, we might expect to find the best calibration. The data of Experiment 3 were reanalyzed to

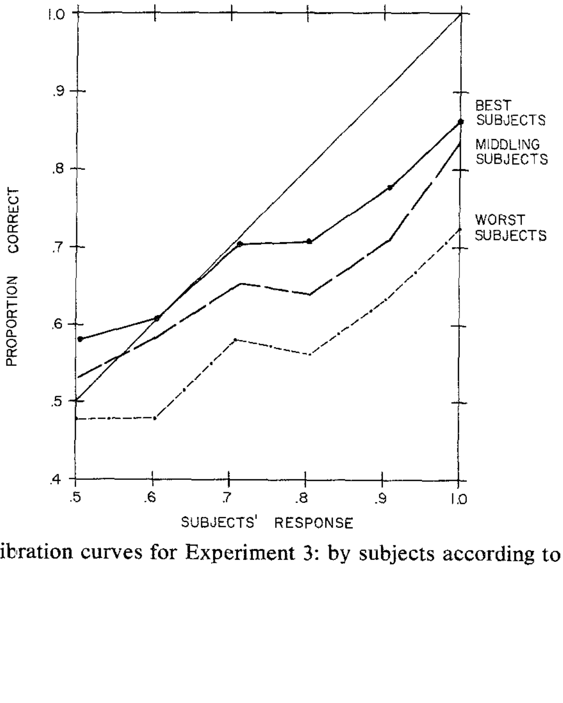


FIG. 5. Calibration curves for Experiment 3: by subjects according to knowledge.

investigate this possibility. Items were sorted into two equal groups according to the percentage of subjects answering them correctly: easy items (67% or more correct) and hard items (less than 67% correct). Each of the three groups of subjects was calibrated for each of the two groups of items, to produce the six calibration curves shown in Fig. 6.

Despite some irregularities in these calibration curves due to the reduced number of responses per data point, a pattern of roughly parallel lines emerged. With low knowledge, substantial overconfidence occurred. However, when the percentage of correct answers was high (85% for the best subjects on the easy items and 80% for the middle subjects on the easy items), substantial underconfidence was seen (e.g., 75% of the .60 responses were correct). As shown in Table 1, the groups whose percentage correct exceeded 80% had slightly worse (higher) calibration scores than the group with 70% correct. The possibility that calibration is curvilinearly related to knowledge will be explored more fully in Experiment 5.

Experiment 4

Although conducted for a somewhat different purpose (see below), Experiment 4 affords a replication of the above analysis.

Method. All on-campus graduate students in the Psychology Department of the University of Oregon were asked to participate in this

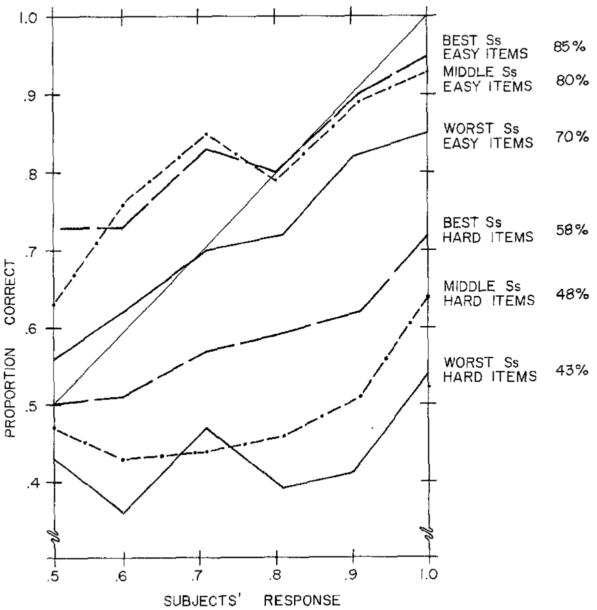


FIG. 6. Calibration curves for Experiment 3: subjects by items.

experiment. Packets with stimuli and instructions were sent to all 64 graduate students; 50 were returned completed.

The stimuli were 50 general knowledge items (30 of those used in Experiment 3 and 20 additional, similar items) and 50 specially written items dealing with psychology [e.g., the Ishihara test is (a) a perceptual test, (b) a social anxiety test; Anna Freud is Sigmund Freud's (a) oldest child, (b) youngest child]. The two types of items were randomly intermixed in the stimulus package.

Results. Separate calibration curves are shown in Fig. 7 for four subsets of responses obtained by splitting the subjects into best and worst at the median (74.5%) of the distribution of percentage correct and by splitting the items into easy (at least 75% correct; 58 items) and hard (fewer than 75% correct; 42 items). For these analyses, no distinction was made between general knowledge and psychology items. The same pattern of almost parallel lines found in Fig. 6 emerged from these data, shown in Fig. 7. Summary statistics are given in Table 1. Again, the group with the highest knowledge did not have the best calibration scores.

In a series of experiments involving four-alternative questions and confidence ratings expressed on a 5-point scale, Nickerson and McGoldrick (1963, 1965) obtained results highly similar to those of Experiments 3 and 4: (a) Calibration curves for subjects who answered more questions correctly lay parallel to, but higher than, the corresponding curves for subjects who knew less. (b) Tests constructed to have different levels of

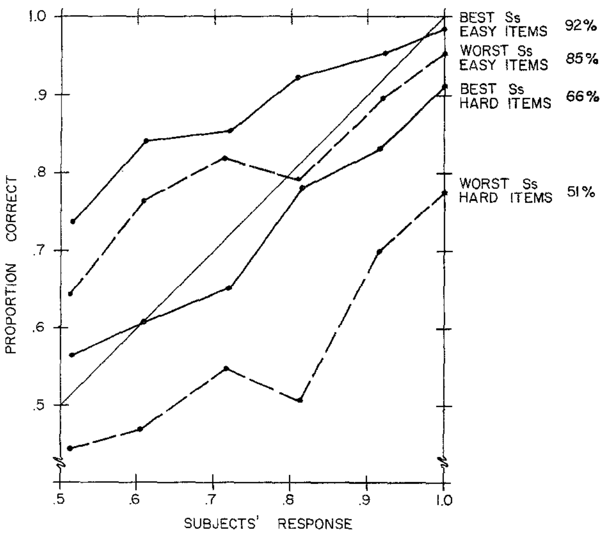


FIG. 7. Calibration curves for Experiment 4: graduate students in psychology, subjects by items.

difficulty had similar, but vertically displaced, curves. (c) Subjects who knew more tended to be more confident than those who knew less.

EFFECTS OF CHANCE FLUCTUATIONS

The analytic technique used in Experiments 3 and 4, in which the data were divided into subsets as a function of item difficulty and subjects' performance, is vulnerable to random fluctuations which could artifactually produce separation between the calibration curves for the subsets. Assume that our subjects were equally knowledgeable and identically calibrated. In any sample of their responses some subjects will appear more knowledgeable by chance. The same chance factors that led them to have a higher overall percentage correct will also lead them to have a higher hit rate for their responses of .5, .6, etc., and thus have an elevated calibration curve.

The extent to which such chance factors could lead to differences in calibration was examined by simulating the results of Experiment 4. For the simulation, all subjects were assumed to have exactly the same calibration, which was taken as the actual calibration derived from pooling their responses to all 5000 items (100 items for each of 50 subjects). Subjects' original probability responses were retained in the simulation. For each response, the correctness of the chosen alternative was simulated in accordance with the overall calibration curve. For example, since in the real data 86% of the .90 responses were correct, in the simulation, each response of .90 received a simulated outcome, either correct with a probability of .86 or incorrect with a probability of .14. These simulated data (the original probability reponses with randomly assigned outcomes) were then partitioned into four subsets: best and worst subjects, easy and hard items. The calibration for each subset was computed. The entire simulation was repeated 50 times.

Figure 8 shows the average calibration curves across the 50 replications. Figure 8 is directly comparable to Fig. 7; it is based on the same data except for the assumption that all subjects have exactly the same calibration. The amount of separation between the calibration curves in Fig. 8 is due solely to chance. This separation is much smaller than the separation found in the original data (Fig. 7). The most extreme of the 50 simulated curves for best subjects/easy items (that with the highest simulated hit rate, 88%) is only slightly higher than the average curve shown in the figure; it is still well below the actual curve for best subjects/easy items. Similarly, the most extreme curve for worst subjects/hard items (lowest simulated hit rate, 57%) is close to the average curve and well above the actual one. We reject the hypothesis that in Fig. 7 all subjects on all items had the same calibration.

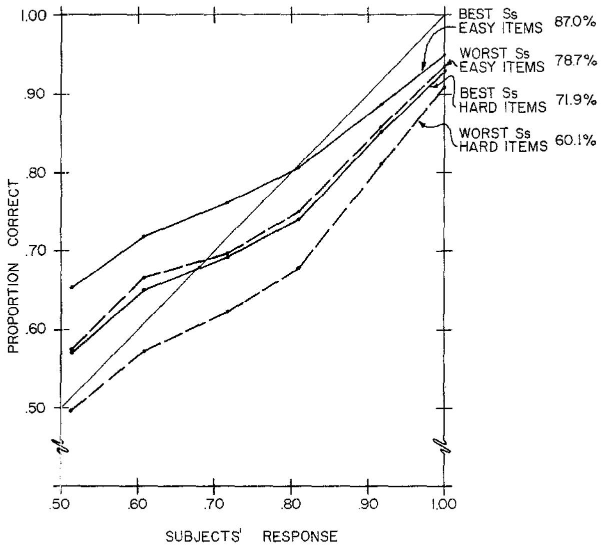


FIG. 8. Simulated calibration curves, using probability responses of data appearing in Fig. 7.

TESTS VARYING IN DIFFICULTY VS SUBTESTS VARYING IN DIFFICULTY

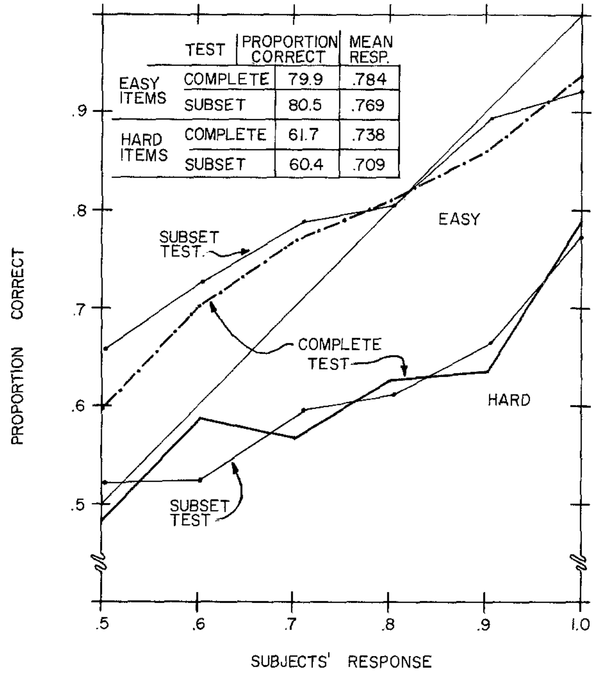
The previous experiments analyzed subsets of items actually contained in a single test. It may be that some adaptation to the overall difficulty of the test might account for the observed overestimation with hard items and underestimation with easy items. This possibility was explored in Experiment 5.

Experiment 5

Method. From the items used in Experiment 3, two tests of 50 items each were compiled. Items were selected in pairs according to the percentage of subjects answering them correctly in Experiment 3. Each item in the hard test was matched with an item in the easy test that had been answered correctly by an additional 20% of subjects. The mean percentage correct for the hard test was 60.4 (range, 46.2 to 77.5); for the easy test 80.5 (range, 66.2 to 97.5).

The two tests were distributed to 93 subjects; 48 received the hard test and 45 the easy test.

Results. Figure 9 compares results from this experiment (the "complete" tests) with those from Experiment 3 using the same items (the "subset" tests). Here, too, the calibration curve depends on test difficulty, with underconfidence on the easy test and overconfidence on the



	TEST	PROPORTION CORRECT	MEAN RESP.
EASY ITEMS	COMPLETE	79.9	.784
	SUBSET	80.5	.769
HARD ITEMS	COMPLETE	61.7	.738
	SUBSET	60.4	.709

Fig. 9. Calibration curves for easy and hard items from complete tests (Experiment 5) and subset tests (Experiment 3).

hard test. The similarity between the calibration curves for the complete tests and the subset tests is striking.

Interpretation of these results is somewhat ambiguous. We know from Fig. 8 that some of the spread between calibration curves for items selected from a larger set on the basis of percentage correct is due to capitalization on chance. Therefore, one would expect a regression effect upon reuse of the the same items; that is, the calibration curves should be somewhat closer together. The fact that the curves for complete tests are so similar to those for the subset tests suggests either that there is a context effect which just offsets the regression effect or that neither effect is strong enough to be perceptible. We lean to the latter interpretation; the regression of mean percentage correct, for example, was quite modest: from 60.4 in Experiment 3 to 61.7 in Experiment 5 for the hard test and from 80.5 to 79.9 for the easy test.

Individual analyses. Because they were preselected on the basis of difficulty, the hard and easy complete tests shown in Fig. 9 provide the cleanest test of the effect of knowledge (percentage correct) on appropriateness of probability response. Knowledge, calibration and resolution scores were calculated for each subject in these groups and compared by t tests. The group given the easy test had significantly lower knowledge scores (t (91) = 9.48), as would be expected from the difference between overall percentages correct (61.7 vs 79.9%). They were also better calibrated (t (91) = 4.54), reflecting the greater proximity of their curve to the identity line. However, there was no appreciable difference in resolution scores (t (91) = 0.84), meaning, roughly, that the two calibration curves had about the same slope. Thus, although those who knew more used the numbers of the probability score more appropriately, they were not able to make finer discriminations in their judgments.

Examination of Figs. 6 and 7 (and the corresponding statistics in Table 1) suggests, however, that calibration may not improve indefinitely with increased knowledge. As percentage correct increases from chance level (50%) to about 80%, overconfidence decreases, producing improved calibration. Above 80%, however, people seem to become underconfident and calibration worsens. We tested this hypothesis with the individual. Brier partition scores of subjects who had participated in the hard and easy groups of Experiment 5. These 93 individuals had percentages correct ranging from 42 to 96%, with the 10th percentile at 54% and the 90th at 86%. Ideally, for this analysis, we would have liked to have had more subjects at the upper extreme. Over all subjects, there was a moderate linear relationship between calibration score and percentage correct ($r = -.48$, $p < .001$). The quadratic relationship between these two variables was, however, much stronger ($R = .62$, $p < .001$); the improvement in variance explained by including a quadratic term was highly significant (F (1, 90) = 22.49; $p < .001$). The quadratic equation was $\hat{y} = 0.72x^2 - 1.13x + 0.47$ (where $\hat{y}$ = calibration score and x = percentage correct), indicating that the best calibration is associated with approximately 78% correct. There was no systematic relationship, either linear ($r = -.06$) or quadratic ($R = .20$), between resolution and percentage correct. This was consistent with our failure to find any difference in resolution between the hard and easy groups.

Individual analyses also confirmed the finding in Experiment 1 that lack of knowledge leads to poor performance. Seventeen of the 48 subjects who took the hard test scored less than 60% correct. The total Brier score for 15 of these 17 was greater than .25, i.e., worse than they could have scored by always responding .5 and answering randomly.

EXPERTISE

Perhaps "percentage of items answered correctly" does not really capture the essence of expertise. Experts might be better calibrated not because they know the correct answer for more items, but because they have thought more about the topic area in question and thus can more readily recognize the extent and the limitations of their knowledge. The following analysis searched for differences in calibration due to any sort of "quality of insight" that experts might have above and beyond their level of knowledge.

Method. The experts were the 50 graduate students in the Department of Psychology mentioned in the description of Experiment 4. This experiment is simply a reanalysis of those data, comparing their calibration on the 50 items pertaining to psychological knowledge with the 50 general-knowledge items.

Results. The psychology subtest and general-knowledge subtest were virtually identical in percentage correct (75.7 vs 76.0) and mean probability response (.780 vs .778). Figure 10 shows that calibration for the two subtests was essentially the same. Dependent group t tests done on individual subjects' calibration and resolution scores showed no significant differences (p = .10) between the two subtests. Thus with percentage correct held constant, there is no evidence that expertise in a particular subject area leads to better calibration or resolution.

INTELLIGENCE

The subjects in Experiment 3 were mostly undergraduate students attending the University of Oregon. Their intelligence scores are probably lower, on the average, than those of the graduate student subjects of Experiment 4, who are highly selected for intelligence by the admissions procedures of the Psychology Department. We assume that we are thus able to investigate the effects of intelligence on calibration.

Method. Subtests of 54 items each, matched item by item in difficulty,

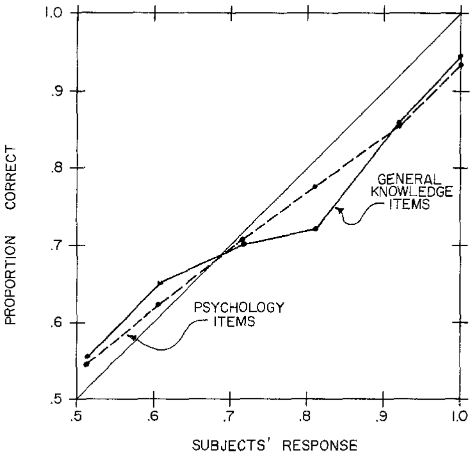


FIG. 10. Expertise. Calibration curves for psychology graduate students responding to psychology and general knowledge items.

were created from the Experiment 3 (regular volunteer subjects) and Experiment 4 (graduate student subjects) data.

Results. Eight items were common to both groups. Responses to them revealed the graduate students' superior knowledge. They averaged 73.5% correct on these items, compared with the regular volunteers' mean of 63.8% correct. The graduate students had a smaller percentage correct for only one of the eight items.

The matching process succeeded in producing subtests with a mean percentage correct of 68.1 for the graduate students and 67.7 for the regular volunteers. Mean probability responses were .742 and .741, respectively.

Figure 11 shows the calibration of the two groups. It appears that the graduate students may have been slightly better calibrated at the extremes. However, comparison of calibration scores for individual subjects revealed no difference between the two groups (*t* (88) = .52). The psychology graduate students did, however, show significantly better resolution [.053 vs .033; *t* (88) = 3.91].

DISTRIBUTION OF RESPONSES

Figure 12 presents the proportion of subjects' probability responses that fell into each response category. These proportions are shown for all groups or subgroups of all experiments, ordered by percentage correct. Subjects showed a definite tendency to make more use of the high end of the response scale for the easiest tests. However, this tendency, while in the right direction, was less than it should have been. While percentages

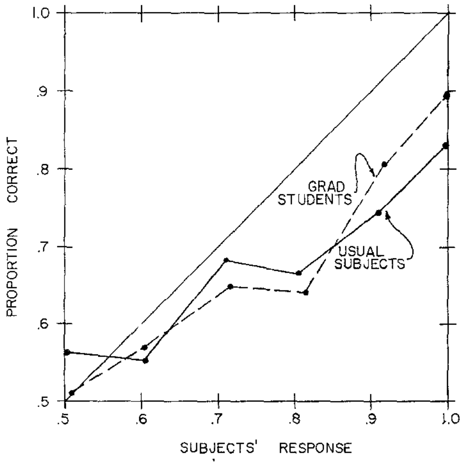


FIG. 11. Intelligence. Calibration curves for psychology graduate students and usual subjects with items matched for difficulty.

correct ranged from 43 to 92, the range of mean probability was only .65 to .86. It is this insufficient discrimination which leads to underestimation with easy tests and overestimation with hard tests.

The other striking attribute of Fig. 12 is the great frequency of extreme responses (.5 and 1.0). While no response category was unused, over all experiments, subjects used the extreme categories for about half their responses. This inclination to treat the task as dichotomous (either "I know the answer"—1.0, or "I don't know the answer"—.50) appears to have been less pronounced in Experiments 1 and 2, with relatively few items all dealing with the same topic, than in Experiments 3, 4, and 5, which used many items concerning diverse topics.

The effect on calibration of the tendency to avoid using probabilities other than .5 and 1.0 was examined with the data of Experiment 3. Subjects were divided into three groups: heavy users of .5 and 1.0 (49 subjects using these two responses more than 50% of the time; mean use, 67.7%); medium users (33 subjects using .5 and 1.0 between 41% and 49% of the time; mean use, 46.9%); and light users (38 subjects using .5 and 1.0 for 40% or less of their answers; mean use, 34.4%). The three groups were similar in percent of items answered correctly (65, 64, 62%, respectively). Their calibration curves (not shown) were highly similar, and one-way ANOVAs on individual calibration and resolution scores showed no significant differences. We thus found no support for the notion that the tendency to avoid using probabilities, as an individual difference, affects calibration. This result is consistent with Phillips and Wright's (1976) finding of quite low (<.30) correlations in three different groups of subjects between a measure of the tendency to use .5 and 1.0 responses and a measure of calibration.

DISCUSSION

Do those who know more . . . ? A summary of the results of this series of experiments can be cast as an answer to the question posed in the title of this article. The nonknowledge components of probability assessments are over- or underconfidence, calibration, and resolution. The primary distinction in our findings is between knowing nothing at all and knowing something. It is clear that those who knew something did outperform those who knew nothing. The latter situation typically led to vast overconfidence, terrible calibration, and no resolution. Indeed, when individuals or groups knew less than 60% of the correct answers to items, they would have been better off (earned a better Brier score) had they always said .5 and flipped a coin to determine the correct answer.

Among those who knew something, the results were more complex. With increasing knowledge comes decreasing overconfidence until, for

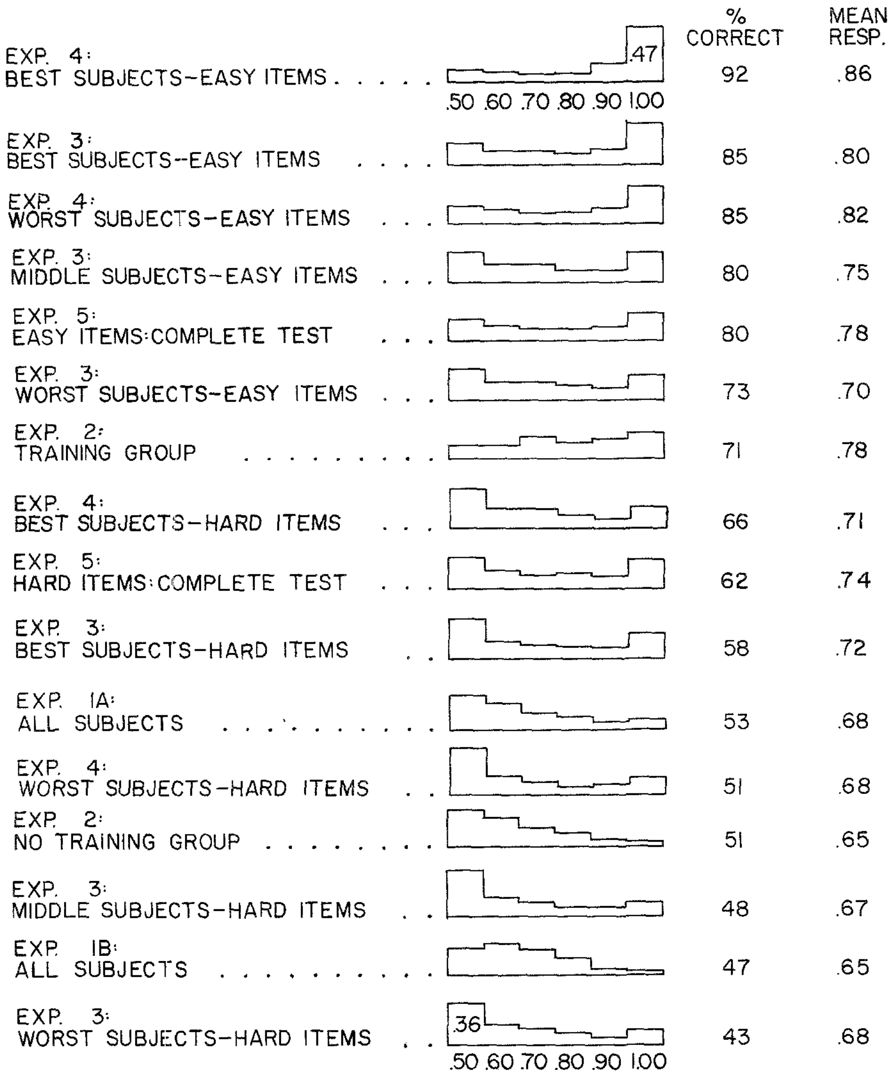


FIG. 12. Distributions of subjects' responses.

those whose percentage correct exceeded 80%, we found moderate underconfidence. This relationship resulted in a nonmonotonic relationship between knowledge and calibration, with the best calibration found at approximately 80% correct.

Resolution scores for those who knew something did not change as knowledge changed.

Thus our answer to the title question is: Those who know more do *not* generally know more about how much they know. Calibration improves up to about 80% correct, and then becomes worse. Resolution does not change at all after about 60% correct. Overconfidence is widespread, but changes to underconfidence with very easy items.

The strikingly different calibration curves for items of varying difficulty are a direct result of subjects' insensitivity to how much they really know. Among the items for which they believe that they have a 50% chance of knowing the correct answer, the appropriate probability may be anywhere between .45 and .85. When they estimate 1.00, the appropriate probability may be between .55 and .95 (Fig. 6). The ease with which different calibration curves were constructed from the fairly representative sets of items used in Experiments 3, 4, and 5, and the large numbers of responses in each category for even the most extreme curves, indicate that subjects' inability to make discriminations is widespread (i.e., there are not only some instances in which, for example, people should be saying .75 when they actually say .50, but many such instances).

Other experiments reported here explored other possible correlates of probabilistic performance. Neither our manipulation of expertise nor variation in intelligence produced any change in performance when knowledge was held constant, except for the finding that the graduate students had better resolution than our regular (mostly undergraduate) subjects.

Implications. Although subjective probabilities have a prominent role in many psychological theories, the study of probabilities themselves has been atheoretical in most cases (including the present study; see also Lichtenstein, Fischhoff, & Phillips, in press). While there have been some suggestions for, or fragments of, process theories of calibration (Pitz, 1974; Slovic, 1972; Tversky & Kahneman, 1974), only Pitz (1974) predicts a decrease in overconfidence as knowledge increases. Some comprehensive theories of the psychology of confidence are urgently needed to account for the data that are rapidly accumulating.

Aside from their theoretical import for the psychologist interested in how people perform judgments under conditions of uncertainty, these results have strong implications for those whose jobs involve actually making and taking responsibility for such judgments. With the development of sophisticated information processing and decision analytic techniques, operations as diverse as intelligence analysis, corporate planning, environmental impact assessment, and nuclear power engineering utilize explicit probability assessments (Fischhoff, 1977). Users of these approaches should consider results like the present ones in determining how much faith to put in the results of their analyses. Similarly, psychologists who elicit subjective probability estimates in the study of behaviorial phenomena might think twice before taking them at face value, or expecting too much of them.

In addition to their cautionary value, these results may also help improve the quality of probabilistic analyses. Assume that in the context of a practical problem using judgments of the type studied here, a judge reports a probability of .90. From Fig. 4, we know that a better estimate of the appropriate probability is .71, and we would do better treating it as such. Although such "correction after the fact" is better than taking biased judgments at face value, the revised assessments may still be inappropriate. In the present example, even though our best guess of the appropriate probability is .71, Fig. 6 suggests that anything between .40 and .90 might be better still, depending on the difficulty of the item involved.

If we know how difficult the item is, then we can make a much more accurate correction. In practice, however, such situations will be rare. To know how difficult an item is, we must know the correct answer. But if we know the correct answer, we will not have any practical need for the judge's assessment. Such assessments are valuable only when the correct answer is not known. Short of knowing the correct answer, the only way to capitalize on the relationship between item difficulty and type of miscalibration seems to be to assume something about the difficulty of the items in the world in which our judge is functioning. The distribution of judges' responses (as shown in Fig. 12) could be exploited for this purpose. Across the 16 groups or subgroups, there was a correlation of .91 between percentage correct (an index of difficulty which is typically unknown in a practical setting) and mean response (which is observable when a number of assessments are made). Thus, inferences about task difficulty could be made when true outcomes are unknown. With some idea of task difficulty, even so indirectly measured, more precise external recalibration of probability assessments is possible. Without it, the present data suggest that we have only a vague idea of whether to recalibrate an assessment by increasing or decreasing it.

In view of these difficulties in recalibration, it is important for future research to explore the possibility that judges can be trained to be better calibrated, thus obviating the need for correction.

The practical implications of our findings on resolution are more problematic. In some ways resolution is a more fundamental aspect of probabilistic functioning, for it reflects the ability to sort items into subcategories whose percentage correct is maximally different from the overall percentage correct. Calibration can then be viewed as the ability to attach appropriate numerical labels to these subcategories. Without resolution, calibration is bound to be terrible, except in the special case of an assessor who gives the overall percentage correct as her or his every response. Neither external recalibration nor training in calibration can be expected to be successful when resolution is lacking. It might be possible to devise a two-stage training procedure, with an initial focus on the nonnumerical discriminative sorting necessary for resolution, followed by training in attaching numbers to the sorted categories, to achieve good calibration.

REFERENCES

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

Atomic Energy Commission. *Reactor safety study: An assessment of accident risks in U.S. commercial power plants* (WASH-1400 Draft). Washington, D.C.: The Commission, 1974.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 1950, **75,** 1–3.

Clarke, F. R. Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests. *Journal of the Acoustical Society of America,* 1960, **32,** 35–46.

Cohen, J. *Chance, skill and luck: The psychology of guessing and gambling.* Baltimore: Penguin, 1960.

Edwards, W., & Tversky, A. *Decision making.* Baltimore: Penguin, 1967.

Feather, N. T. Subjective probability and decision under uncertainty. *Psychological Review,* 1959, **66,** 150–163.

Fischhoff, B. Cost–benefit analysis and the art of motorcycle maintenance. *Policy Sciences,* 1977, **8,** 177–202.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing with certainty. *Journal of Experimental Psychology: Human Perception and Performance,* in press.

Fishbein, M. A behavior theory approach to the relations between beliefs about an object and the attitude toward the object. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967.

Jones, E. E., & Davis, K. E. From acts to dispositions: the attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology.* New York: Academic Press, 1965. Vol. 2.

Kellogg, R. *Analyzing children's art.* Palo Alto, Calif.: National Press, 1970.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement.* New York: Academic Press, 1971. Vol. 1.

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

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

Murphy, A. H. A sample skill score for probability forecasts. *Monthly Weather Review,* 1974, **102,** 48–55.

Nickerson, R. S., & McGoldrick, C. C. Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Perceptual and Motor Skills,* 1963, **17,** 159–167.

Nickerson, R. S., & McGoldrick, C. C. Confidence ratings and level of performance on a judgmental task. *Perceptual and Motor Skills,* 1965, **20,** 311–316.

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. *Psychological Bulletin,* 1967, **68,** 29–46.

Phillips, L. D., & Wright, G. N. *Group differences in probabilistic thinking.* (Tech. Rep. 76-4). Brunel Institute of Organisation and Social Studies, 1976.

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

Pollack, I., & Decker, L. R. Confidence ratings, message reception, and the receiver operating characteristic. *Journal of the Accoustical Society of America,* 1958, **30,** 286–292.

Raiffa, H. *Decision analysis.* Reading, Mass.: Addison Wesley, 1968.

Shuford, E. H., Albert, A., & Massengill, H. E. Admissible probability measurement procedures. *Psychometrica,* 1966, **31,** 125–145.

Slovic, P. From Shakespeare to Simon: Speculations—and some evidence—about man's ability to process information. *Oregon Research Institute Research Monograph,* 1972, **12,** 2.

Slovic, P., Kunreuther, H., & White, G. F. Decision processes, rationality and adjustment to natural hazards. In G. F. White (Ed.), *Natural hazards, local, national and global.* New York: Oxford University Press, 1974.

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

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

Weiner, B. *Achievement motivation and attribution theory.* Morristown, N.J.: General Learning Press, 1974.

Wyer, R. S. *Cognitive organization and change: An information processing approach.* Potomac, Md.: Erlbaum, 1974.

RECEIVED: July 12, 1976