Categorical Confidence

Baruch Fischhoff, Don MacGregor and Sarah Lichtenstein

Decision Research
A Branch of Perceptronics
1201 Oak Street
Eugene, Oregon 97401

April 1983

Running head: Categorical Confidence

Abstract

People tend to be inadequately sensitive to the extent of their own knowledge. This insensitivity typically emerges as overconfidence. That is, people's assessments of the probability of having answered questions correctly are typically too high compared to the portion of questions they get right. Few debiasing procedures have proven effective against this problem. Those that have worked seem to be directive in character. Rather than improving subjects' feeling for how much they know, such procedures may have suggested to subjects how their probability assessments should be changed. These successful manipulations include giving feedback and requiring subjects to provide reasons contradicting their chosen answers. The present study attempted to improve the appropriateness of confidence with a nondirective method. Subjects were asked to sort items into a specified number of piles according to their confidence in the correctness of their answers. Subsequently, they assigned a number to each pile expressing the probability that each item in the pile was correct. It emphasizes confidence assessment over fact assessment; it forces the comparison of knowledge levels for different questions, it deemphasizes the need to produce numbers; it gives different hints as to the fineness of the discrimination that assessors can make. This procedure differed from its predecessors in many respects; nonetheless, performance here was indistinguishable from that observed elsewhere. Although some small pockets of improvement were noted, confidence was largely resistant to this manipulation. Such robustness is discouraging for the developer of elicitation procedures, encouraging for the student of judgmental processes.

Categorical Confidence

Whenever there is uncertainty, it can be important to know its extent. In business, technology, or everyday life, understanding the limits to one's knowledge is an essential ingredient for wise decisions regarding when and how to act, as well as what additional information to acquire and what advisors to trust. As a result, explicit assessments of uncertainty play a central role in theories regarding processes as diverse as auditing financial statements [12], designing nuclear power plants [22], describing strategic balances [8], analyzing business decisions [20], interpreting eye witness testimony [3, 24], estimating physical constants [21], monitoring the operation of one's memory [9], and learning how to learn [25]. If there is any degree of communality to the way in which uncertainty is assessed in these different contexts, then the study of uncertainty assessment per se could produce widely useful results. This hope has generated a substantial body of research examining how people assess the limits of their own knowledge. Although the "how" of these processes has only been incompletely understood, the "how well" is fairly clear. People seem to have difficulty appraising uncertainty accurately. After describing these difficulties, the present article offers a new procedure for reducing them.

In a typical uncertainty assessment task, participants first ponder some question of fact and then assess the likelihood that the answers they have produced are correct. Observation of such individuals suggests that they spend considerably more time on the first of these operations than on the second. A variety of possible reasons come to mind: (a) answers are harder to produce than probabilities; therefore they require more time; (b) people are more experienced in answering questions, hence, they can spend more time

profitably on that task; (c) until an answer is produced, one cannot even begin to assess its accuracy; or (d) people are more accustomed to having answers evaluated than probabilities and want to take greater care that the former are in order.

Given these reasons for deemphasizing the probability assessment task, it should perhaps come as no surprise then that its quality is often poor. The most commonly observed result is that the magnitude of probability assessments is only roughly predictive of the actual likelihood that the associated answers will be correct. In most cases, correctness does increase as confidence increases. However, it increases too slowly. In many tasks, as people's assessed probabilities of being correct increase from .5 to 1.0, their actual probability of being correct increases from .5 to only about .8. People believe that they can distinguish between a greater range of states of knowledge than is actually the case.

When tasks are difficult, a contrast between people's overall confidence and their overall accuracy reveals overconfidence; they make too many high confidence assessments. With easy tasks, one finds underconfidence. These patterns are very robust; they can be found with a variety of response modes, question topics, and levels of expertise (for reviews, see [5, 16, 23]). People have, moreover, considerable confidence in these confidence assessments (e.g., [7]).

The few experimental manipulations that have managed to improve the appropriateness of confidence assessments have typically involved focusing people's attention on the assessment task in a fairly directive manner. For example, the quality of assessment improves when assessors are given extensive personalized feedback (e.g., [15, 18]). Another effective manipulation has

been requiring people to list explicitly reasons supporting and contradicting their choice of answer, prior to assessing the likelihood that it is correct [13].

The secret of even these partial successes is, however, still unclear. It would be theoretically interesting anad practically useful if such simple manipulations were able to enhance people's ability to appraise their own knowledge. However, the improvement observed with these manipulations might come not from helping people focus on the assessment task but from some unintended cues as to how subjects should change their assessments. For example, because people do not ordinarily list contradicting reasons, the requirement to do so might be interpreted by some subjects as a hint to reduce their confidence. Feedback shows what assessments one should have used; it may be tempting just to reduce one's probability assessments mechanically, which would lead to better performance if training and test items were of equal difficulty.

An obvious danger with such directive procedures is that whatever is learned may prove to be task specific, leaving one no better (or even more poorly) prepared to face a new task differing, say, in difficulty level. Learning that one is overconfident on a hard task might, in fact, induce exaggerated underconfidence on a subsequent easy task. These fears are alleviated somewhat by Lichtenstein and Fischhoff's [15] finding of modest generalization of training to some other tasks. Nonetheless, it would be comforting to know that confidence assessment could be improved by a technique that affected response usage only as a by-product of affecting understanding of how much one knows.

One simple, non-directive way to focus attention on the assessment task would be to provide people with a detailed lecture on the nature of the response mode, the properties of good assessments, and the kinds of biases that may be observed. Such instruction would prepare people for assessment in general, not just for one particular task. Unfortunately, however, it does not seem to work (see [15]; Note 1).

The present experiment explores an alternative non-directive approach that differs in many ways from its predecessors. In it, judges first answer an entire set of two-alternative forced-choice questions. Then they sort the questions into a prescribed number of piles, each reflecting a different degree of confidence that the answers chosen for the items assigned to it are correct. Finally, after reviewing the results of the sorting procedure, judges assign a number to each pile expressing the probability that each item in the pile is correct. This procedure should emphasize confidence assessment over question answering. Moreover, within the assessment task it should focus attention on appraising one's feeling of knowing rather than on the production of a numerical expression of that feeling that the experimenter will find acceptable. Some explicit response is, of course, needed to communicate one's degree of confidence, but the careful formulation of a feeling of knowing should take precedence over the more technical task of translating it into a number [1].

One respect in which the present procedure is directive is in its specification of the number of categories that subjects are to use. That number might be reasonably interpreted by subjects as an indication of how many distinct categories they can reliably use. There is probably no way to avoid giving some direction on this topic. For example, the non-categorical

half-range probability scale [.5, 1.0] used in many studies seems to suggest to subjects that they can and should use all the "round" responses (.5, .6, .7, .8, .9, 1.0). One might even attribute the hypersensitivity observed in such studies to this implicit suggestion that they are able to make the discriminations corresponding to these six distinct levels of knowledge [4].

A final feature of this procedure that might have a salutary effect is that it forces subjects to read the entire set of questions before assessing their confidence in any. Upon entering an experiment, subjects may have some expectation regarding how difficult the questions will be. If that expectation is erroneous, it might artificially buoy or depress their confidence levels until they had completed enough questions to realize that their assumption was in error [19].

Within the judgment and decision making literature, perhaps the most closely related study is one in which Gray [11] had third graders assess the percentage of questions that they would answer correctly from each of six piles of arithmetic problems. Although she did not evaluate the accuracy of these estimates, Gray did find that they were systematically related to her subjects choices of problems to tackle (undervarying rewards schemes).

A more general reason in the literature for believing that this procedure will be efficacious is the proliferation of techniques such as the Q-sort [2, 10]. Many investigators have found that when subjects have difficulty making direct qualitative estimates, that their task can be simplified by changing it to one requiring primarily comparative judgments. The Q-sort procedures accomplish this by having subjects sort alternatives into equivalence classes, akin to those being created here. That literature would lead one to expect more robust and valid assessments with our sorting task.

Experiment 1

Method

Design. The experimental design involved four groups of subjects, each asked to sort 50 two-alternative questions into a prespecified number of piles (3, 4, 5, or 6) according to their degree of confidence in knowing the correct answer to each. After the sorting was completed, they assessed the probability that each answer in each pile was correct.

Procedure. The details may be best understood by verbatim citation of relevant portions of the experimental instructions:

> For this task, we have prepared 50 general-knowledge items. Each item has two alternative answers, one of which is correct and one incorrect. Each item appears on a card. Your job is to:
>
> Step 1--Separate the 50 cards, tearing them along the dotted lines (there are six (6) cards on each page).
>
> Step 2--Go through the cards and circle the letter a or the letter b to indicate which of the alternatives you think is the correct alternative. If you have no idea which alternative is correct, circle one of the two letters anyway--just guess.
>
> Step 3--Sort the cards into 3, [or 4, 5, or 6] piles according to how sure you are that you have circled the correct alternative.
>
> * One pile should contain all the cards for which you feel least confident;
>
> * One pile should contain all the cards for which you feel most confident;
>
> * The other pile(s) will have cards for which you have an intermediate feeling(s) of sureness.

Keep sorting and resorting until all the cards in a particular pile are those for which you feel the same level of certainty or uncertainty.

You may, if you wish, do steps 2 and 3 at the same time. That is, you could take the first card, circle an answer, and immediately use that card to start one pile. Then take the second card, mark an answer on it, and then put it in a pile. And so on.

Do not hesitate to rearrange the cards, moving them from pile to pile as needed.

Step 4--When you are satisfied with your sorting, you must assign a number to each pile. This number expresses the probability, for each card in the pile, that you have indeed circled the correct alternative. This number expresses numerically the degree of certainty or uncertainty that you feel about each of the cards in the pile.

The number you assign to each pile may be any number from .5 to 1.0. ".5" means that, for each card in the pile, you felt completely uncertain as to which of the two answers is the correct answer. The number ".6" means that for each card in the pile, you felt 60% sure that you selected the correct answer and so forth. The number "1.0" means that you are completely sure that you have selected the correct answer for every card in the pile.

* All the cards in one pile must be assigned the same probability.

* Every pile must have a different probability.

* You must use numbers from .5 to 1.0 inclusive, but you may pick any numbers from that range that seem appropriate. You do not have to use the numbers 1.0 and .5, but you may if they adequately express your degree of certainty/uncertainty for your most extreme piles, the ones you feel least and most confident about.

* You may use two-digit numbers (like .55 or .75) if you wish.

* Do not use numbers like .4 or 1.2 that are outside the range .5 to 1.0.

Steps 5 and 6 told subjects how to write their responses, reemphasizing several key points and informing them that they would have 40 minutes to complete the task. In studies using the usual numerical response format, answering 50 questions typically consumes about 15 minutes, once instructions have been completed.

Items. In order to facilitate comparisons between these responses and those produced by the usual response format, an item set was used that had been tested previously on subjects drawn from the same pool. Specifically, it was the "complete test/hard items" set, reported in Figure 9 of Lichtenstein and Fischhoff [15]. Subjects there knew the answers to 61.7% of the items and responded with a mean probability of .758, reflecting substantial overconfidence.

Administration. One hundred seventy-five individuals participated, distributed over the four experimental groups according to their preference for the time at which the different groups were conducted. This task was the first of several unrelated tasks presented in sessions lasting approximately 1-1/2 hours. Subjects were paid $6, and were recruited through an advertisement in the University of Oregon student newspaper.

Results

Response usage. When the original group of subjects [15] responded to these items, the great majority (35 of 48) used six response categories. All but one of these individuals used the six "natural" responses (.5, .6, .7, .8, .9, 1.0). In the entire group of 48, all but two subjects used .5, indicating

"guess;" all but one used 1.0, indicating complete confidence. The bottom rows of Table 1 describe the responses of those subjects, both for the entire group and for those who used just six response categories.

The top four lines of Table 1 show how the subjects in the present experiment coped with the constraint of not being able to make all possible responses. For those who sorted into six piles, this should have been a minimal constraint. In some respects, they used the response scale similarly to the unconstrained subjects. In particular, most availed themselves of the .5 and 1.0 options. Nonetheless, the constraint did have some effect, in that 22 of the 45 six-pile subjects did not use the six "natural" responses, preferring other intermediate values between .5 and 1.0. The subjects who were allowed five categories typically gave up one of the intermediate responses, rather than one of the extreme responses, each of which was still used by about 90% of the subjects. The most pronounced effect was seen with the three-pile group which significantly reduced usage of 1.0.[1] Usage of .5, however, remained very high, indicating that "guess" was a more essential response than "certain." When subjects in the five- and six-pile groups (and in the original study) failed to use 1.0, their highest response was always in the .90-.99 range. A number of the subjects in the three- and four-pile groups had highest responses less than .9.

Performance. Given these differences in response usage, there is some reason to expect differences in performance. Figure 1 and the remainder of Table 1 provide pertinent details. The calibration curves in Figure 1 show the percentage of correct responses associated with each level of confidence expressed by subjects (after collapsing those expressions into the categories, .5-.59, .6-.69, .7-.79, .8-.89, .9-.99, and 1.0 and representing each by its

mean). The similarities between these curves are more striking than are any differences. The curves for the various sort-and-label groups closely resemble one another; perhaps more importantly, they also resemble the curve for the unconstrained group from Lichtenstein and Fischhoff [15]. If the four sort-and-label groups are pooled, the resulting curve falls very close to the unconstrained group's curve. Sorting per se seems to have had no effect.

This conclusion is generally borne out by the summary statistics of Table 1. The proportion correct column suggests that the focus on probability assessment may have slightly reduced the attention subjects paid to question answering, as they answered 2.2% fewer questions correctly than the unconstrained group. This difference (.595 vs. .617) was not significant, however. Constrained subjects' mean confidence was correspondingly lower (.716 vs. .737). As a result, the sort and non-sort groups have similarly high levels of overconfidence, which is computed as the signed difference between mean confidence and proportion correct. The various groups expressed confidence that was too confident by .11 to .14 on the average. In all these results, the only significant difference was between the 3-pile group, which was less confident than the 5-pile and unconstrained groups.

"Calibration" is a statistic characterizing curves such as those in Figure 1. It is the mean squared distance between each point in a curve and the identity line representing perfect calibration, weighted by the number of responses summarized in each point. Ideally, it should be 0. These levels, too, are similar in the constrained and unconstrained groups, confirming the visual impression from the figure.

Certainty. In previous research, the most extreme overconfidence has been observed with responses of 1.0. All of these should be associated with

correct answers; however, it is not uncommon to find 20% of those answers being incorrect. The third column shows that the sorting procedure significantly reduced the percentage of subjects using 1.0 at all. As one would expect, the percentage of responses that were 1.0 was down sharply from the unconstrained group where they comprised one quarter of all responses. Nonetheless, saying 1.0 less frequently did not increase the probability of being correct when subjects did so. Constrained subjects were still wrong about 20% of the time when they expressed certainty that they were right. Perhaps the clearest evidence that present procedure changed the frequency but not the manner in which subjects used 1.0 may be seen in the 3-pile group. Only half of these subjects used that response at all. As a group, though, they used the response half as much, indicating that the rate of use among users was just as high as before. And the probability of being correct was the same, indicating further that manner of use was unchanged.

Fischhoff and MacGregor [6] observed in an unconstrained task that subjects who never used the 1.0 response were somewhat better calibrated than other subjects. This was not the case in the present study. Pooling across the four constrained groups, the 37 non-users of 1.0 were not appreciably better calibrated than the 138 users (figure not shown). Unfortunately for the sake of this comparison, non-users also had a lower proportion of correct answers than did users (.566 vs. .603). Because calibration typically deteriorates as task difficulty increases [15], comparisons are somewhat ambiguous when difficulty varies.

Discussion

Although the sorting task affected subjects' choice of responses, it does not seem to have affected the appropriateness of those responses. Perhaps the

only glimmer of an effect, although not significant, is the slight superiority of the groups using fewer categories. Subjects in the three-pile and four-pile groups had a bit better overall calibration than subjects using five- or six-piles, despite having a slightly lower percentage of correct answers. Considering the variety of ways in which the present task differed from its predecessors, this is a meager haul. Accepting it at face value would lead one to believe that the appropriateness of people's confidence cannot be improved by any of the changes from the usual assessment procedure embodied in the sorting task: focusing attention on confidence assessment, comparing knowledge levels on different items, reducing the number of responses used, and eliminating whatever implicit cues are provided by the usual response format.

Before accepting this conclusion, we decided to repeat the study using small-group rather than large-group administration, with the experimenter close at hand to answer any questions that arose. The tradeoff between these two designs is that proximity raises the risk of experimenter interference, but reduces the risk of subjects deviating from the prescribed task. Although subjects in Experiment 1 appeared to work quite hard, the groups were too large to ensure that every subject performed the task in the desired sequence. Group size may also have inhibited some subjects from asking clarifying questions regarding what might have seemed a moderately complicated procedure.

Experiment 2

Method

Experiment 2 repeated the three- and six-pile groups of Experiment 1 in order to see what, if any, effect would be obtained with the most extreme versions of the sort-and-label manipulations. Instead of large-group admin-

istration, groups of about five people were brought to a small conference room. The experimenter read the instructions with them, discussed any questions, and remained during the course of the task. The continual presence of the experimenter made it possible to ensure that subjects were following the instructions. The presence of other, hardworking subjects seemed to encourage them to do so. Every attempt was made to avoid any communication between subjects or any hint from the experimenter regarding what responses to use.

Subjects were recruited through the local state employment office. All had at least one year of higher education, making them generally comparable in educational background to the subjects in Experiment 1. Each individual was paid $8 for working two hours on completing this and a number of subsequent unrelated judgment tasks. Most subjects completed this task within 20 minutes, not including the 10-15 minutes required for the experimenter to read and discuss the instructions.

Results

Response usage. The basic patterns of Experiment 1 were repeated. Of the 30 six-pile subjects, only 9 did not use the natural responses (.5, ... , 1.0); of these 9, only three did not use one of the extreme categories (.5, 1.0). As before, three-pile subjects made somewhat less use of .5 and significantly less use of 1.0. They used a wide variety of response sets; even the most popular (.5, .7, 1.0) was chosen by only 5 people. Details appear in Table 2.

Performance. The various performance statistics show the sorting groups as a whole to be quite similar to the unconstrained group. Treated together, the constrained groups were slightly more often correct, and slightly more confident, leading to a similar level of overconfidence and similar overall

calibration. Performance when using 1.0 was also similar. Indeed, the one difference of note that does emerge is between the two sort groups. The three-pile group was better calibrated and less overconfident than the six-pile group, as can be seen in both the summary statistics of Table 2 in the graphic representation of Figure 2. The six-pile group here actually performed worse than the unconstrained group, most of whom used six responses spontaneously. Although these differences are statistically significant, the level of significance ($p<.05$) is too low to attribute very much importance to them, especially given the welter of non-differences observed here and in Experiment 2.

Unlike Experiment 1, subjects here who did not use 1.0 were somewhat better calibrated than those who did. Their calibration curves are compared in Figure 3, which replicates a pattern observed by Fischhoff and MacGregor [6]. The 47 subjects who used 1.0 expressed, on the average, slightly greater confidence in the correctness of their answers than the 12 who did not (.765 vs. .750), but got a smaller portion right (.619 vs. .647). As a result, users of 1.0 were more overconfident than non-users (.146 vs. .103). Although this does replicate a previous result, the sample size (12 subjects) and effect size are too small for it to be of more than speculative interest.

Discussion

The overall message of these data is that this rather drastic change in procedure had little effect on confidence assessment. The constraints of the procedure did induce sorting subjects to adopt somewhat different response patterns; however, the accompanying calibration was indistinguishable from that observed elsewhere. The only differences of any note were a weak suggestion that calibration may improve as the number of categories decreases,

and feeble support for the previous observation that people who do not use 1.0 tend to be better calibrated.

From a practical perspective, these results have both good news and bad news. The bad news is that despite a rather concerted effort, we were no more successful than our predecessors in devising a simple, nondirective scheme for improving the quality of confidence assessments. The good news is that this robustness of confidence assessments gives elicitors great freedom in how they extract assessments from themselves or their clients. From a large set of methods, elicitors can choose whatever method seems most comfortable and natural without fear that the method itself will bias the results.

From a theoretical perspective, these results are informative and even encouraging. They point to the robustness of confidence effects and the generality of previous results. We have also learned some things about the cognitive processes involved in confidence assessment. The fact that putting the emphasis on probability, rather than fact, assessment had no effect indicates that the probabilities are not neglected in the usual task. The fact that reading all items prior to responding had no effect indicates that even in the usual task, respondents are able to extract quickly as much information about the overall difficulty of the items they face as they can with a more thorough review. The fact that postponing the assignment of numbers had no effect indicates that producing quantitative assessments *per se* is not an obstacle. The fact that varying number of response possibilities had no effect indicates that the reason for poor calibration elsewhere was not that the natural response scale communicated a misleading message to subjects regarding the fineness of the discriminations that they can make between different degrees of uncertainty.

The sort-and-label procedure differed from traditional procedures on a number of dimensions. Had it had an effect, subsequent research would have been directed to assessing which dimension provided the effective element. Some of those dimensions are still of interest. For example, what determines how fine are the discriminations in level of knowledge that people believe they can make? How do people appraise the overall difficulty of a set of items and how does that appraisal affect how people create equivalence classes for feelings of knowing? Do they first make a crude partition (e.g., don't know, may know, certain) and then refine it into subsidiary categories, or do they build categories by matching items for which their knowledge levels seem equivalent? Variations on the present procedure could shed light on these questions. For the moment, though, the dominant impression is that confidence is determined by powerful psychological processes which have resisted the present attempts to manipulate them, just as they have resisted most previous efforts.

Reference Note

1. Lichtenstein, S. & Fischhoff, B. The effect of gender and instructions on calibration. Decision Research Report 81-5, 1981.

Footnotes

This research was supported by the Office of Naval Research under Contract N00014-80-C-0780 to Perceptronics, Inc. We thank Nancy Collins, Geri Hanson, and Peggy Roecker for much technical help. Correspondence may be addressed to Decision Research, 1201 Oak Street, Eugene, Oregon 97401.

1. All significant tests reported in the text and tables are t-tests conducted with the generalized jackknife procedure [17]. Unlike procedures that compute statistics on individual subjects, jackknifing computes statistics on the entire set of data after each subject is deleted. It is a suitable procedure when the sampling distribution is unknown or when statistics computed for individual subjects are unstable. A particular problem with calibration statistics, which jackknifing addresses, is that they are sensitive to the distribution of responses across categories. The fewer responses per category, the less stable the estimate of percentage correct, The higher the calibration score will tend to be. This would introduce an inevitable confound in comparisons between subjects using different numbers of categories (as they were forced to do here). Because the groups as a whole produced similar distributions of responses, this problem is greatly reduced with pooled data.

References

1. Beyth-Marom, R. How probable is probable? Numerical translation of verbal probability expressions. Journal of Forecasting, 1982, 1, 257-269.

2. Brown, S. R. Bibliography on Q technique and its methodology. Perceptual and Motor Skills, 1968, 26, 587-613.

3. Deffenbacher, K. A. Eye witness accuracy and confidence. Law and Human Behavior, 1980, 4, 243-260.

4. Ferrell, W. R. & McGoey, P. J. A model of calibration of subjective probabilities. Organizational Behavior and Human Performance, 1980, 26, 32-53.

5. Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

6. Fischhoff, B. & MacGregor, D. Subjective confidence in forecasts. Journal of Forecasting, 1982, 1, 155-172.

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

8. Fisher, G. W. Conceptual models, judgment, and the threat of uncertainty in nuclear threat assessment. Journal of Social Issues, in press.

9. Glucksberg, S. & McCloskey, M. Decisions about ignorance: Knowing that you don't know. Journal of Experimental Psychology: Human Learning and Memory, 1981, 7, 311-325.

10. Gottschalk, L. A. & Averbach, A. H. (Eds.), Methods of research in psychotherapy. New York: Appleton-Century-Crofts, 1966.

11. Gray, C. A. Factors in student's decisions to attempt academic tasks. Organizational Behavior and Human Performance, 1975, 13, 147-165.

12. Johnson, W. B. The impact of confidence interval information on probability judgments. Accounting Organizations and Society, 1982, 7, 349-367.

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

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

15. Lichtenstein, S. & Fischhoff, B. Training for calibration. Organizational Behavior and Human Performance, 1980, 26, 149-171.

16. Lichtenstein, S., Fischhoff, B. & Phillips, L. D. Calibration of probabilities: 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, 1982.

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

18. Murphy, A. H. & Winkler, R. L. Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? National Weather Digest, 1977, 2, 2-9.

19. Poulton, E. C. Quantitative subjective assessments are almost always biased, sometimes completely misleading. British Journal of Psychology, 1977, 68, 409-425.

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

21. Roos, M., Hietanen, M. & Luoma, J. A new procedure for averaging particle properties. Physica Fennica, 1975, 10, 20-33.

22. U.S. Nuclear Regulatory Commission. Fault tree handbook. NUREG-0492. Washington, D. C.: Author, 1981.

23. Wallsten, T. & Budescu, D. Encoding subjective probabilities. Management Science, in press.

24. Wells, G. L. Applied eyewitness testimony research: Systems variables and estimator variables. Journal of Personality and Social Psychology, 1978, 36, 1546-1557.

25. Wilkinson, A. C. Partial knowledge and self-correction: Developmental studies of a quantitative concern. Developmental Psychology, 1982, 18, 876-893.

Table 1

Summary Statistics

Experiment 1*

Group	N	Percentage Using .5	Percentage Using 1.0	Prop. Correct	Mean Confidence	Over-Confidence	Calibration	1.0 Responses % of Total	1.0 Responses % Correct
Sort-and-label									
3 piles	50	88.0	50.0[a,b,c]	.586	.694[a,b]	.109	.0212	12.7[a,b,c]	77.1
4 piles	42	88.0	80.0[a]	.592	.715	.122	.0236	21.6[a]	77.4
5 piles	38	92.1	89.5[b]	.601	.743[a]	.142	.0281	21.6[b]	81.0
6 piles	45	93.3	91.1[c]	.604	.717	.113	.0261	14.5[d]	78.7
All Sort	175	90.3	76.6[d]	.595	.716	.120	.0235	17.2[e]	78.6
Unconstrained									
All**	48	95.8	97.9[a,d]	.617	.737[b]	.121	.0238	24.6[c,d,e]	80.4
Using 6 Categ.	35	100	100	.625	.746	.121	.0227	23.6	82.8

* Entries with a common superscript by column differ significantly ($p<.01$). All tests are jackknifed t-tests.

** Data from Lichtenstein & Fischhoff [15].

Table 2

Summary Statistics

Experiment 2*

Group	N	Percentage Using .5	Percentage Using 1.0	Prop. Correct	Mean Confidence	Over-Confidence	Calibration	1.0 Responses % of Total	1.0 Responses % Correct
				Sort-and-label					
3 piles	29	97.3	65.5[a,b]	.639	.749	.110[a]	.0212[a]	20.9	78.4
6 piles	30	90.0	93.3[a]	.612	.778[a]	.166[a,b]	.0411[a,b]	27.1	79.4
All Sort	59	84.8	79.7[c]	.625	.764	.139	.0262	24.1	79.0
				Unconstrained					
All**	48	95.8	97.9[b,c]	.617	.737[a]	.121[b]	.0238[b]	24.6	80.4

* Entries with a common superscript by column differ significantly ($p<.01$); those for mean confidence, overconfidence, and calibration differ at $p<.05$.

** Data from Lichtenstein & Fischhoff [15].

Figure Captions

Figure 1. Calibration curves for the 3-, 4-, 5-, 6-pile groups of Experiment 1, compared with the calibration of subjects in Figure 9 of Lichtenstein and Fischhoff [14].

Figure 2. Calibration curves for the 3-pile and 6-pile groups of Experiment 2, compared with subjects from Lichtenstein and Fischhoff [14].

Figure 3. Calibration curves for users and non-users of 1.0 in Experiment 2. (Note: For the non-users group, the data in the range .6-.69 comprised so few cases that they were aggregated with the data in the range .5-.59.)

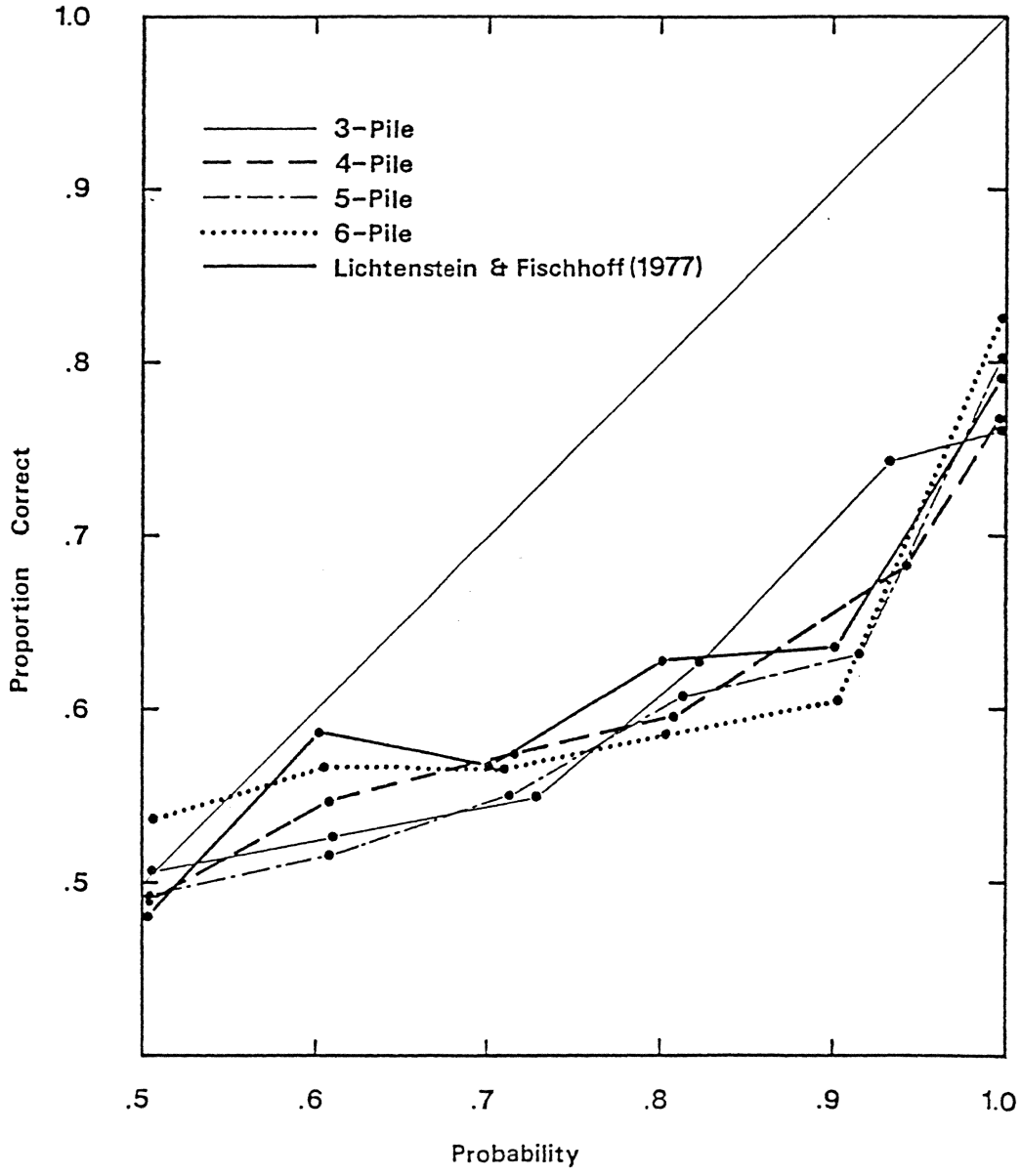
1.0
.9
.8
.7
.6
.5
Proportion Correct
3-Pile
4-Pile
5-Pile
6-Pile
Lichtenstein & Fischhoff (1977)
.5
.6
.7
.8
.9
1.0
Probability

Figure 1.

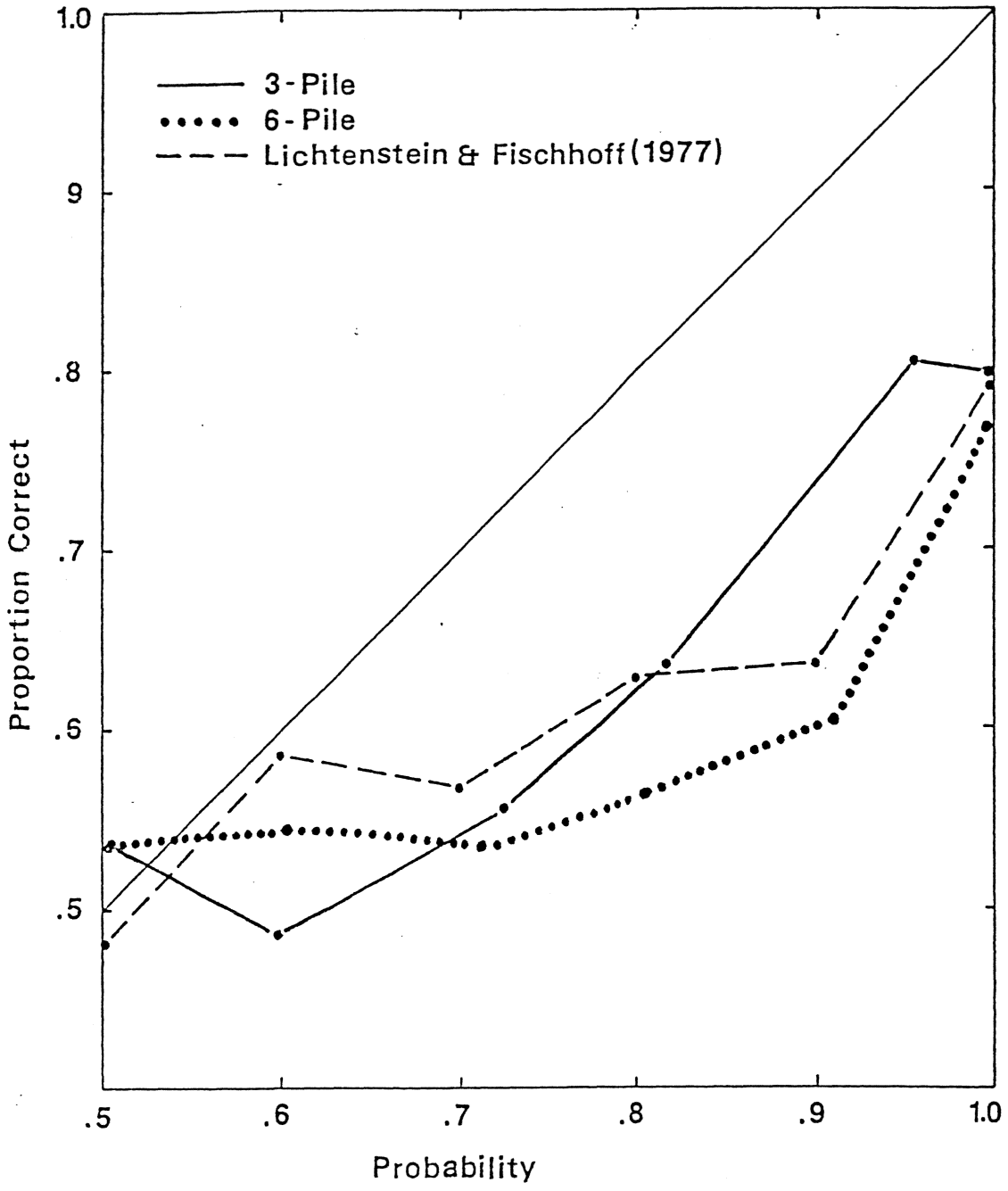
1.0
9
.8
.7
.6
.5
Proportion Correct
3-Pile
6-Pile
Lichtenstein & Fischhoff (1977)
.5
.6
.7
.8
.9
1.0
Probability

Figure 2.

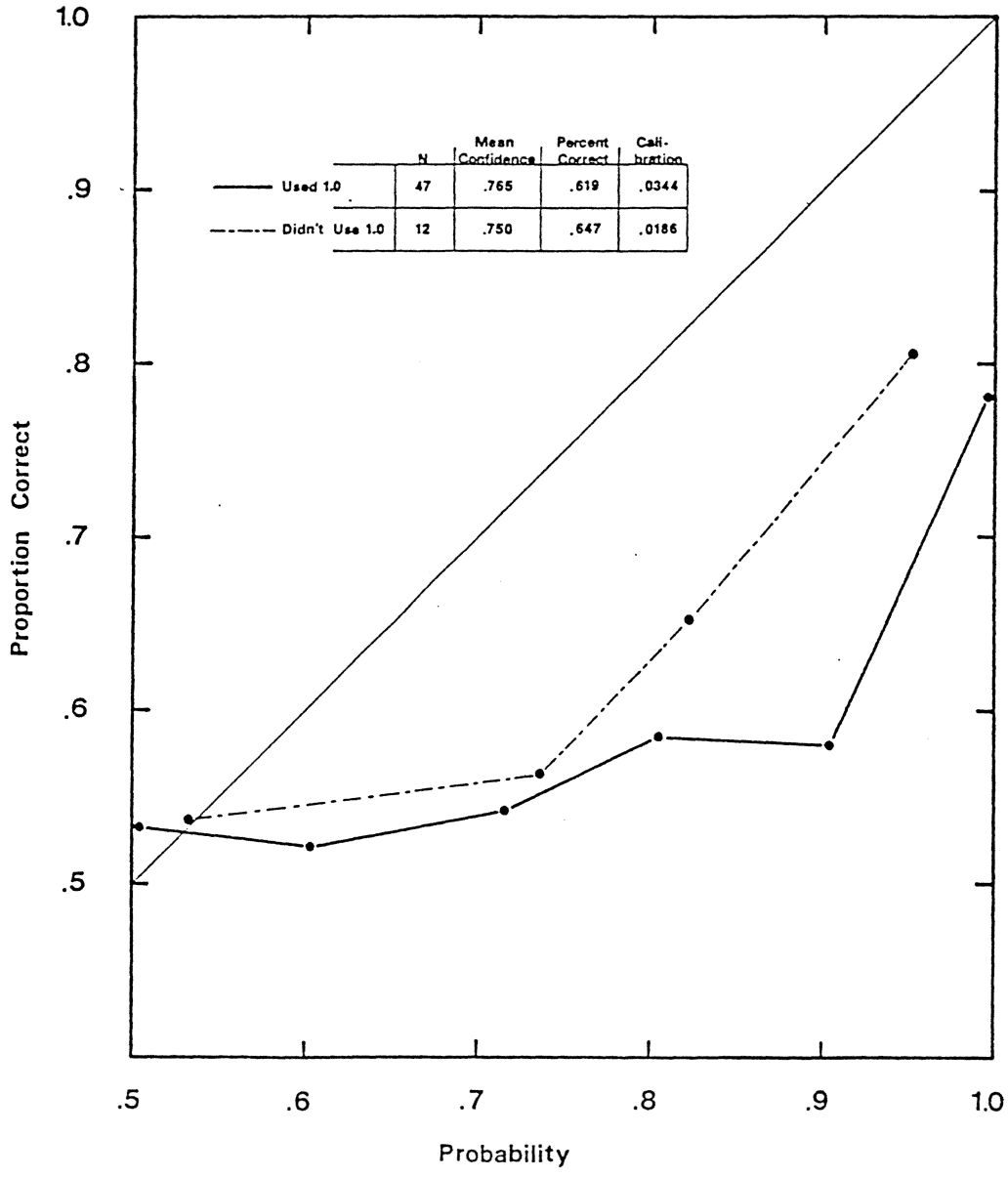
1.0
.9
.8
.7
.6
.5
Proportion Correct
.5
.6
.7
.8
.9
1.0
Probability
N
Mean Confidence
Percent Correct
Cali-bration
Used 1.0
47
.765
.619
.0344
Didn't Use 1.0
12
.750
.647
.0186

Figure 3.