(12)

TECHNICAL REPORT PTR-1092-82-7
CONTRACT NUMBER: N00014-80-C-0150
WORK UNIT NUMBER: NR197-064
July 1982

CATEGORICAL CONFIDENCE

Baruch Fishhoff
Don MacGregor
Sarah Lichtenstein

Prepared for:

OFFICE OF NAVAL RESEARCH
800 North Quincy Street
Arlington, VA 22217

PERCEPTRONICS
6271 VARIEL AVENUE • WOODLAND HILLS • CALIFORNIA 91367 • PHONE (213) 884-7470

NOTES

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any office of the United States Government.

12

TECHNICAL REPORT PTR-1092-82-7
CONTRACT NUMBER: N00014-80-C-0150
WORK UNIT NUMBER: NR197-064
July 1982

CATEGORICAL CONFIDENCE

Baruch Fishhoff
Don MacGregor
Sarah Lichtenstein

Prepared for:

OFFICE OF NAVAL RESEARCH
800 North Quincy Street
Arlington, VA 22217

PERCEPTRONICS
6271 VARIEL AVENUE • WOODLAND HILLS • CALIFORNIA 91367 • PHONE (213) 884-7470

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A120 555 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Categorical Confidence	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER PTR-1092-82-7
7. AUTHOR(s) Baruch Fischhoff, Don MacGregor and Sarah Lichtenstein	8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0150
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Research A Branch of Perceptronics 1201 Oak Street, Eugene, Oregon 97401	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217	12. REPORT DATE July 1982
	13. NUMBER OF PAGES 16
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

decision making
confidence
calibration
probability assessment

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

People tend to be inadequately sensitive to the extent of their own knowledge when asked to assess the probability that each of their answers to a set of questions is correct. This insensitivity typically emerges as overconfidence. That is, their assessments are typically too high compared to the portion of items they get right. Few prescriptions have proven effective against this problem. Those that have worked might be thought of a directive in character. Rather than improving subjects' feeling for how much they know, they may have suggested to subjects how their probability assessments should have changed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

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 seemingly non-directive tack. 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. Even though this procedure differed from its predecessors in many respects, performance here was indistinguishable from that observed elsewhere. Though small pockets of improvement were noted, confidence was largely resistant to this manipulation. Some implications of these results for attempts to study confidence and eliminate overconfidence are discussed.

Table of Contents

List of Tables

List of Figures

Summary

When asked to assess the probability that each of their answers to a set of questions is correct, people typically exhibit overconfidence; the proportion of answers correct for the probability values are too small. The present study attempted to improve the appropriateness of confidence judgments by having people sort their responses to a group of general knowledge items into a number of piles, each reflecting a different level of confidence in their answers. However, this procedure had no consistent effect on overconfidence, even though it differed in many ways from previous unsuccessful efforts to reduce this bias. Implications for future studies of the overconfidence phenomenon are discussed.

Categorical Confidence

In a typical probability assessment task, participants first ponder some question of fact and then assess the likelihood that the answers they have produced are correct. Casual 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 spring to mind: (a) answers are harder to produce than probabilities; therefore they require more time, (b) we are more experienced in answering questions; hence, we can spend more time profitably on that task, (c) until an answer is produced, one cannot even begin to assess its accuracy, or (d) we are more accustomed to having our answers evaluated than our 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 Fischhoff, 1982; Lichtenstein, Fischhoff & Phillips, 1982). People have, moreover, considerable confidence in these confidence assessments (e.g., Fischhoff, Slovic & Lichtenstein, 1977).

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., Lichtenstein & Fischhoff, 1980; Murphy & Winkler, 1977). 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 (Koriat, Lichtenstein & Fischhoff, 1980).

The secret of even these partial successes is, however, still unclear. It would be theoretically interesting and 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. Because one does 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.

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 (1980) 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 (Lichtenstein & Fischhoff, 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 more than on the production of some 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.

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 to 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.

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.

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 that 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 numerical 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 by Lichtenstein and Fischhoff (1977). Subjects there knew the answers to 61.7% of the items, and responded with a mean probability of .758, reflecting substantial overconfidence.

Subjects. 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 1½ 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 (Lich-

tenstein & Fischhoff, 1977) responded to these items, the great majority (35 of 48) used six response categories. Moreover, all but one of these individuals used the six "natural" responses (.5, .6, .7, .8, .9, 1.0). In the entire group, 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 these subjects, both for the entire group and for those who used just six response categories. The first columns are devoted to response usage.

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. Indeed, most did avail 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 chose to give up one of the intermediate responses, rather than one of the extreme responses, each of which was still used by 92.1% of the subjects. The increasing constraints on the four-pile and three-pile groups led to reduced usage of 1.0, but not of .5. That is, "guess" proved to be 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 the mean confidence for each level of confidence expressed by subjects (after collapsing those

Table 1

Summary Statistics

Experiment 1

Group	N	Percentage Using		Prop. Correct	Mean Confidence	Over-confidence	Calibration	1.0 Responses	
		.5	1.0					% of Total	% Correct
Sort-and-label									
3 piles	50	88.0	50.0	.586	.694	.109	.0212	12.7	76.3
4 piles	42	88.0	81.0	.592	.715	.122	.0236	21.6	76.8
5 piles	38	92.1	92.1	.601	.743	.142	.0281	21.6	80.4
6 piles	45	93.3	91.1	.604	.717	.113	.0261	14.4	82.5
Unconstrained[a]									
All	48	95.8	97.9	.617	.737	.121	.0238	24.6	79.1
Using 6 Categ.	35	100	100	.625	.746	.121	.0227	23.6	82.8

[a] Data from Lichtenstein & Fischhoff, 1977.

expressions into the categories, .5-.59, .6-.69, .7-.79, .8-.89, .9-.99, and 1.0). 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 (1977). 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; the mean for all sort groups was .595, compared with .617 for the unconstrained group. Their mean confidence was correspondingly lower (.717 vs. .737). As a result, the sort and non-sort groups have similarly high levels of overconfidence, which is computed as the signed differences between mean confidence and proportion correct. The various groups expressed confidence that was too confident by .11 to .14 on the average.

"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 sort groups and unconstrained group, confirming the visual impression from the figure.

Certainty. The most extreme overconfidence has typically been observed with responses of 1.0, all of which should be associated with correct answers. The final two columns show that the sorting procedure did reduce the usage of 1.0 (as was shown by the third column), which comprised one quarter of the unconstrained group's responses. However, it did not affect the cor-

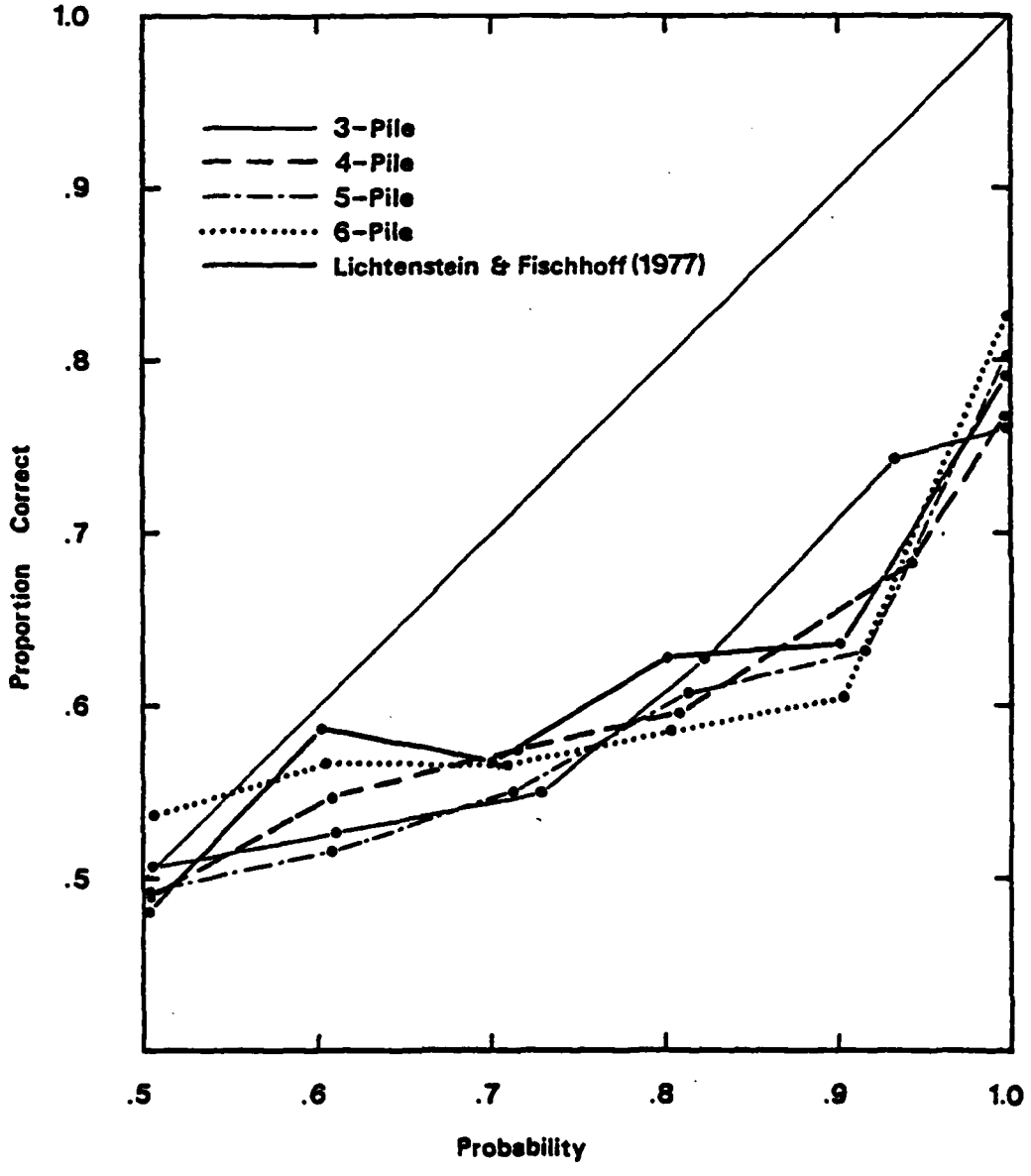

Figure 1. Calibration curves for the 3-, 4-, 5-, and 6-pile groups of Experiment 1, compared with the calibration of subjects in Figure 9 of Lichtenstein and Fischhoff (1977).

rectness of the associated answers. Subjects were still wrong about 20% of the time when they expressed certainty that they were right.

Fischhoff and MacGregor (in press) observed in an unconstrained task subjects who never used the 1.0 responses were somewhat better calibrated than other subjects. This was not the case in the present study. 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 of 1.0 also had a lower proportion of correct answers than did users (.566 vs. .603). Because calibration typically deteriorates as task difficulty increases (Lichtenstein & Fischhoff, 1977), 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 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. Although such proximity raises slightly the risk of

experimenter interference, it also reduces the risk that subjects deviated 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 administration, 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.

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 considerably 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. Though the proportion of correct answers for both sort groups was slightly superior to the unconstrained group, this difference was also reflected in a somewhat higher level of confidence for the sort groups. The sorting and unconstrained groups were compatible on the remaining measures. 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. This can be seen in the summary statistics of Table 2 and 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.

A second modest effect is that the 22 subjects (20 from the three-pile group and two from the six-pile group) who did not use 1.0 were somewhat better calibrated than the 47 who did. Their calibration curves are compared in Figure 3. Those who used 1.0 expressed, on the average, slightly greater confidence in the correctness of their answers than those 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).

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 are 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.

Table 2

Summary Statistics

Experiment 2

Group	N	Percentage Using		Prop. Correct	Mean Confidence	Over-confidence	Calibration	1.0 Responses	
		.5	1.0					% of Total	% Correct
Sort-and-label									
3 piles	29	82.8	65.5	.639	.749	.110	.0212	20.9	79.8
6 piles	30	93.3	93.3	.612	.778	.166	.0411	27.7	76.8
All sort	59	88.1	79.7	.625	.764	.139	.0262	24.0	78.1
Unconstrained[a]									
All	48	95.8	97.9	.617	.737	.121	.0238	24.6	79.1

[a] Data from Lichtenstein & Fischhoff, 1977.

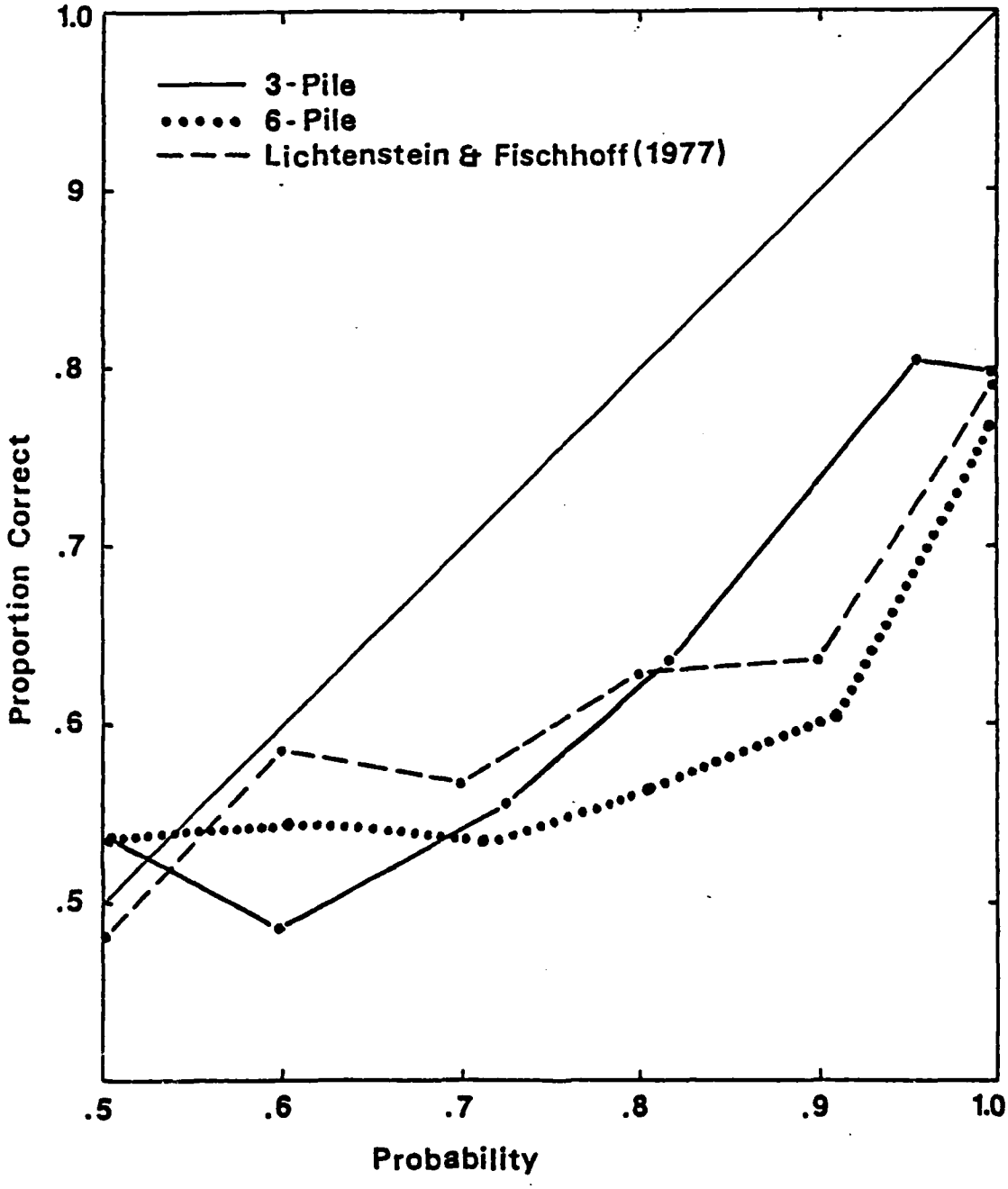

Figure 2. Calibration curves for the 3-pile and 6-pile groups of Experiment 2, compared with subjects from Lichtenstein and Fischhoff (1977).

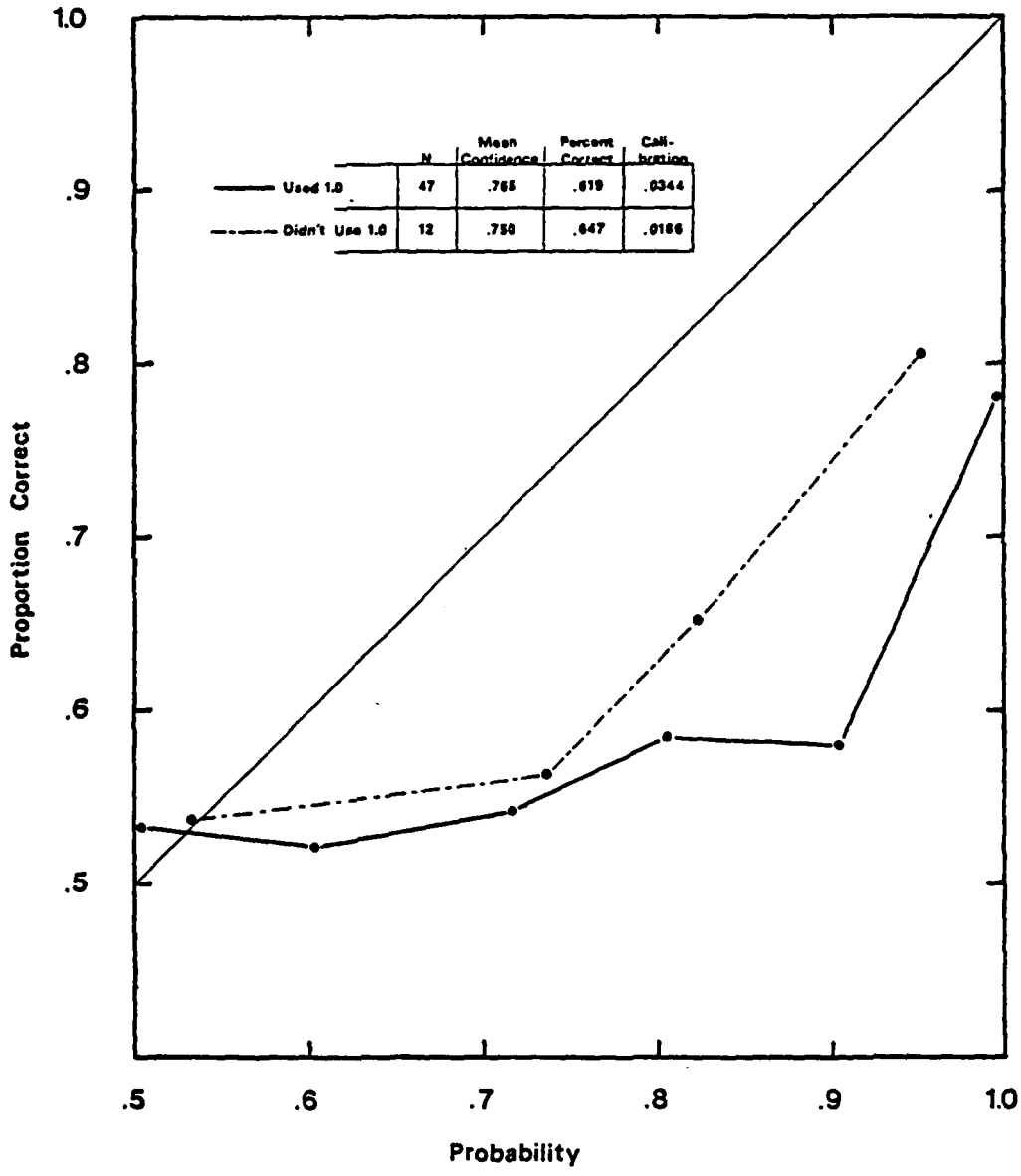

Figure 3. Calibration curves for users and non-users of 1.0 in Experiment 2. (Note: For the non-users group, the data in the range .6-.69 comprised so few cases that they were aggregated with the data in the range .5-.59.)

From a practical perspective, these results are disappointing. Despite a rather concerted effort, we were no more successful than our predecessors in devising a simple scheme for improving the quality of confidence assessments. From a theoretical perspective, however, such negative results are informative and even encouraging. They point to the robustness of confidence effects and the generality of previous results.

As noted in the introduction, 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? 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 Notes

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

References

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

Fischhoff, B., & MacGregor, D. Subjective confidence in forecasts. Journal of Forecasting, in press.

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

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

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.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

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

Acknowledgement

This research was supported by the Office of Naval Research under Contract N00014-80-C-0780 to Perceptronics, Inc. We thank Nancy Collins, Gerry Hanson, and Peggy Roecker for much technical help. Correspondence may be addressed to either Don MacGregor or Sarah Lichtenstein at Decision Research, 1201 Oak, Eugene, Oregon 97401; or to Baruch Fischhoff, MRC/APU, 15 Chaucer Road, Cambridge, CB1 1EF, U.K.

OFFICE OF NAVAL RESEARCH

TECHNICAL REPORTS DISTRIBUTION LIST

CDR Paul R. Chatelier
Office of the Deputy Under Secretary
of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Engineering Psychology Programs
Code 422
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217 (5 cys)

Manpower, Personnel & Training
Programs
Code 270
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Operations Research Programs
Code 411-OR
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics & Probability Program
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Systems Program
Code 411-IS
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

CDR K. Hull
Code 410B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Physiology & Neuro Biology Programs
Code 441
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Commanding Officer
ONR Eastern/Central Regional Office
ATTN: Dr. J. Lester
Bldg. 114, Section D
666 Summer Street
Boston, MA 02210

Commanding Officer
ONR Western Regional Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
American Embassy, Room A-407
APO San Francisco, CA 96503

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D.C. 20375

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis & Evaluation
Group
Naval Training Equipment Center
Code N-OOT
Orlando, FL 32813

Dr. Albert Colella
Combat Control Systems
Naval Underwater Systems Center
Newport, RI 02940

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Dr. A.L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

Mr. Arnold Rubinstein
Naval Material Command
NAVMAT 0722 - Rm. 508
800 North Quincy Street
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 340F
Washington, D.C. 20361

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bathesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Department
Naval Medical Research Institute
Bethesda, MD 20014

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Dr. James McGrath
CINCLANT FLT HQS
Code 04E1
Norfolk, VA 23511

Navy Personnel Research &
Development Center
Planning & Appraisal Division
San Diego, CA 92152

Dr. Robert Blanchard
Navy Personnel Research &
Development Center
Command & Support Systems
San Diego, CA 92152

LCDR Stephen D. Harris
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA L8974

Dr. Julie Hopson
Human Factors Eningeering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

CDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, CA 93940

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations &
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks, AFB, TX 78235

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
ENGLAND

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsview, Ontario M3M 3B9
CANADA

Dr. A.D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
ENGLAND

Dr. Robert T. Hennessy
NAS-National Research Council
2101 Constitution Ave., N.W.
Washington, D.C. 20418

Dr. M.G. Samet
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91367

Dr. Robert Williges
Human Factors Laboratory
Virginia Polytechnic Institute
& State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Alphonse Chapanis
Department of Psychology
The Johns Hopkins University
Charles & 34th Streets
Baltimore, MD 21218

Dr. Ward Edwards
Director, Social Science Research
Institute
University of Southern California
Los Angeles, CA 90007

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, D.C. 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Christopher Wickens
University of Illinois
Department of Psychology
Urbana, IL 61801

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 cys)

Dr. Judith Daly
System Sciences Office
Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goleta, CA 93017

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, CO 80309

Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, VA 22202

Dr. T. B. Sheridan
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Harry Snyder
Department of Industrial Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. W. S. Vaughan
Oceanautics, Inc.
422 6th Street
Annapolis, MD 21403

Dr. Richard W. Pew
Information Sciences Division
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Hillel Einhorn
University of Chicago
Graduate School of Business
1101 E. 58th Street
Chicago, IL 60637

Dr. John Payne
Duke University
Graduate School of Business Administration
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Andrew P. Sage
University of Virginia
School of Engineering and
Applied Science
Charlottesville, VA 22901

Dr. Leonard Adelman
Decisions and Designs, Inc.
8400 Westpark Drive, Suite 600
P.O. Box 907
McLean, VA 22101

Dr. Lola Lopes
Department of Psychology
University of Wisconsin
Madison, WI 53706

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Industrial Park
Burlington, MA 01803

Dr. Rex Brown
Decision Science Consortium
Suite 721
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090