MEMORANDUM
RM-6118-PR
MARCH 1970

THE DELPHI METHOD, IV: EFFECT OF PERCENTILE FEEDBACK AND FEED-IN OF RELEVANT FACTS

N. Dalkey, B. Brown and S. Cochran

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

THIS DOCUMENT HAS BEEN APPROVED FOR PUBLIC RELEASE AND SALE; ITS DISTRIBUTION IS UNLIMITED.

This study is presented as a competent treatment of the subject, worthy of publication. The Rand Corporation vouches for the quality of the research, without necessarily endorsing the opinions and conclusions of the authors.

Published by The RAND Corporation

MEMORANDUM
RM-6118-PR
MARCH 1970

THE DEPHI METHOD, IV: EFFECT OF PERCENTILE FEEDBACK AND FEED-IN OF RELEVANT FACTS

N. Dalkey, B. Brown and S. Cochran

This research is supported by the United States Air Force under Project RAND—Contract No. F41620-67-C-0045—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the United States Air Force.

DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

PREFACE

This Memorandum is one of a series reporting the results of a set of experiments evaluating the Delphi techniques for formulating group judgments. It augments the results reported in RM-5888-PR, RM-5957-PR and RM-6115-PR.

The primary goal of the studies reported in these RM's is the design of improved techniques for the use of expert opinion by decisionmakers. For a number of basic military concerns, the best information available is the judgment of knowledgeable individuals. This is especially true in the assessment of long-range technological developments, and the evaluation of long-range future threats. Thus, the military has an important stake in ensuring that the procedures used for obtaining judgments are adequately designed to elicit the most accurate estimates possible from the community of experts.

Experimental studies at RAND have demonstrated that there are identifiable deficiencies in the ways in which expert judgment has been used in the past. The studies have also shown that it is possible to design improved procedures that guard against some of these deficiencies, and that can increase the accuracy of judgments furnished by a group of experts.

Studies making use of the procedures developed at RAND have been made, or are being made, in each of the military services, as well as the DIA and CIA. Problems

being approached by these techniques include, in addition to long-range technological developments and threat assessment, forecasting innovations in reliability and maintenance, and evaluation of development programs. The techniques have also been employed by numerous industrial contractors, including TRW, Martin-Marietta, McDonnell Douglas, and Monsanto Chemical, primarily for long-range technological forecasting as an input to planning.

Professor Samuel Cochran of the Department of Psychology, East Texas State University, Commerce, Texas, is a consultant to The RAND Corporation.

SUMMARY

Two variations in the form of feedback in Delphi exercises are investigated: the effect of feeding back the percentile location of an individual's answer in the group distribution and the effect of presenting the respondent with a single additional relevant fact.

Experiments were set up to compare the improvement in responses to 20 general information type questions between two groups that received different feedback.

In the first experiment, the control group was fed back the median and the quartiles of the group response; the experimental group received individual percentile ratings of Round 1 answers. The results show no difference in either the number of improved responses or in the amount of improvement between the groups. In addition, the control and experimental groups show no difference in the number of changes made in Round 2.

In the second experiment both groups were fed the median and quartiles of the group response, and, in addition, the experimental group received one relevant fact for each question. The result was decisive improvement in the experimental group in both the number of questions showing improved accuracy and in the amount of numerical improvement. This is the first experiment in the series in which numerical improvement in accuracy has reached statistical significance. The fact that the number of responses which

were changed by the experimental group was also much greater lends credence to the hypothesis that the introduction of a relevant fact strengthens the motivation for revising.

A comparison of the performance of men and women confirms earlier results. Females are less accurate on their initial estimates but they are able to make as good, or better use of both feedback and feed-in to improve their estimates. Given statistical feedback the women's responses in Round 2, though greatly improved, are still less accurate than the men's in Round 1. However, when relevant facts are introduced, the women's revised responses are significantly better than the men's Round 1 responses and are not significantly worse than their Round 2 responses.

CONTENTS

I. INTRODUCTION

One of the basic features of the Delphi procedures for formulating group opinion is iteration with controlled feedback. The iteration step is generally associated with convergence (smaller dispersion of answers on the second round) and increased accuracy on a majority of questions for which answers change. In general, a sizeable fraction of the answers do not change upon iteration.

In the most elementary form of Delphi exercise, the only material fed back between iterations is some statistic of the group's responses on the previous round--e.g., the median and the upper and lower quartiles. There is obviously a very large number of potential variants on this procedure. In previous experiments [1] the variants investigated were (1) no feedback at all, which led to no improvement on Round 2; (2) feedback of reasons for extreme responses (those outside the interquartile range) which led to decreased accuracy on Round 2; and (3) feedback of supplementary relevant opinions from the group which produced no clear-cut results. In this Memorandum we record the results of two additional experiments in which other forms of feedback were investigated.

Because of the virtually limitless set of possibilities, it was considered desirable to have in mind specific hypotheses suggested by previous results, which would be given relatively definitive tests by the experiments. One such

hypothesis is expressed in [1]. In experiments where the median and quartiles are fed back, the respondents exhibit overconvergence, i.e., the ratio of error to standard deviation increases between Round 1 and Round 2. This result suggests that a form of feedback that presents a less specific "target" for the changes might lead to greater accuracy on Round 2. It was hypothesized that feeding back simply the percentile location of the individual's response within the group would be sufficiently less specific than the median, but still convey enough information to motivate revised estimates. Hence, one hypothesis to be tested was that feeding back percentile locations would lead to greater improvement than feeding back medians and quartiles.

The second hypothesis needs some introductory remarks. It is generally taken for granted that the more information on a given question available to an individual, the more accurate his answer will be. However, results reported in [1] indicate that little improvement, or even degradation, will result if the information is in the form of additional opinions on the part of the group. This left open the question of the effect of furnishing additional information in the form of "hard facts." Since we do not have a clear picture of the kind of information that is self-generated by an individual member of the group, there is no way to predict how the self-generated material will interact with

externally supplied information. There was the possibility that for many questions an externally supplied fact would conflict with self-generated material, and act as noise in the production of an estimate. The hypothesis to be tested, then, was that feed-in of a single "hard fact" on Round 2 would lead to across-the-board improvement on all questions.

The two experiments described below show that only the second of these hypotheses is correct. Feeding back the percentile location of a respondent does not lead to greater improvement than feeding back the median and two quartiles. On the other hand, feeding in a single relevant hard fact leads to a dramatic improvement in the accuracy of responses on Round 2 almost independently of the kind of additional fact.

Although only one of the two original hypotheses was confirmed by the experiments, both results are significant with regard to the design of Delphi procedures. The percentile feedback result suggests that Delphi exercises in the closed information mode--i.e., with no feed-in of external information--are relatively insensitive to the form of feedback. The outcome of the relevant fact experiment raises some basic questions concerning the interaction of Delphi panels and external sources of information.

II. COMPARISON OF PERCENTILE AND QUARTILE FEEDBACK

In previous investigations of Delphi procedures, there has been fairly good evidence that the amount of convergence attendant upon feedback is excessive in the simple sense that the bias (average error divided by standard deviation) of the group's responses increases between Round 1 and Round 2. Figure 1 is taken from [1] and shows the relationship between average (log) error and standard deviation for both Round 1 and Round 2. The line expressing the relationship for Round 2 lies uniformly above the line for Round 1. These results were obtained where the information fed back between rounds was the median and the two quartiles of the group's answers.

This consideration suggests the possibility that weakening the tendency toward convergence might allow fuller play for "rethinking" on the second round. (Compare [1] pp. 46-48.) One possibility for reducing the amount of convergence, but at the same time giving the respondent useful information about the location of his response in the group distribution of responses, would be to report to each respondent his percentile rank, i.e., the percent of answers that were numerically less than his. Speaking loosely, this piece of information might be expected to remove the hard "target" for those who change their answers; but at the same time create motivation for change, especially for those respondents at the extremes of the distribution.

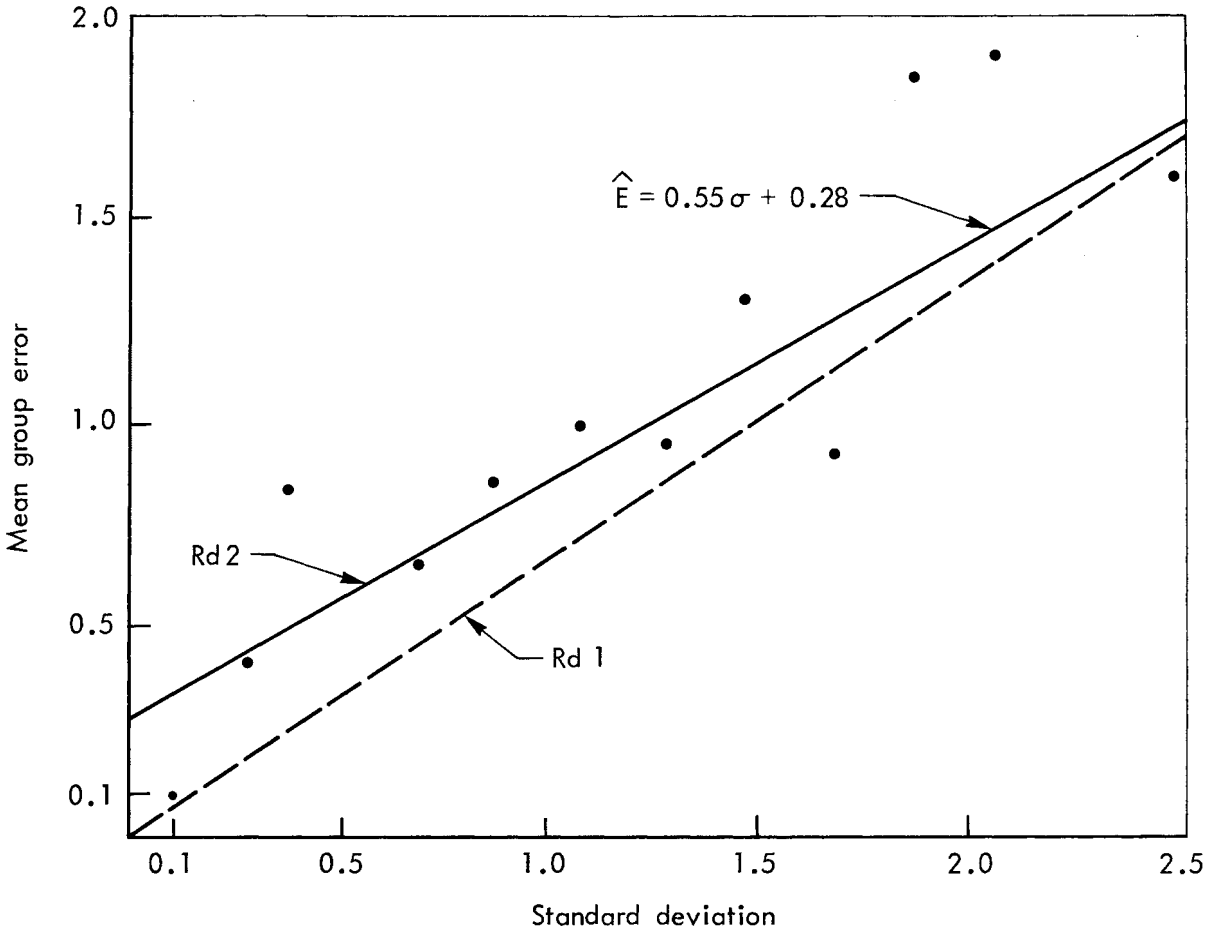

Fig. 1— Bias, Round 2

PURPOSE

The purpose of the experiment was to test the hypothesis that feedback of individual percentiles on each question would produce more improvement than feedback of the median and quartiles.

METHOD

One hundred thirty-seven students from the University of California at Los Angeles were paid to serve as respondents in this experiment. Of these, 67 were male, 70 female; 33 were graduate students, 104 upper division students. Eight groups of 15 to 20 respondents were formed; four of these were designated as control groups (Nos. 1, 3, 5, and 7) and four experimental (Nos. 2, 4, 6, and 8). In March 1969, on each of four days, one of the control groups and one of the experimental groups were tested. The design of the experiment was as follows:

	Control group	Experimental group
Round 1	Give self-rating for each question. Answer 20 questions.	Give self-rating for each question. Answer 20 questions.
Interim Period	Take Terman's "Concept Mastery Test."	Take Terman's "Concept Mastery Test."
Round 2	Feed back 3 group quartiles for each question. Revise answers to 20 questions.	Feed back individual percentiles on each question. Revise answers to 20 questions.

Note: Terman's "Concept Mastery Test," form T, was administered to all subjects in both experiments reported in this Memorandum. An analysis of the CMT data will be reported in another publication.

A different set of questions was used each day. The feedback for the control groups consisted of a computer printout giving the median and the two quartiles Q_1, Q_3 of the group response in Round 1. The feedback for the experimental groups consisted of a computer printout sheet for each respondent that gave the percentile rating of his Round 1 answer to each question. The percentile rating was defined as the number of respondents giving a lower estimate expressed as a percentage of the number of respondents in the group. The individual giving the lowest answer was thus fed back a zero percentile for that question.

ANALYSIS OF RESULTS

Improvements in Group Estimates

We have used the absolute value of the logarithm of the ratio of group median to true answer as a measure of the accuracy of the group response, i.e., accuracy is measured by $|\ln(\text{Md}/\text{True})|$. Using this measure, we looked at the group response to the 80 questions which were presented to the four control groups (Groups 1-3-5-7) and the four experimental groups (Groups 2-4-6-8). We inspected Round 1 and Round 2 responses on each question and determined whether the accuracy increased, decreased, or remained unchanged in Round 2. The summary of this accuracy count is given in Table 1.

Table 1

CHANGE IN ACCURACY OF GROUP RESPONSES FROM ROUND 1 TO ROUND 2

(80 Questions; 160 Group Responses)

Control Groups				Experimental Groups			
Group No.	Number of Questions Whose Accuracy in Round 2--			Group No.	Number of Questions Whose Accuracy in Round 2--		
	Increased	Decreased	Was Unchanged		Increased	Decreased	Was Unchanged
1	11	6	3	2	6	10	4
3	7	4	9	4	7	2	11
5	6	8	6	6	7	4	9
7	5	2	13	8	8	6	6
Total	29	20	31	Total	28	22	30

It is clear that feeding back individual percentiles to the experimental group did not produce a greater number of improvements. The totals for control and experimental are very much alike.

The next question to be considered is the amount of improvement in each of the groups. If we look at the 29 questions in the control group and the 28 questions in the experimental group that showed improvement in accuracy, we find that the improvement from Round 1 to Round 2 averaged about .21 per question in both cases. It appears that there is no difference in amount of improvement. Seven of the questions in the group of 29 appeared also in the group of 28. The variation in amount of improvement among questions was great--for the control group from .001 on question 3 in Group 1 to 1.07 on question 8 in Group 5. For the experimental group, the improvement on individual questions varied from .025 on question 11 in Group 8 to .92 on question 11 in Group 4.

If we look at all 80 responses in the control and experimental groups and sum the errors over all questions using the absolute value of ln Md/True, we find that net improvement per question is .045 for the control group as opposed to .001 for the experimental group. All control groups (1, 3, 5, and 7) show slight net improvement. Among the experimental groups only Group 4 shows improvement; Groups 2, 6, and 8 all show a net loss. The control and experimental groups seem undistinguishable on this measure of accuracy.

Number of Changes of Estimates on Round 2

The number of changes made by respondents in Round 2 is summarized in Table 2. The control groups made 707 changes out of a possible 1380, or 51%. The experimental group made 746 out of a possible 1360, or 55%. These results do not appear to differ.

A further analysis of responses was made to see if there was any evidence that respondents in the experimental groups who made changes were motivated to do so by the occurrence of an unusually high or unusually low percentile point.

We examined the 1360 responses in Groups 2, 4, 6, and 8 and made a distribution of percentile ratings for the 746 responses in which changes were made and the 614 in which no change was made. The two frequency distributions of percentile ratings are shown in Table 3.

A comparison of the two frequency distributions, one in which the Round 1 response was changed in Round 2 and the other in which no change was made, shows clear differences in the distribution patterns. These patterns are shown in Fig. 2. The changed responses approximate a U-shaped curve with fewer changes for questions in which the percentile feedback was between 40 and 60 and more changes when percentile ratings were either high or low. The curve for the unchanged responses approximates an inverted V curve. The total distribution of responses is approximately uniform, with about 10% of the ratings in each category,

Table 2

NUMBER OF CHANGES OF ESTIMATES IN ROUND 2

Control Groups				Experimental Groups			
Group No.	Number of Responses	Number of Changes	Average Number of Changes per Respondent	Group No.	Number of Responses	Number of Changes	Average Number of Changes per Respondent
1	20	211	10.55	2	18	191	10.61
3	18	166	9.22	4	17	224	13.18
5	16	178	11.12	6	18	149	8.28
7	15	152	10.13	8	15	182	12.13
Total	69	707	10.25	Total	68	746	10.97

Table 3

DISTRIBUTION OF PERCENTILE RATINGS FOR RESPONSE OF EXPERIMENTAL GROUPS

Percentile Rating	Total Responses	No Change Made	Changes Made	
			Number	Percent
0-09	187	43	144	77
10-19	139	48	91	65
20-29	159	68	91	57
30-39	119	58	61	51
40-49	127	93	34	27
50-59	143	110	33	23
60-69	145	86	59	41
70-79	120	44	76	63
80-89	147	43	104	71
90-99	74	21	53	72
Total	1360	614	746	

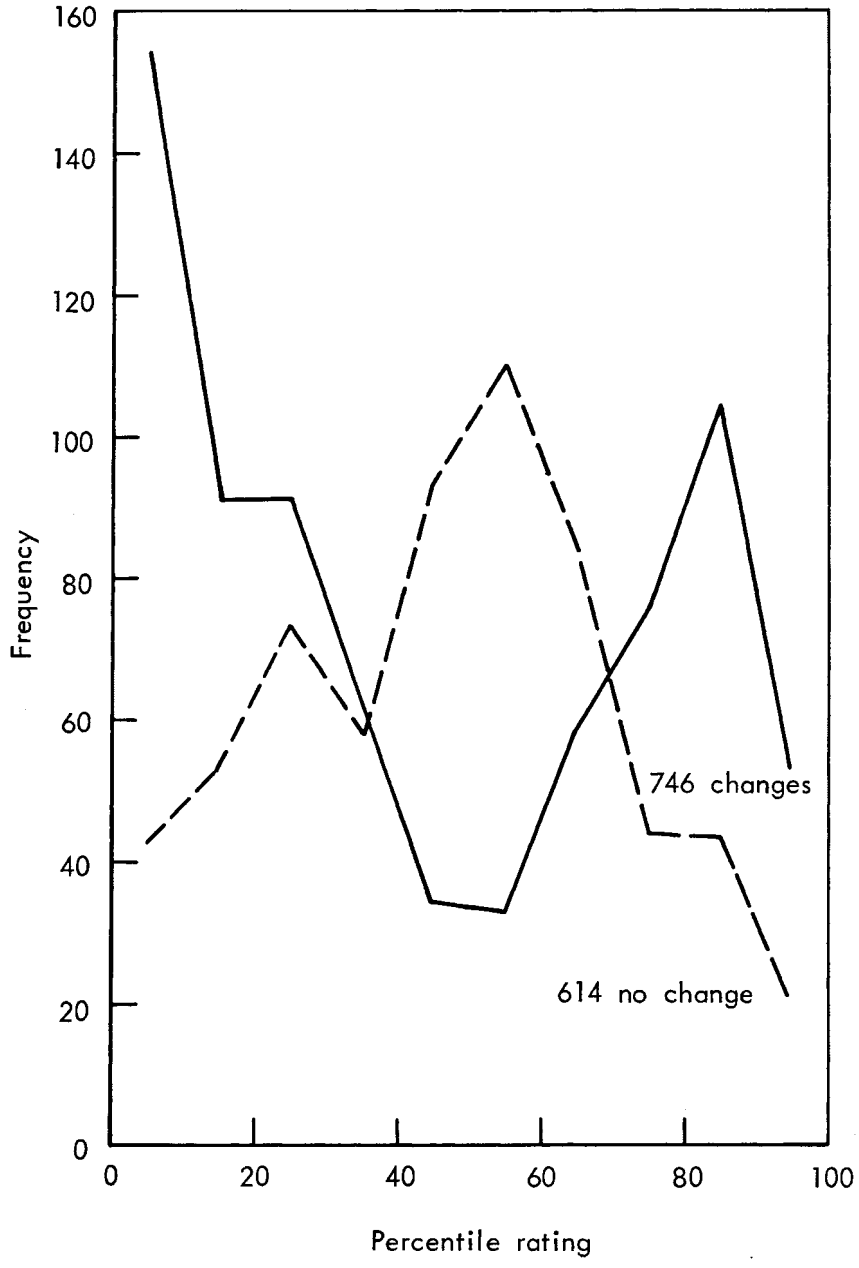

Fig. 2—Distribution of percentile ratings for changed and unchanged response in Round 2

but there was a slight bias in favor of very low ratings (0-9) and against very high (90-99) because of the definition used for percentile ratings. About 14% of the ratings were in the lowest group and only 5% in the highest. There were no ratings above 94 by our arbitrary definition and there were 91 zero ratings in the distribution.

In Fig. 3 we have shown the percent of changed responses in each category as a function of the percentile ratings. A freehand curve was drawn neglecting the two end points. It appears that about 15% of the responses would be changed if the percentile rating is near 50 and about 65% changed if percentiles are less than 20 or more than 80.

The hypothesis that feedback of individual percentiles would produce more improvement than feedback of median and quartiles was not upheld by the data of this experiment. The control group and experimental group show no difference in the number of improvements or in the amount of improvement.

In addition, the control and experimental groups show no difference in the number of changes made in Round 2. The relationship between likelihood of change and percentile rank is very similar to the relationship between likelihood of change and distance from the median previously found for quartile feedback. This relationship is displayed in Fig. 4 taken from [1].

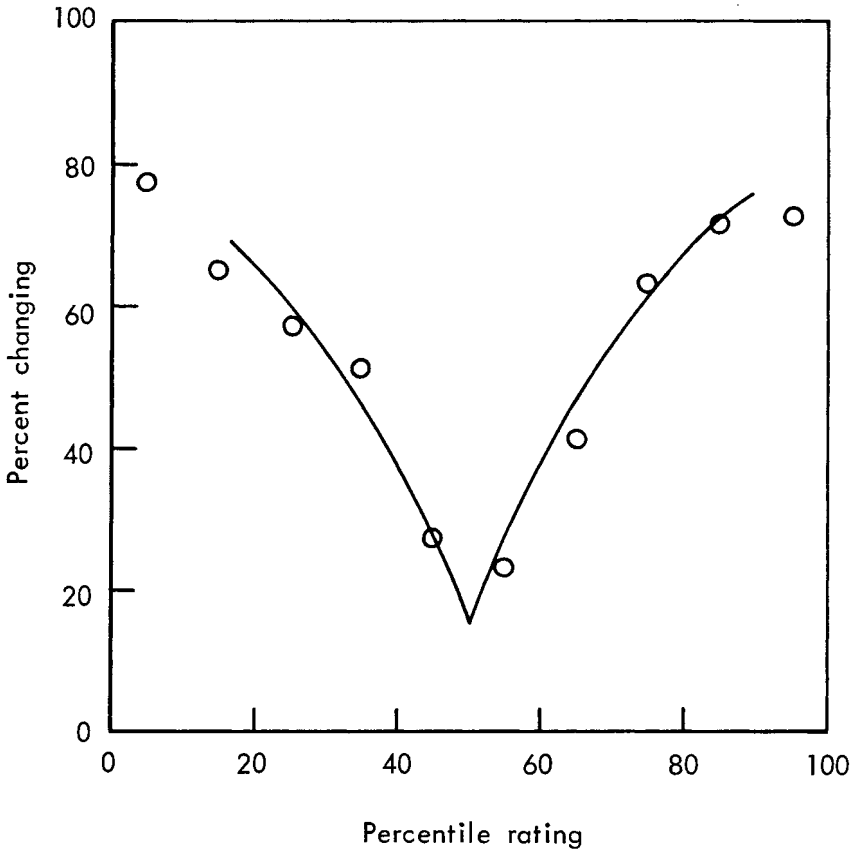

Fig. 3—Likelihood of change in response as a function of percentile rank.

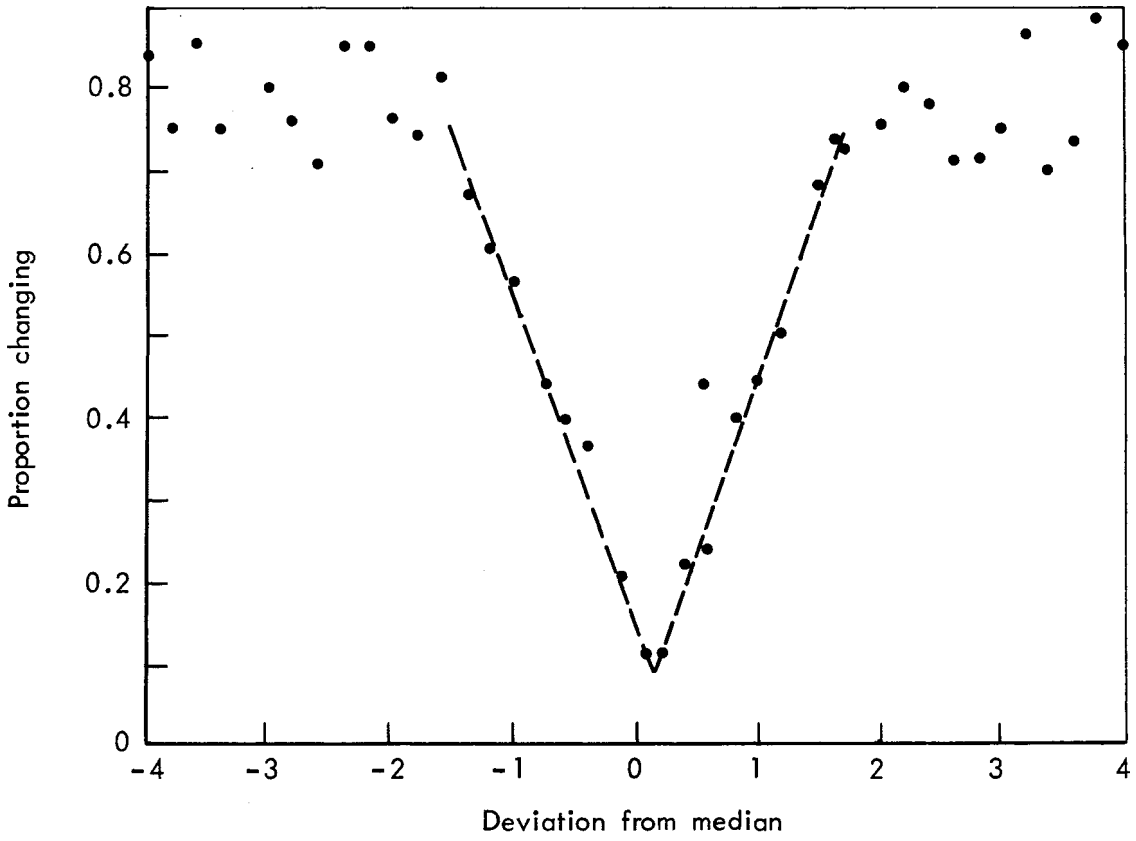

Fig. 4—Likelihood of change in response as a function of distance from median

III. EFFECT OF FEED-IN OF RELEVANT FACT

PURPOSE

The purpose of the experiment was to determine the effect of feeding in a relevant fact in addition to the feedback of quartiles.

METHOD

One hundred forty-five students from the University of California at Los Angeles were paid to serve as subjects in April 1969. Of these 74 were women and 71 men; 30 were graduate students, 115 upper division undergraduates. The subjects were divided at random into eight groups of 17 to 20 per group. The control groups (Nos. 9, 11, 13, and 15) included 37 males and 35 females and 14 of these were graduate students. The experimental groups (Nos. 10, 12, 14, and 16) included 34 males and 39 females, of whom 16 were graduate students.

Each day of the experiment one control group and one experimental group were asked to answer 20 general information type questions. These questions and the relevant fact fed in are listed in the Appendix.

The design of the experiment was follows:

	Control group	Experimental group
Round 1	Give self-rating for each question Answer 20 questions.	Give self-rating for each question Answer 20 questions.
Interim Period	Take Terman's "Concept Mastery Test."	Take Terman's "Concept Mastery Test."
Round 2	Feed back 3 group quartiles for each question Revise answers to 20 questions.	Feed back 3 group quartiles for each question. Feed in a relevant fact for each question Revise answers to 20 questions.

ANALYSIS OF RESULTS

There were 160 group responses to the 80 questions on Round 1 and the same number on Round 2. The accuracy of the group response was measured, as in the previous experiment, by the absolute value of the logarithm of the ratio of group median to true answer.

We looked at the pair of responses for Round 1 and Round 2 on each question and counted the number of cases in which the second round error was smaller than the first (improved accuracy), the number of questions for which Round 2 response was not an improvement, and the number in which Round 1 and 2 were the same. These results are shown in Table 4.

It is clear that there were a greater number of improvements in the responses of the experimental group than in those of the control group. The number of group responses that remained unchanged from Round 1 to Round 2 is much greater for the control group. We assume that the introduction

Table 4

COMPARISON OF ROUND 2 ACCURACY WITH ROUND 1 ACCURACY

Median of Group Response

Control Groups				Experimental Groups			
Group No.	Number of Questions Whose Accuracy in Round 2--			Group No.	Number of Questions Whose Accuracy in Round 2--		
	Improved	Decreased	Was Unchanged		Improved	Decreased	Was Unchanged
9	10	5	5	10	16	2	2
11	7	4	9	12	15	5	0
13	8	5	7	14	14	1	5
15	5	5	10	16	12	5	3
Total	30	19	31	Total	57	13	10

of a relevant fact was a strong motivation for revising the response. If we omit the responses that did not change and look at the ratio of number of improvements to total changed responses, the results show that the control group improved 61% and the experimental group, 81%. Chi-square for one degree of freedom was calculated as 5.98 ($.02 > P > .01$). This evidence would indicate a rejection of the hypothesis that the number of improvements was the same in both groups.

Comparisons in terms of mean error per question are shown in Table 5. The two groups did not differ on Round 1 response but with the second round response, the experimental group shows a smaller mean error which is bordering on statistical significance at the .05 level. The control group showed no change in mean error from Round 1 to Round 2, but the experimental group decreased its mean error in Round 2 significantly. Apparently the feeding in of a relevant fact was an aid to the group in reducing the average error.

Table 5

GROUP COMPARISONS ON MEAN ERROR PER QUESTION

	Control Group	Experimental Group	Difference	Standard Error or Difference	t
Round 1	1.197	1.282	.085	.230	.370
Round 2	1.144	.808	.336	.200	1.680
Difference	.053	.474			
Standard Error Of Difference	.225	.205			
t	.235	2.312			

The facts which were used as information feedback for Round 2 could not be characterized by a single descriptive phrase. (See Appendix.) The relationship of the fact to the original question varied widely from a tenuous hint to a direct help. A few were nonnumerical facts. We classified these 80 relevant facts into 3 groups. These are shown in Table 6.

The following are illustrative examples of the three classes of facts:

1. Upper and Lower Bound Facts

   Question: How many home runs did Babe Ruth hit in 1927?

   Fact (upper bound): Babe Ruth hit a total of 714 home runs.

   Question: What was the total world diamond production in carats in 1965?

   Fact (lower bound): The diamond production in USSR in 1965 was 3.5 million carats.

2. Qualitative Fact

   Question: How many years ago was gunpowder invented?

   Fact: Gunpowder was invented by the Chinese.

3. Other Type Fact (Comparison, Analogy, etc.)

   Question: What was the number of people in the Royal Netherlands Air Force in 1965?

   Fact: The number of people in the Royal Norwegian Air Force in 1965 was 9000.

The comparison shown in Table 6 is between experimental and control group responses on Round 2. The numbers are simply counts of the number of questions in each category for which the experimental group was better or worse or the same as the control group. Note that about 31% of the questions used boundary values as facts and that about 13% were qualitative facts, leaving about 56% of the questions for the "other" classification.

The chi-square test supports the hypothesis that the three categories of relevant facts are not differentiated by the counts shown here.

Table 7 shows the change between Round 1 and Round 2 for both the control groups and the experimental groups in terms of the number of changes which led to increased accuracy, decreased accuracy, and no change.

If we omit the 10 questions for which a qualitative fact was used for feedback and also omit the questions on which Round 1 and Round 2 resposes were the same, we will have a 2 x 2 table for the control group and one for the experimental. The chi-square value* for 1 degree of freedom (d.f.) for the control group is .48 and for the experimental is .09, indicating that the percent of improvements does not differ between the set of questions having boundary values as relevant facts and those that do not.

The average amount of error per question for both groups and both rounds is shown in Table 8. This table shows the quantitative accuracy measures that correspond

Table 6

COMPARISON OF EXPERIMENTAL AND CONTROL GROUP RESPONSES ON ROUND 2
(By Description of Relevant Fact)

Description of Relevant Fact	Number of Questions	Number of Questions in Which Experimental Group was--		
		Better than Control	Worse than Control	Same as Control
Upper and Lower Bound	25	18	6	1
Qualitative	10	4	4	2
Others	45	23	18	4
Total	80	45	28	7

χ^2 for 4 d.f. = 6.0, P ~ .20

Table 7

COMPARISON OF THE NUMBER OF CHANGES BETWEEN ROUND 1 AND ROUND 2
FOR EXPERIMENTAL AND CONTROL GROUPS

Description of Relevant Fact	Number of Questions	Control Groups			Experimental Groups		
		Number of Questions Which in Round 2 were--			Number of Questions Which in Round 2 were--		
		Better	Worse	Same	Better	Worse	Same
Upper and Lower Bound	25	5	10	10	19	3	3
Qualitative	10	6	2	2	2	4	4
Others	45	19	7	19	36	6	3
Total	80	30	19	31	57	13	10

χ^2 for 4 d.f. 8.8, P ~ .09 15.5, P < .01

Table 8

COMPARISON OF CONTROL AND EXPERIMENTAL GROUPS ON ROUND 1 AND ROUND 2 ERRORS, BY RELEVANT FACT CATEGORIES

Description of Relevant Fact	Round 1 Error Mean Error per Question		Round 2 Error Mean Error per Question		Number of Questions
	Control	Experimental	Control	Experimental	
Upper or Lower Bound	1.34	1.48	1.33	.75	25
Qualitative Facts	.72	.62	.53	.58	10
Others	1.23	1.32	1.18	.89	45
Average for all questions	1.20	1.28	1.14	.81	80

t (Experimental group, Round 1 and Round 2) = 2.00

t (Round 2, Upper and Lower Bound questions, Experimental and Control Groups) = 2.00

to the counts in Table 6. The average error for the set of 80 questions is the same as shown in Table 5.

Round 2 average errors in each category except the qualitative facts are less for the experimental group than for the control. The 25 questions in the first category (Upper and Lower Bound) show an average per question change from Round 1 to Round 2 which is almost imperceptible in the control group, but in the experimental group there is an average improvement of .73 per question. This is approximately a 50% decrease in error. The 45 questions in the category designated "other" show an average decrease in error of .05 for the control group and .43 for the experimental. This means the average error decreased about 4% for the control group as against 30% for the experimental. Over the total set of questions, the control group reduced their Round 1 error by about 5% in Round 2 while the experimental group reduced their Round 1 error by 37%.

The respondents in the control group on Round 2 made 920 changes out of a possible 1440 (64%). Respondents in the experimental group made 1119 changes out of a possible 1460 (77%). It may be that the feedback of group quartiles provides a signal to the respondent to revise his estimate and that the addition of a relevant fact is a stronger signal and gives direction to the revision.

In a previous set of experiments significant differences were found between the performance of men and women subjects. Table 9 displays the comparison of male and female performance in both the percentile experiment and the relevant fact experiment. These are averages of the medians of male and female subgroups. This comparison introduces a small distortion for Round 2, since the material fed back was for the total group, and not for male and female subgroups separately.

The table reconfirms results previously obtained: On Round 1 the females are distinctly less accurate than the males. Given feedback of the median and quartiles, the increase in accuracy of the females is significantly greater than that of the males. Nevertheless, the female estimates are still less accurate on Round 2 than are those of the males on Round 1.

The situation is quite different in the relevant fact experiment. Although the females are again less accurate on Round 1, their improvement in Round 2 is much greater than the improvement of the males. Furthermore, their average error in Round 2 is significantly lower than the average error of the males in Round 1 and is not significantly greater than the average of the males in Round 2.

In short, although the females are less accurate on their initial estimates, they are able to make as good (or better?) use of the relevant fact to improve their estimates.

Table 9

GROUP ERROR

(Average Over 80 Questions)

Type of Experiment	Group	Round 1			Round 2		
		Males	Females	Total	Males	Females	Total
Percentile	Control	.72	1.30	.85	.69	.95	.80
	Experimental	.87	1.18	1.00	.84	1.12	1.00
Relevant Fact	Control	1.05	1.51	1.20	1.04	1.21	1.14
	Experimental	1.09	1.51	.85	.78	.85	.81

IV. DISCUSSION

The results of the experiment with percentile feedback would appear to indicate that the Delphi process is not very sensitive to the form of feedback as long as it involves some relatively precise summary of the group response on the previous round. The percentile feedback appears to be slightly less effective than medians and quartiles with respect to numerical improvement (average error), but neither form of feedback is very effective on this measure.

The results look entirely different for the experiment involving feed-in of a relevant fact. Both the number of questions showing improved accuracy on Round 2, and the amount of numerical improvement, are decisively greater than for statistical feedback alone. In fact, this is the first experiment in the series in which the numerical improvement in accuracy has reached statistical significance.

The ability of the subjects to use essentially any fact, whatever the nature of its relevance to the question involved, is suggestive of the great flexibility of the human mind in dealing with fragmentary information. Also interesting is the fact that the women were able to use the factual information to improve their answers to the point where they were essentially as accurate as the men. In all previous exercises in which statistical feedback alone was used, the responses of the female subjects on Round 2 have not been as accurate as those of the male subjects on Round 1.

The Relevant Fact results do not have a simple implication for applications. For most practical problems, the volume of potentially relevant material is enormous; an attempt to make this material available to all respondents would undoubtedly lead to excessive communications--both with respect to the task of the exercise managers and with respect to the ability of the respondents to absorb the material. There is some hope that before too long on-line computer networks, with access to large data stores, may lead to feasible handling of the communications problem from the standpoint of the exercise manager. It is not clear that on-line interaction with large data stores will solve the problem of absorption by the respondent.

There are two issues here. One is the issue of saturation. One, or a few, relevant facts may lead to improvement, but more than some small number may lead to degradation of estimates. This issue can be studied experimentally, and will be examined in future experiments.

The second issue is more one of study design for applications. Undoubtedly, one of the strongest appeals of the Delphi procedures is that it tends to rely on the information already "available" to the respondents in terms of their present expertise, plus whatever material they have readily at hand. Interaction of the panel with large exogenous sources of information is a different kind of exercise, with different criteria of excellence. Although

the success of the Relevant Fact experiment suggests the payoff from such an extended exercise may well be worth the additional costs in time and respondent effort, it should be pointed out that the problem of design and the measure of its effectiveness will be difficult.

Appendix

QUESTIONS AND RELEVANT FACTS

1. What was the total world diamond production in carats in 1965?
   Fact: The diamond production in U.S.S.R. in 1965 was 3,500,000 carats.

2. How many years ago was gunpowder invented?
   Fact: Gunpowder was invented by the Chinese.

3. How tall in inches was the winner of the Miss America Pageant in 1945?
   Fact: Miss America of 1945 weighed 135 pounds.

4. How many people living in the U. S. in 1960 were born in Norway?
   Fact: 15,723 people born in Norway were living in California in 1960.

5. How many home runs did Babe Ruth hit in 1927?
   Fact: Babe Ruth hit a total of 714 home runs.

6. The 1968–1969 edition of Who's Who in America has a total of how many names?
   Fact: The number of names in the 1968–69 edition of Who's Who in America is eight times the number of names appearing in the first edition in 1899.

7. What is the greatest number of people ever employed at one time in the U.S.?
   Fact: The number of unemployed people in October 1967 in the U.S. was 3,367,000.

8. How many U.S. and USSR manned orbital flights were accomplished through April 1967?
   Fact: The first manned orbital flight was in April 1961 (USSR).

9. In 1967, in how many of the 50 states was the annual salary of the governor less than 25,000 dollars?
   Fact: The annual salary of the governor of California is 44,100 dollars.

10. What was the number of people in the Royal Netherlands Air Force in 1965?
    Fact: The number of people in the Royal Norwegian Air Force in 1965 was 9000.

11. What is the estimated number of Frenchmen who die each year of acute chronic alcoholism?
    Fact: During 1965 France produced 18,080,000,000 gallons of wine.

12. What was world production of soybeans in tons in 1965?
    Fact: Japan produced 253,000 tons of soybeans in 1965.

13. How many women earned Master's Degrees in American Universities and Colleges in 1928?
    Fact: 47,588 women earned Master's degrees in American Universities and Colleges in 1966.

14. What was the total U. S. foreign economic aid commitments in 1967?
    Fact: The Food for Freedom program totalled 1,040,000,000 dollars in 1967.

15. How many people were on the teaching staff at Ohio State University, Columbus in school year 1966–67.
    Fact: 1392 people were on the teaching staff at UCLA during the school year 1966–67.

16. How many people in the United States died in 1966?
    Fact: The number of registered live births in the U. S. in 1966 was 3,629,000.

17. What was the U.S. total defense expenditures in 1967?
    Fact: The number of people on active duty in Army, Navy, and Air Force in 1967 was 3,090,000.

18. In how many of the 50 states do toll turnpikes or toll throughways exist?
    Fact: The total mileage of toll turnpikes and throughways is 3419 miles.

19. In the U. S. during 1963 how many people were employed in meat slaughtering plants?
    Fact: 8,726 of the employees in Meat Slaughtering Plants in 1963 were classified as nonproduction workers.

20. What was the total membership in the AFL–CIO in 1967?
    Fact: The civilian labor force (persons 16 years of age and over) in the U. S. in 1967 was 78,402,000.

21. What is the highest price ever paid for a painting by a living artist?
    Fact: The picture was "Mother and Child", by Picasso, painted in 1902 and auctioned at Sotheby's in 1967.

22. In 1967, what was the total number of U.S. travelers to the West Indies and Central America?
    Fact: In 1967 the total number of travelers to South America was 175,000.

23. How many one family houses were started in the U.S. in 1966?
    Fact: Total private and public new housing units started in the U.S. in 1966 was 1,251,900.

24. How many artificial satellites have been successfully orbited through 1968?
    Fact: There were approximately 125 artificial satellites still orbiting at the end of 1968.

25. In 1960 how many people in the United States were living in towns of from 1,000 to 2,500 population?
    Fact: The number of people in the U.S. living in towns of 2500 to 10,000 was 13,274,424 in 1960.

26. What is the height (in feet) of the tallest TV tower in the world?
    Fact: The height of the Eiffel Tower including the TV antenna is 1056 feet.

27. How many automobiles were registered in California in 1966?
    Fact: The number of licensed car, truck and bus drivers in California in 1966 was 10,356,000.

28. How many times has the Western football team won the Rose Bowl game in the period from 1920 through 1968?
    Fact: There were 3 ties in the Rose Bowl game in the period from 1920–1968.

29. What was the number of school buses in North Carolina in 1966?
    Fact: There were 1,183,690 pupils enrolled in public elementary and secondary schools in North Carolina in 1966.

30. What is the estimated number of new cases of cancer of the mouth or lips that develop each year?
    Fact: About 7000 people die in the U.S. each year from cancer of the mouth and lips.

31. How many dams 220 feet or more in height were in the U.S. in 1966?
    Fact: There were 29 dams 220 feet or more in height in California in 1966.

32. In the 1968–69 edition of Who's Who In America how many names appear for the first time?
    Fact: The 1968–69 edition of Who's Who In America has a total of 66,000 names.

33. What was the total number of state & local government employees in the U.S. in 1966?
    Fact: The number of civilian employees in Federal government in 1966 was 2,612,000.

34. How many whooping cranes were there in North America in 1965?
    Fact: There were fourteen whooping cranes in North America in 1939.

35. What is the wingspan in feet of the Boeing 707?
    Fact: The maximum speed of the Boeing 707 is 600 mph.

36. How many years have elapsed since the first MD degree was conferred on a negro in the United States?
    Fact: During 1960, there were 215 negro physicians practicing medicine in Chicago.

37. In 1966 how many earned doctorates were awarded at UCLA?
    Fact: In 1966, 18,239 earned doctorates were awarded in the U.S.

38. How many deaths as a result of riots occurred in the U.S. from 1964 through 1967?
    Fact: There were about 20,000 arrests made as a result of riots occurring in the U.S. from 1964 through 1967.

39. In 1928 how many students were enrolled in U.S. colleges?
    Fact: In 1967, 6,963,687 students were enrolled in U.S. colleges.

40. In 1966, Americans consumed an average of how many pounds of beef per person?
    Fact: Total sales in the meat industry in the U.S. in 1966 amounted to 18 billion dollars.

41. How many law students were there in state universities in France in 1964?
    Fact: There were 42,114 medical students in state universities in France in 1964.

42. What was the median years of school completed by the population of U.S. non-white males, age 75 years and over in 1950?
    Fact: The median years of school completed for all U.S. males, age 25 years and over in 1950 was 9.0.

43. What is the maximum number in any one year of strikes and lockouts in the U.S. since 1900?
    Fact: The minimum number of strikes and lockouts in any one year since 1900 in the U.S. was 841 in 1932.

44. How many bowling teams were there in the U.S. in 1955?
    Fact: There were 58,203 bowling lanes in the U.S. in 1955.

45. In California, during 1966, how many females held valid drivers licenses?
    Fact: The total number of licensed drivers in Illinois in 1966 was 5,820,735.

46. What is the world's record rainfall in inches for any 24 hour period?
    Fact: The world's record rainfall for any 24 hour period occurred in the Indian Ocean in 1952.

47. How many dollars did U.S. Consumers spend on radio and T.V. repairs in 1957?
    Fact: The amount spent by U.S. consumers on radio and T.V. repairs in 1947 was 140,000,000 dollars.

48. How many physicians were there in Israel in 1965?
    Fact: The population of Israel in 1965 was 2,525,600.

49. How many $2.00 bills were in circulation in 1966?
    Fact: There were 551,000,000 five dollar bills in circulation in 1966.

50. What was the U.S. dollar value of the English pound in 1943?
    Fact: The current U.S. dollar value of the English pound is $2.45.

51. How many Counties are there in the state of Texas?
    Fact: There are 99 counties in the state of Iowa.

52. How many were employed in the manufacture of tobacco products in the United States during 1966?
    Fact: The total payroll in the manufacture of tobacco products in the U.S. in 1966 was 356,000,000 dollars.

53. How many children 7 to 13 years old were enrolled in school in October 1965?
    Fact: Approximately one-fourth of the population of the U.S. is in the age group 5-17 years.

54. What was the population of California in 1890?
    Fact: The population of California in 1960 was 15,717,204.

55. What was the number of passenger cars in the United States using highways during 1967?
    Fact: There were 16,500,000 trucks and buses using U.S. highways during 1967.

56. What was the U.S. production in tons of copper ore in 1965?
    Fact: Canada produced 516,120 tons of copper ore in 1965.

57. How long has Harvard been in existence?
    Fact: Yale University was established in 1701.

58. During 1966 how many square yards of cotton material were produced in the United States?
    Fact: During 1966, U.S. exported 336,632,000 square yards of cotton material.

59. How many accredited U.S. Universities and Colleges were there in 1967?
    Fact: There were 87 accredited universities and colleges in California in 1967.

60. What was the average number of deaths per day from motor vehicle accidents in the U.S. in 1966?
    Fact: The average number of deaths per day from all accidents in the U.S. was 288 in 1966.

61. How many lawyers in private practice were there in South Carolina in 1966?
    Fact: In the U.S. there were 212,662 lawyers in private practice during 1966.

62. What was the U.S. ski jump record in 1920?
    Fact: The U.S. ski jump record in 1967 was 335 feet.

63. How many women have been elected to the Hall of Fame for Great Americans?
    Fact: Ninety-three people have been elected to the Hall of Fame for Great Americans.

64. During 1966 how many pairs of adult men's shoes were manufactured in the U.S.?
    Fact: The total production of shoes in the U.S. in 1966 was 646,327,000 pairs.

65. What was the amount of the U.S. Public Debt in 1966?
    Fact: The Gross National Product for 1966 was 743,288,000,000 dollars.

66. How many selective service draftees examined during 1966 were disqualified because of failure to meet mental requirements?
    Fact: During 1966, 1,609,000 selective service draftees were examined.

67. What is the population of the Republic of the Philippines?
    Fact: The population density of the Republic of the Philippines is 289 per sq. mile.

68. What was the estimated property damage from Hurricane "Betsy" in 1965?
    Fact: The number of deaths in the U.S. from Hurricane "Betsy" was 75.

69. How many of the automobiles registered in the District of Columbia were owned by Civilian branches of the Federal Government?
    Fact: In Washington D.C. in 1966, there were 290,200 Federal civilian employees.

70. What is the height (in feet) of the tallest Cathedral spires in the world?
    Fact: The Cathedral with the tallest spires in the world is in Germany.

71. How many deaths in the U.S. in 1966 resulted from motor vehicle non-collision accidents?
Fact: The total number of deaths from motor vehicle accidents in 1966 was 53,000.

72. How many pages does Webster's Third New International Dictionary (unabridged) contain?
Fact: Wesbster's Third New International Dictionary (unabridged) has a vocabulary of over 450,000 words.

73. In the period from 1932 through 1967 how many times have the Green Bay Packers been the Western Conference football winners?
Fact: The Los Angeles Rams have been Western Conference football winners 4 times in the period from 1932-1967.

74. What was the total steel consumption of Japan, in tons, in 1964?
Fact: The steel consumption of the United States, in tons, during 1964 was 129,874,000.

75. In the U.S. in 1966 how many male children were under five years of age?
Fact: The population of the U.S. in 1966 was 196,842,000.

76. What was the cheese production of France in 1965 in tons?
Fact: The cows, goats and sheep of France produced 30,500,000 tons of milk in 1965.

77. How many submarines did the U.S. military have as of June 30, 1967?
Fact: As of June 30, 1967 the U.S. Military had 662 destroyers, frigates and their escorts.

78. What was the total number of votes cast for the presidential candidates in the 1964 election?
Fact: The population of voting age in the U.S. in 1964 was 113,931,000.

79. How many years has the Bureau of the Budget been established?
Fact: The Council of Economic Advisors was established in 1946.

80. What is the record price paid for an impressionist painting?
Fact: The painting is by Monet - La Tervasse a' Sainte-Adresse.

REFERENCE

1. Dalkey, N., The Delphi Method: An Experimental Study of Group Opinion, The RAND Corporation, RM-5888-PR, June 1969.