IMPROVING THE RELIABILITY OF ESTIMATES
OBTAINED FROM A CONSENSUS OF EXPERTS

Bernice Brown
Olaf Helmer

September 1964
IMPROVING THE RELIABILITY OF ESTIMATES
OBTAINED FROM A CONSENSUS OF EXPERTS

Bernice Brown*
Olaf Helmer

The RAND Corporation, Santa Monica, California

Introduction

This is a report on the outcome of an experiment in the use of expert opinions. The experiment involved both the so-called Delphi technique, which seeks to induce opinion convergence through a sequence of questionnaires interspersed with controlled opinion feedback, and the computation of a consensus based on self-appraised competence ratings.

By way of a preface to our report, we would like to point out why we think that studies directed toward the improved utilization of expertise are important for operations research. The objective of operations research is not so much to find things out—as the pure scientist tries to do—but to help arrive at efficient operating decisions. This pragmatic attitude implies that the operations analyst, in dealing with phenomena for which no well-established scientific theory is currently available, must nevertheless construct a model as best he can, using whatever intuitive insight limited practical experience may have yielded in

*Any views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.
order to choose the appropriate structure for the model and
to estimate appropriate values for the input parameters.
Consequently, the quality of the output of such an analysis
depends very definitely on the intuitive wisdom that went
into its design and into its underlying numerical assumptions,
and therefore on expert opinion. Thus, the decision-making
process, even if it does not rely merely on informally
expressed expert opinion—as it often does—but is based on
the results of an operations—analytical model, still cannot
fail to depend indirectly on the expert judgment that went
into the construction of the model. It is for this reason
that it seems to us to be the proper concern of operations
research to seek methods of improving the quality of such
statements as are based on expert opinion.

One important step in this direction is the pursuance
of consensus research. This is the study of techniques to
determine the most reliable consensus if several expert
opinions are offered for a particular estimate. Our recent
experiment, on which we are here reporting, was a contribu-
tion to this effort.

Suppose a number of experts are asked independently to
estimate some quantity. Their estimates will cover a certain
range. A convenient way to describe this pattern of opinions
is in terms of the median (the most convenient single number
to be used as representative of the group's collective opinion)
and the interquartile range* (as a measure of the divergence
of opinion among the experts). Our experiment was concerned
with three desiderata: (i) to cause convergence of opinions
in the sense of shrinking the opinion spread as expressed
by the interquartile range; (ii) to cause convergence in
the sense of more closely approximating the true value by
the median; and (iii) to find a formula for determining a
consensus that would be a more reliable estimator of the
true value than the group median.

*The interquartile range is the interval containing
the middle 50% of the responses.
Description of the Experiment

In the experiment we submitted 20 questions, all having numerical answers, to 23 respondents drawn from the RAND research staff. The questions covered a variety of subject matters. Eighteen of them were of a kind for which the answers can be found in the World Almanac; the remaining two were mathematical questions whose answers could be computed but only with some effort. The subjects were told to give their answers without using any reference material and without spending more than a few minutes on each question.

The list of questions used was as follows:

1. How many randomly selected persons must there be in a group so that the probability is $\frac{1}{2}$ that at least three of them have their birthday on the same day of the year?

2. How many million board feet of lumber were produced in the United States in 1962?

3. What is the surface of the Moon in thousand square miles?

4. What was the average price received by the United States farmer for a bushel of apples in 1940?

5. What is the distance in geographical miles between Capetown and the geographical point antipodal to Los Angeles?

6. What is the area in square miles of Los Angeles County?

7. What was the total payroll, in million dollars, of employees in the automobile industry in the United States in 1962?

8. What was the total tonnage, in millions, shipped through the Port of New York in 1962?

9. How many divorces, in thousands, were there in the United States in 1960?

10. How many seconds after launch does a Minuteman (the model now in military use) reach an altitude of 100,000 feet?

11. What is the specified operational gross weight, in pounds, of the Gemini capsule (exclusive of occupants)?
12. How many prime numbers are there between 1000 and 2000?

13. What was the year of birth of the inventor of FM-radio?

14. How many dozens of ordinary lead or graphite pencils did RAND (Santa Monica) buy in 1963?

15. What was the dollar value, in millions, of United States exports to Canada in 1962?

16. What is the mean distance, in million miles, between Earth and Mars?

17. What is the basic fly-away price of a DC-9, in thousand dollars (including seats, radios, and galleys)?

18. How many popular votes were cast for Lincoln when he first ran for the presidency?

19. How many accidental deaths, in thousands, were there in the United States in 1960?

20. In the alphabetical listing of the current Santa Monica telephone directory the mid-name is "Lancaster"; the Manhattan telephone directory's alphabetical listing begins on p. 17 and ends on p. 1803; what is the number of the page on which the name "Lancaster" appears?

In the first-round questionnaire, each respondent was asked to indicate, along with the estimated answer to each question, a self-rating, based on a scale of 1, 2, 3, or 4, giving an evaluation of his own degree of expertise on each question.

In a second questionnaire, each respondent was informed of the group median of the first-round responses and of their interquartile range, as well as of the frequency distribution of the self-ratings. The participants were instructed to reconsider each answer, make a revised estimate and, if the new answer lay outside the indicated interquartile range, briefly state a reason for this opinion.

A third-round questionnaire summarized the reasons thus elicited that had influenced respondents' opinions, and the subjects were invited to criticize these reasons if they found them unconvincing. The medians and quartiles of the
second-round responses were also included in the feedback information, and the respondents were asked again to give, possibly revised, answers to the 20 questions.

In the final fourth round, in addition again to medians and quartiles, the respondents were given a summary of some majority and minority responses as obtained from the critique of Round 3. They now had one last opportunity to revise their 20 answers.

**Medians and Quartiles of Responses**

The medians and interquartile ranges of the responses, together with the true answers are listed in Table 1.

A graphic representation of the same information is given in Fig. 1. Since we have very different units of measurement, it was convenient to introduce a standard yardstick of some kind. Millions of board feet, dozens of pencils purchased, and the year of birth of an inventor are not measurements one would choose for simple comparisons. Therefore we used the first-round results to standardize the measurements. We took \( \max \{ |Q_1-T|, |Q_3-T| \} \) as a unit of measure for each question, where \( T \) is the true value and \( Q_1 \) and \( Q_3 \) are the first and third quartiles of the first-round responses, and expressed the deviation from the true answer of medians and quartiles in Rounds 2, 3 and 4 in terms of this unit. (Note that this unit is the deviation from \( T \) of the worst response within the interquartile range, and thus represents some intuitive measure of how difficult the respondents regarded each question when first confronted with it.)

**Analysis of Data**

In analyzing the results of the experiment it is well to bear in mind that we had hardly any real experts regarding any of the questions among our respondents. This is perhaps best evidenced by the fact that, in the first round, among all 460 responses only 21% were even in the right ballpark,
<table>
<thead>
<tr>
<th>Question No.</th>
<th>Median and Interquartile Range</th>
<th>Median and Interquartile Range</th>
<th>Median and Interquartile Range</th>
<th>Median and Interquartile Range</th>
<th>&quot;True&quot; Answer [and unit of measure]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>122 (50–549)</td>
<td>122 (80–547)</td>
<td>122 (80–530)</td>
<td>122 (90–400)</td>
<td>87 [number]</td>
</tr>
<tr>
<td>2</td>
<td>2000</td>
<td>2000</td>
<td>2500</td>
<td>3500</td>
<td>33174 [mill. board ft]</td>
</tr>
<tr>
<td>3</td>
<td>5000–5000</td>
<td>1000–5000</td>
<td>1000–5000</td>
<td>2000–7200</td>
<td>14657 [thousand sq mi]</td>
</tr>
<tr>
<td>4</td>
<td>12566</td>
<td>12000</td>
<td>12560</td>
<td>13500</td>
<td>56 [cents]</td>
</tr>
<tr>
<td>5</td>
<td>4300–50200</td>
<td>6000–15000</td>
<td>12000–14700</td>
<td>12000–14660</td>
<td>1950 [miles]</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>4060 [sq miles]</td>
</tr>
<tr>
<td>7</td>
<td>50–180</td>
<td>60–150</td>
<td>60–150</td>
<td>60–140</td>
<td>5828 [million dollars]</td>
</tr>
<tr>
<td>8</td>
<td>2500</td>
<td>3000</td>
<td>2500</td>
<td>2500</td>
<td>145 [million tons]</td>
</tr>
<tr>
<td>10</td>
<td>1600</td>
<td>1800</td>
<td>2000</td>
<td>2000</td>
<td>90 [seconds]</td>
</tr>
<tr>
<td>11</td>
<td>500–4000</td>
<td>1200–4000</td>
<td>1500–3000</td>
<td>1500–2500</td>
<td>5877 [pounds]</td>
</tr>
<tr>
<td>12</td>
<td>4000</td>
<td>4000</td>
<td>5000</td>
<td>5000</td>
<td>135 [number]</td>
</tr>
<tr>
<td>14</td>
<td>100–1000</td>
<td>100–750</td>
<td>350–750</td>
<td>400–500</td>
<td>1841 [dozens]</td>
</tr>
<tr>
<td>15</td>
<td>40</td>
<td>40</td>
<td>35</td>
<td>35</td>
<td>3830 [million dollars]</td>
</tr>
<tr>
<td>16</td>
<td>16–200</td>
<td>20–42</td>
<td>20–42</td>
<td>20–42</td>
<td>142 [million miles]</td>
</tr>
<tr>
<td>17</td>
<td>100</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>3067 [thous. dollars]</td>
</tr>
<tr>
<td>18</td>
<td>3000–10000</td>
<td>5000–8000</td>
<td>5000–7000</td>
<td>5000–6500</td>
<td>1866 [thousands]</td>
</tr>
<tr>
<td>19</td>
<td>3000–10000</td>
<td>5000–8000</td>
<td>5000–7000</td>
<td>5000–6500</td>
<td>94 [thousands]</td>
</tr>
<tr>
<td>20</td>
<td>12–80</td>
<td>20–60</td>
<td>22–55</td>
<td>25–40</td>
<td>916 [page number]</td>
</tr>
</tbody>
</table>

*Rough approximation based on unclassified data.*
Fig. 1—Deviation of medians and quartiles from the true answers
if we arbitrarily define "ballpark" as anything within 25% of the true value*. On the other hand, in assessing the possibilities of applying the method of this experiment to real-life situations, that is, to cases where decisions depend on the opinions of expert advisors, their estimates may often be just as unlikely to fall in the right ballpark. The conditions of the experiment therefore may have simulated reality in this respect without too much distortion, in view of which some of the results of it are all the more noteworthy.

We begin by discussing the question of convergence of the estimates.

First of all, the spread of opinions shrank considerably from round to round, which was to be expected, especially as we had virtually forced convergence between Rounds 1 and 2 by demanding a statement of a reason for any "extreme" estimate in the second round, that is, one that remained outside the interquartile range of responses given in Round 1. If, indeed, we measure the opinion spread by the width of the interquartile range, we observed a shrinkage in 19 out of the 20 cases. The median amount by which the range shrank between Rounds 1 and 2 was about one third of the width, and again about one third in the two steps between Rounds 2 and 4 combined.

Next we had to ask ourselves, of course, whether the observed shrinkage in opinion spread represented a convergence in the right direction, namely toward the true values. In one sense it did not. While the first-round interquartile ranges were large enough so that 13 out of the 20 contained the true value, in Rounds 2 and 3 only 10 still contained it, and in Round 4 the number of interquartile ranges containing the true answer had decreased to 7 out of the total of 20. While this looked discouraging and indicated that the experimental procedure had induced perhaps

*Except for Questions 13 and 20, where there is no natural zero value; here we stipulated that the "ballpark" consists of the interval \(<T-10,T+10>\).
more convergence than was desirable, it turned out nevertheless that the opinion pattern as a whole tended to move in the right direction, if we describe this motion in terms of its central, or median, element. Comparing, in particular, the median of the first-round responses with that of the fourth round, we found that in $13\frac{1}{2}$ out of 20 cases the latter was closer to the true answer than the former:

$$M_4 \text{ vs } M_1: \ 13\frac{1}{2} / 6\frac{1}{2} = 2.1.$$  

Another way of describing the same phenomenon is in terms of what we chose to call the ballpark: While of the first-round medians only 6 were in the right ballpark, by the time we had progressed to the final round 9 medians were in the right ballpark. (Incidentally, while initially only 21% of all 460 responses were in the right ballpark, in the terminal round this figure rose to 38%.)

**Analysis of Data: Consensus**

Let us turn now to the question of how good a group consensus the median represents. In representing the group's collective opinion by a single number, the median has the evident advantage over, say, the mean of being independent of the metric. Moreover it has the obvious and appealing quality that it is that value for which half the group thinks that the true answer is less than or equal to it and the other half that it is greater than or equal to it. But how about the quality of the median, not as a representative group opinion, but as a reliable estimator of the true value?

Comparing the performance of the final-round medians with that of the individual respondents in the final round, we obtained the following scores:

$$M_4 \text{ vs } R_i: \ 7\frac{1}{2} 8 8 8 9 9\frac{1}{2} 9\frac{1}{2} 10 11\frac{1}{2} 12\frac{1}{2} 12\frac{1}{2} 12\frac{1}{2} 13 13 \ 13\frac{1}{2} 13\frac{1}{2} 14\frac{1}{2} 15 15\frac{1}{2} 15\frac{1}{2} 16 17 17$$

where each figure represents the number of questions (out of a total of 20) in which the median was closer to the true
answer than the individual respondent. Taking the median of these scores (circled figure), we note that the group median scored \(12\frac{1}{2}/7\frac{1}{2}\) against a randomly chosen respondent. This is good but far from spectacular. We also note that 7 of the 23 respondents in the final round actually did better than the median, and if only we had known their identity a priori we could have obtained better estimates than through the median.

But this observation is a clue as to how to go about improving our estimates. In the absence of objective information about the competence of the respondents, we have their own self-appraisals to go by which we had obtained in Round 1 and which thus far had not been utilized. These self-ratings had been solicited by an instruction worded as follows:

"Write one of the numbers 1, 2, 3, or 4 indicating how relatively confident you feel about your answer, using 1 for most confident. More specifically, imagine that the answers from all respondents (chosen from various departments of the RAND research staff) are ranked according to their distance from the true answer; then your number should indicate whether you think your answer falls in the first, second, third, or fourth quartile in this ranking."

Using the subjects' self-ratings, we selected for each question a subgroup of about one third of the participants, namely those who had given themselves the relatively highest rating. This group—a different subgroup for each question—we called E, which represents in a sense the (self-appointed) elite among the experts.

Using just the fourth-round responses of the elite, and comparing its median with that of the entire group, we observed a considerable improvement:

\[
E_4 \text{ vs } M_4: \quad 15\frac{1}{2} / 4\frac{1}{2} = 3.4.
\]

(We mention parenthetically that there was also a significant absolute improvement in the ballpark sense: While \(M_4\) had landed only 9 times out of 20 within 25% of the true values, \(E_4\) did so 14 times.)
Table 2 gives a round-by-round comparison of the medians $M_r$ of the entire group and $E_r$ of the elite subgroup, where $r$ indicates the number of the round.

Table 2

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_2$ vs $M_1$</td>
<td>$9/11 = 0.8$</td>
</tr>
<tr>
<td>$M_3$ vs $M_2$</td>
<td>$13\frac{1}{2}/6\frac{1}{2} = 2.1$</td>
</tr>
<tr>
<td>$M_4$ vs $M_3$</td>
<td>$13/7 = 1.9$</td>
</tr>
<tr>
<td>$M_4$ vs $M_1$</td>
<td>$13\frac{1}{2}/6\frac{1}{2} = 2.1$</td>
</tr>
<tr>
<td>$E_2$ vs $E_1$</td>
<td>$11/9 = 1.2$</td>
</tr>
<tr>
<td>$E_3$ vs $E_2$</td>
<td>$13/7 = 1.9$</td>
</tr>
<tr>
<td>$E_4$ vs $E_3$</td>
<td>$12/8 = 1.5$</td>
</tr>
<tr>
<td>$E_4$ vs $E_1$</td>
<td>$13\frac{1}{2}/6\frac{1}{2} = 2.1$</td>
</tr>
<tr>
<td>$E_1$ vs $M_1$</td>
<td>$14/6 = 2.3$</td>
</tr>
<tr>
<td>$E_2$ vs $M_2$</td>
<td>$14\frac{1}{2}/5\frac{1}{2} = 2.6$</td>
</tr>
<tr>
<td>$E_3$ vs $M_3$</td>
<td>$13\frac{1}{2}/6\frac{1}{2} = 2.1$</td>
</tr>
<tr>
<td>$E_4$ vs $M_4$</td>
<td>$15\frac{1}{2}/4\frac{1}{2} = 3.4$</td>
</tr>
<tr>
<td>$E_4$ vs $M_1$</td>
<td>$16/4 = 4.0$</td>
</tr>
</tbody>
</table>

Comparing the performance in the final round of the elite median with that of the individual subjects, we arrived at the following scores:

$$E_4 \text{ vs } R_i: \ 12 \ 12 \ 12\frac{1}{2} \ 12\frac{1}{2} \ 12\frac{1}{2} \ 14 \ 14 \ 14\frac{1}{2} \ 15 \ 15 \ 15 \ \frac{15}{2} \ 15\frac{1}{2} \ 16 \ 16 \ 16 \ \frac{1}{2} \ 17 \ \frac{1}{2} \ 17 \ \frac{1}{2} \ 17 \ \frac{1}{2} \ 17 \ \frac{1}{2} \ 18$$

Thus, in the median, the elite scored $15\frac{1}{2}/4\frac{1}{2}$ against a randomly chosen respondent, and moreover the elite is seen to have produced an estimator which performs 50% better,
namely 12/8, than the very best individual participant.*

We compared, in retrospect, the subjective self-appraisals with actual objective performance, describing the latter too by one of the numbers 1, 2, 3, or 4, according as the distance of the fourth-round response from the true answer lay in the 1st, 2d, 3d, or 4th quartile of all responses. The rank correlation, for each question, turned out to be positive, with values ranging from .02 to .46. The median rank correlation was .30.

Summary

Summarizing, it may be said that (i) the subjects were relatively inexpert with regard to the questions posed; (ii) convergence of opinions was quite noticeable—the spread as defined by the interquartile range shrank to less than one half during the experiment—but it may have been induced to an undesirably large extent by the experimental procedure; (iii) convergence of the medians to the true values occurred in the majority of cases; and (iv) the use of self-appraised competence ratings in forming a consensus appeared to be a powerful tool for increasing the reliability of the group estimates.

*As an afterthought, since $E_4$ in the majority of cases is closer to the true value than $M_4$, we constructed the following estimator $E'_4$, which represents a further improvement over $E_4$: Starting from $M_4$, proceed in the direction of $E_4$, and go either twice the distance from $M_4$ to $E_4$, or to the next quartile, whichever is closer. It turns out that $E'_4$ vs the individual respondents were as follows:

$E'_4$ vs R: 13 13½ 14 14 14½ 15½ 15½ 16 16 16½ 16½ 16½ 17 17 17 17½ 17 18 18½ 18½ 19
In view of these findings, if any further experimentation in this field were to be conducted, we would want to make the following recommendations:

Vary the degree of expertise among the respondents and examine the influence of this variation on convergence and on the effectiveness of selecting an elite subgroup.

Vary the amount and kind of information feedback in an effort to maximize proper as opposed to specious convergence.

Vary the methods of obtaining competence self-ratings.

Test the pure questionnaire technique, which avoids face-to-face confrontation, against both free roundtable discussions and pre-structured group meetings.

REFERENCES


