RM-6115-PR
November 1969

THE DELPHI METHOD, III: USE OF SELF RATINGS TO IMPROVE GROUP ESTIMATES

N. Dalkey, B. Brown and S. Cochran

prepared for
UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

Rand maintains a number of special, subject bibliographies containing abstracts of Rand publications in fields of wide current interest. The following bibliographies are available upon request:

Africa • Arms Control • Civil Defense • Combinatorics
Communication Satellites • Communication Systems • Communist China
Computing Technology • Decisionmaking • East-West Trade
Education • Foreign Aid • Health-related Research • Latin America
Linguistics • Long-range Forecasting • Maintenance
Mathematical Modeling of Physiological Processes • Middle East
Policy Sciences • Pollution • Procurement and R&D Strategy
Program Budgeting • SIMSCRIPT and Its Applications • Southeast Asia
Systems Analysis • Television • Urban Problems • USSR
Water Resources • Weather Forecasting and Control

To obtain copies of these bibliographies, and to receive information on how to obtain copies of individual publications, write to: Communications Department, Rand, 1700 Main Street, Santa Monica, California 90406.

Published by The Rand Corporation

RM-6115-PR
November 1969

THE DELPHI METHOD, III: USE OF SELF RATINGS TO IMPROVE GROUP ESTIMATES

N. Dalkey, B. Brown and S. Cochran

This research is supported by the United States Air Force under Project Rand—Contract No. F44620-67-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 Rand or of the United States Air Force.

Rand
SANTA MONICA, CA. 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 and RM-5957-PR.

The work reported is one facet of RAND's continuing study of methods of improving decisionmaking; and has applications to a wide range of military and civilian areas where the best information available on which to base a decision is the judgment of a group of experts.

SUMMARY

This report examines the possibility of using respondent self-ratings as a criterion for selecting more accurate subgroups in applications of the Delphi procedures for eliciting group judgments. The results of a series of experiments (16 groups, 20 subjects per group, 20 questions per subject) with upper-class and graduate college students answering almanac type questions indicate that significant improvements in accuracy of group estimates can be obtained with proper use of the self-ratings.

CONTENTS

THE DELPHI METHOD, III: USE OF SELF-RATINGS TO IMPROVE GROUP ESTIMATES

1. INTRODUCTION

The Delphi method is a set of procedures for formulating a group judgment for subject matter where precise information is lacking. In general, the procedures consist of obtaining individual answers to preformulated questions either by questionnaire or some other formal communication technique; iterating the questionnaire one or more times where the information feedback between rounds is carefully controlled by the exercise manager; taking as the group response a statistical aggregate of the final answers.*

In previous studies it has been shown that the Delphi procedures lead to increased accuracy of group responses more often than not, and that both the spread of answers (standard deviation of responses on a given question) and a self-rating index (average of individual self-ratings on a given question) are valid indicators of the mean accuracy of group responses.

On the other hand, in earlier studies inconclusive results were obtained with respect to using self-ratings as a technique for selecting a more accurate subgroup. Brown and Helmer (3) found a definite improvement

*A more complete discussion of the Delphi procedures and the rationale of their use is contained in (1) and (2).

by selecting the responses of "elite" subgroups based on self-ratings. Campbell (4) obtained somewhat more complex results where a single index of self-confidence, or one of self-rated competence, did not allow the selection of more accurate subgroups, but a combination of these two indices did under some circumstances. In a somewhat more extensive series of experiments we conducted last year (1), no consistent results were obtained when selecting subgroups on the basis of self-ratings. In the present report, a more complete analysis of this issue is presented, based on additional experiments. The major conclusion from this analysis is that when certain elementary safeguards are invoked, more accurate subgroups can be selected for a large proportion of questions. In addition, answers to the remaining questions improve upon feedback, so that a combination of subgroup selection and feedback produces a significantly larger number of improved group responses than could be obtained by feedback alone.

Figure 1 reproduces a curve from RM-5888 showing the empirical relationship between average self-ratings and group error. (Data points not shown.) Since average group error monotonically decreases with average self-rating, it should follow that if a group (C) with an intermediate rating is divisible into two subgroups (A and B), one of which (A) has a higher self-rating than the total group and the other (B) has a lower self-rating, then subgroup A should (on the

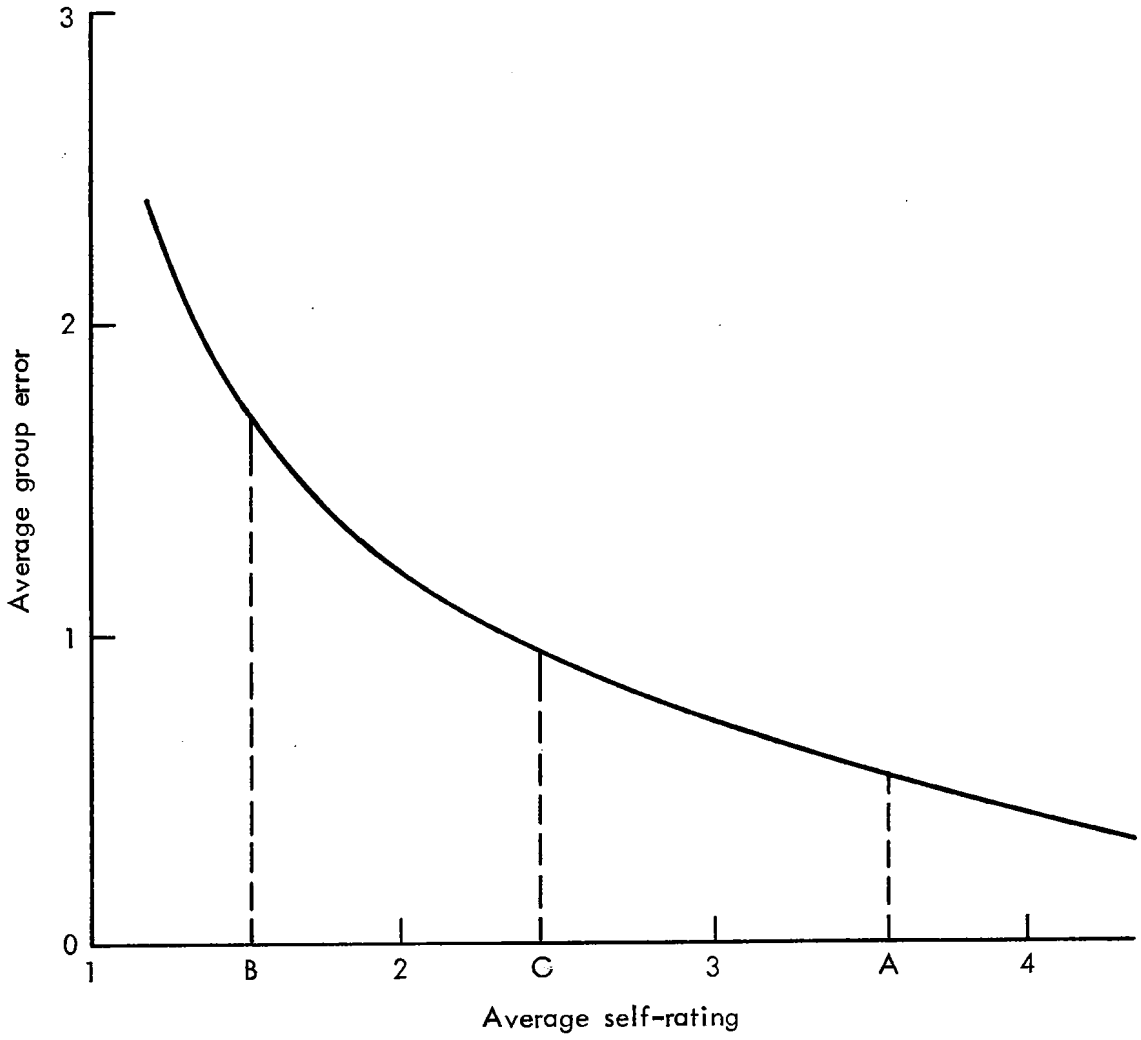

Fig. 1—Group self-rating

average) be more accurate than C, and subgroup B should be less accurate.

There are two considerations that could interfere with such a result: (1) The curve in Fig. 1 is rather "noisy",* i.e., the dispersion around the mean values shown in the figure is large. If the difference between the average self-ratings of the two subgroups is not substantial, any difference in accuracy might be lost in the noise. (2) By dividing the total group into subgroups, a countereffect is introduced. Figure 2 indicates the relationship between group size and accuracy obtained in the previous set of experiments. It is clear that reducing the size of the group from say, twenty respondents to three or four, would result in substantial reduction in average accuracy. The size effect, then, could mask any improvement resulting from the self-rating effect. It seems likely that these two considerations explain the negative results in attempting to select more accurate ("elite") subgroups in our previous analysis of Delphi data.

These considerations suggest two conditions that should be imposed on the selection of subgroups for increasing accuracy: (1) The difference in average self-rating between the subgroups should be substantial. (2)

---

*See the comparable curve in Fig. 3, for the present set of experiments, where quartile confidence limits are presented.

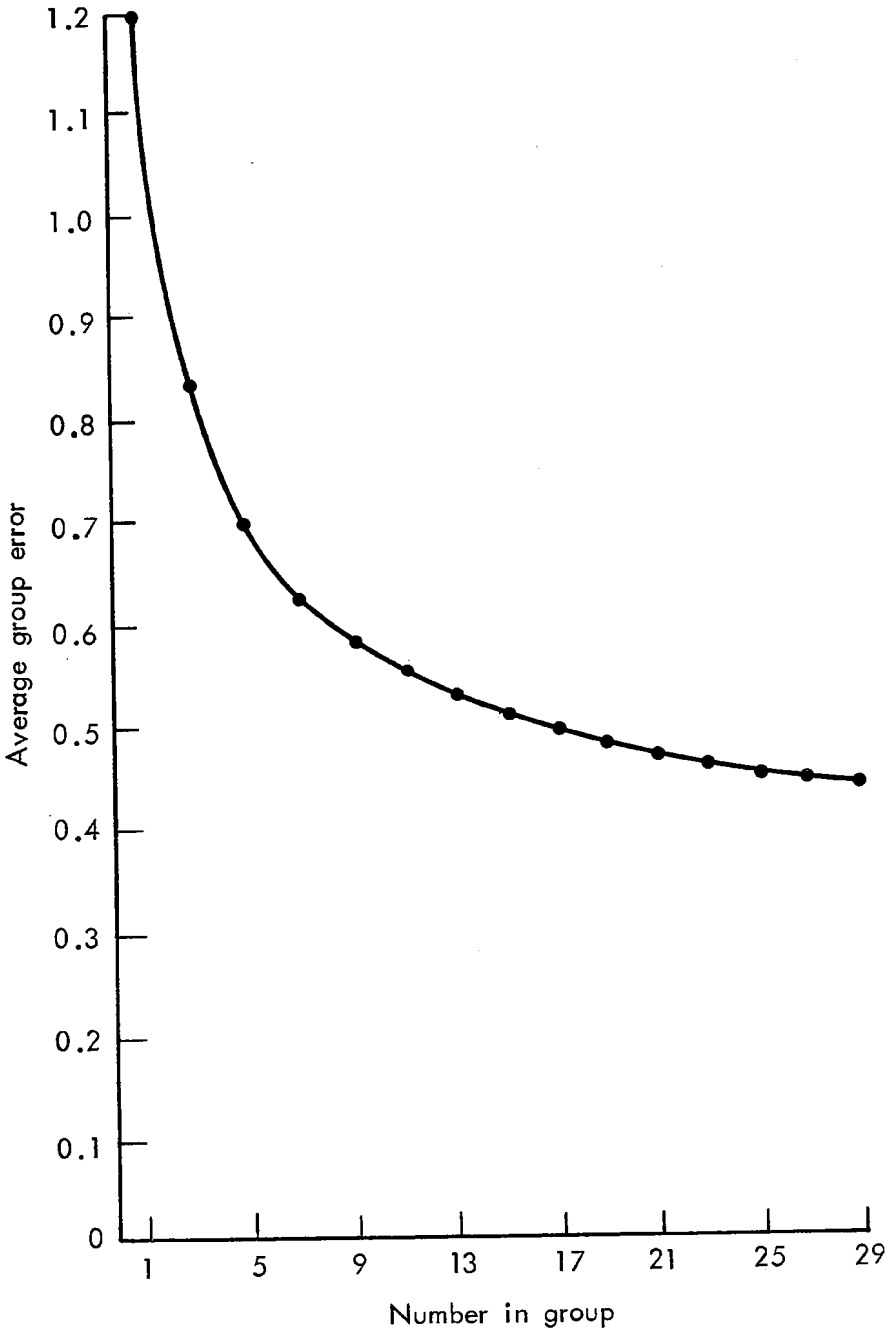

Fig. 2—Effect of group size

The size of the subgroups should be substantial for both the higher and lower self-ratings subgroups. The size that should be taken as substantial for this purpose is not sharply determined by the curve in Fig. 2. We selected 7 as the lower limit on the grounds that it was roughly in the middle of the "knee" of the curve. To implement the substantiality condition for group self-ratings, the rule was laid down that there should be a complete separation between the self-ratings of the two subgroups, i.e., the lower subgroup should contain all answers accompanied by a low self-rating, and the upper subgroup should contain all answers accompanied by a self-rating one degree higher or more. This guaranteed that the group self-rating for the two subgroups would differ by at least one degree, usually more.

The results of the analysis, described in detail below, show that these conditions are effective. The degree of improvement (proportion of improvements to total changes) using the self-rating subgroups is greater than the degree of improvement achieved by feedback, and the total number of improvements is also greater.

2. METHOD

Subjects

A total of 282 subjects—219 juniors and seniors and 63 graduate students—all enrolled at the University of California, Los Angeles were paid to serve in the experiment. The group contained 138 males and 144 females. Among undergraduates, the ratio of males to females was 5 to 6; among graduates, the ratio was 2 to 1.

Procedure

Subjects were randomly divided into 16 groups, 15–20 per group; each subject, working independently, answered 20 questions of a general information type, ten questions in a set. Almanac type questions were selected for which it was believed the subjects would not know the exact answer, yet would have enough general knowledge to allow them to make an estimate—an informed guess—of the answer.

Two groups were scheduled on each day of the experiment with both groups answering the same 20 questions. Different questions were prepared for each of the eight days the experiment was run, resulting in 160 different questions and 320 group responses supplied for each round.

Prior to answering the questions, subjects were given the following instructions and allowed five minutes to give their self-rating to a set of ten questions:

> First, you are asked to rate the questions with respect to the amount of knowledge you feel you have concerning the answer. Do this as follows: Before giving any answers, look over the first ten questions, and find the one that you feel you know the most about. Give this question a rating of 5 in the box on the left labelled "self-rating." Then find the one you feel you know the least about, and give this a rating of 1. Rate all the other questions relative to these two, using a scale of 1, 2, 3, 4, or 5. Thus, a question about which you know almost as much as the one you rated 5 could also get a 5 rating. One that you feel is roughly halfway between the one on which you are least informed and the one on which you are best informed would be rated 3, and so on. Notice that the rating is purely relative, and depends only on how much you feel you know about the question. Do not try to make refined estimates of these ratings, but be impressionistic. Follow the same procedure for questions 11-20.

An average group self-rating for each question was obtained by dividing the sum of the individual self-ratings by the number of subjects in the group. This resulted in a numerical index for each question, representing the relative amount of knowledge the group felt that it had about the question.

A measure of the group's accuracy in answering each question was obtained by dividing the median estimate of the group by the true answer, taking the natural logarithm,* and converting to the absolute value; i.e., the accuracy measure was

$$\left| \ln \frac{Md}{T} \right|,$$

where Md is the median estimate given by members of the group, and T is the true answer to the question.

---

*The reason for using a logarithmic transformation is given in (1), p. 25.

3. RESULTS, PART I—GROUP SELF-RATING

Group self-ratings for different questions ranged almost the full span of the five-point scale, 1.05 to 4.75, indicating a wide variation in the amount of knowledge groups felt they had about different questions. The average group self-rating for all questions was 2.53.

Rank correlations of the group self-ratings for groups answering the same 20 questions were:

Groups	Correlation
1 and 2	.87
3 and 4	.81
5 and 6	.83
7 and 8	.85
9 and 10	.96
11 and 12	.83
13 and 14	.92
15 and 16	.84

Average correlation for all groups was .875.

Correlations of round 1 error for groups answering the same questions were:

Groups	Correlation
1 and 2	.59
3 and 4	.80
5 and 6	.86
7 and 8	.87
9 and 10	.86
11 and 12	.79
13 and 14	.98
15 and 16	.84

Average correlation for all groups was .855.

Each of the 320 questions used in the experiments was placed into one of seven categories, based on the group self-rating, and an average error obtained for the questions falling into each of the seven categories.

Group self-rating	Number of questions in each category	Average group error for questions in each category
1.00 - 1.49	30	1.92
1.50 - 1.99	59	1.54
2.00 - 2.49	65	1.47
2.50 - 2.99	70	.71
3.00 - 3.49	53	.57
3.50 - 3.99	28	.56
4.00 and up	15	.38
	320	

Using the midpoints of the seven group self-rating categories and the average error obtained from the questions falling into each of the seven categories, the correlation is -.92.

Figure 3 plots the relationship between group self-rating and the median error for questions in each category. The median is used for this graph rather than the average because the distribution of error for a fixed group self-rating interval is highly skewed; hence the interquartile range is a better measure of dispersion than the standard deviation.

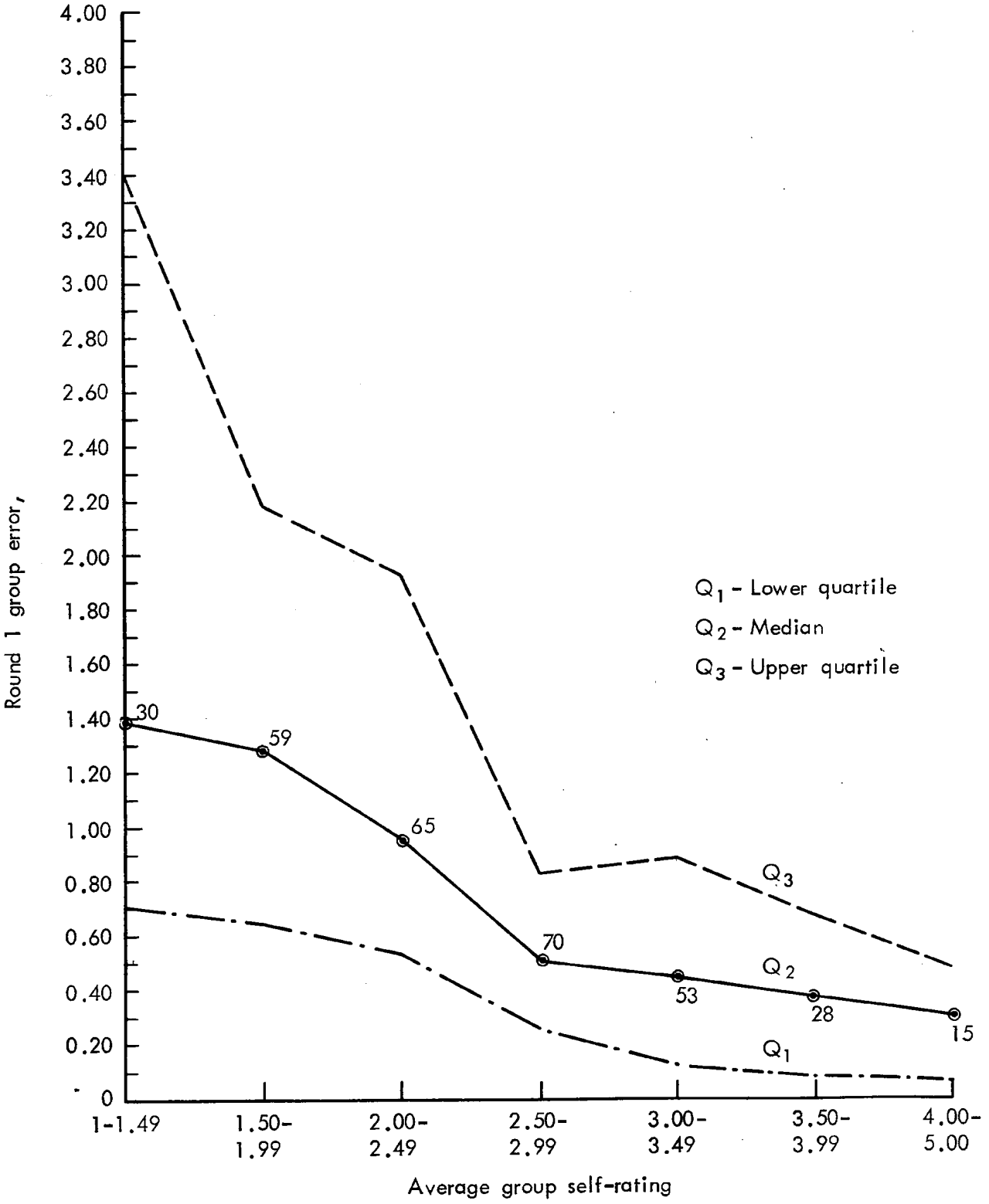

Fig. 3—Relationship between average group self-rating and group accuracy

4. RESULTS, PART II—SUBGROUP SELF-RATING

While most of the members of the group agreed on the level of difficulty of each question, the decision usually was not unanimous. An analysis was made to determine whether subgroups giving a higher self-rating to a question than the entire group self-rating would also give estimates closer to the true value than estimates obtained from the entire group.

Groups

Data for the analysis of Part II were supplied by Groups 1 through 9, 11, 13 and 15. Groups 10, 12, 14 and 16 were treated differently and consequently were eliminated from this analysis.

Procedure

Self-ratings for each of the 240 questions were examined with the aim of dividing the subjects answering the question into a high and a low subgroup, based on their individual self-ratings. The following rules were applied: First, at least seven subjects must be in each of the high and low subgroups. Second, the individual self-ratings in the low subgroups must all be lower than any in the high subgroup. For example, on a specific question, six subjects might give the question a self-rating of 5 and one subject a self-rating of 4; these would be placed in the high subgroup. The remaining subjects giving

self-ratings to the question between 1 and 3, would comprise the residual subgroup, the group designated as low in this case. A group accuracy measure was obtained for both the high and low subgroups.

The subgroup estimate given by the high subgroup and the estimate derived from the total group for a specific question were compared with the correct answer to the question to determine which estimate was closer.

Employing standard Delphi techniques, the group estimates to the 240 questions considered in this analysis changed from round 1 to round 2 as follows:

More Accurate	Less Accurate	Same	Total
86	61	93	240

High and low subgroups, based on self-ratings, were identified for 156 of the 240 questions. Using round 1* estimates, the relationship between the estimates of the high subgroup and the estimates of the total group was determined:

More Accurate	Less Accurate	Same	Total
95	52	9	156

*The results were essentially the same when using round 2 estimates.

The most striking difference between the effect of iteration and feedback and selection of subgroup answers is the reduction in the number of same answers for the latter.* This contributed largely to the increased number of more accurate answers for the subgroups. In addition, the ratio of more accurate to less accurate (.65) for the subgroups is greater than the ratio for feedback (.59). Although this difference in ratio is not statistically significant, it is in the desired direction.

Since adequate subgroups could be identified for only 156 out of the 240 questions, the issue remains whether a combination of subgroup selection for the 156 and feedback on the 84 remaining questions will still lead to an overall improvement.

For the 84 questions where subgroups could not be identified, group estimates changed from round 1 to round 2 as follows:

More Accurate	Less Accurate	Same	Total
33	21	30	84

The result of combining the two procedures, then is:

*The difference in the two patterns is significant at better than the .01 level on a χ^2 test.

	More Accurate	Less Accurate	Same	Total
High subgroup compared to total group on Round 1	95	52	9	156
Round 2 compared to Round 1	33	21	30	84
Total	128	73	39	240

The relationship between the average total group self-rating and the accuracy of the group's estimates on round 1, round 2 and the high subgroup is shown for the 156 questions for which high and low subgroups, based on self-ratings could be identified.

Group self-rating	Number of questions in each category	Average error for questions in each category		
		Round 1	Round 2	High (Round 1) subgroups
1.00 – 1.99	32	1.178	1.112	1.055
2.00 – 2.99	69	.875	.869	.865
3.00 – 3.99	45	.577	.541	.495
4.00 – up	10	.467	.423	.384
Total	156			

These results are shown in graphical form in Fig. 4.

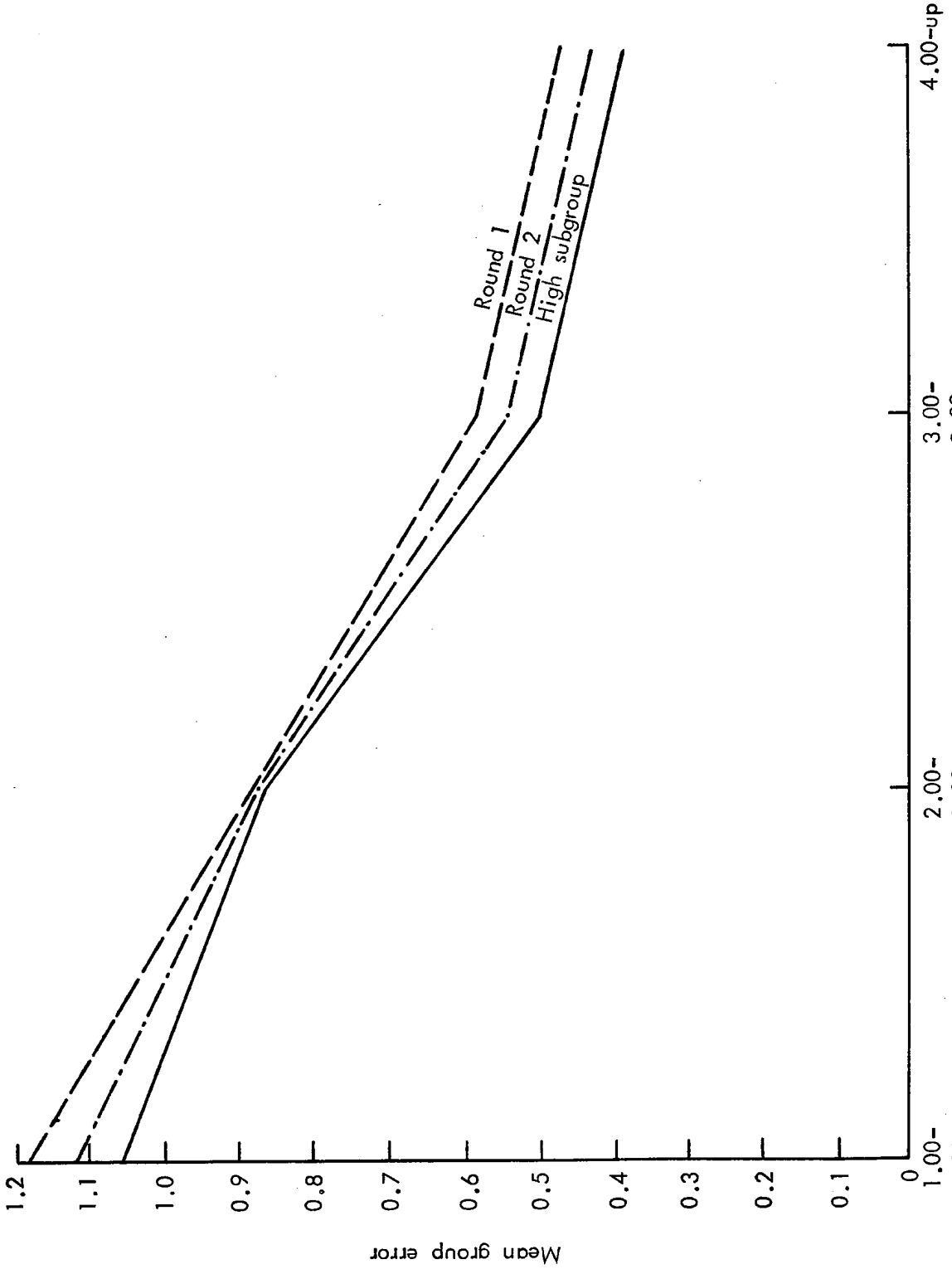

Fig. 4—Relationship between group self-rating and average accuracy

5. DISCUSSION

The results reported here are based on a separate set of experiments from those described in RM-5888 and thus act as an independent confirmation of the earlier results. This includes both the relationship between accuracy and group self-ratings and the improvement of group estimates with feedback.

In addition, the present study shows that a significant improvement in the effectiveness of the Delphi procedures can be obtained by using self-rating information to select more accurate subgroups. This fact reconfirms the relationship between group self-rating and accuracy.

Several other results of interest should be noted. The high and consistent correlations on accuracy between two groups answering the same question (p.10) are in line with the discussion of reliability in RM-5888 (pp. 10-12). The correlation of group self-ratings across groups on the same questions, averaging .875, is quite high, and indicates that the self-ratings are measuring some fairly well-defined property of the questions (for the given class of respondents.) This property is not immediately apparent from the instructions with which the students were asked to rate the questions, i.e., in terms of the two questions concerning which they knew the most and knew the least. Considering the diverse characteristics of the students with respect to major, age, school year, intelligence

scores, sex, and the like, the degree of uniformity in self-ratings is rather surprising, and appears to warrant further study.

There are a number of questions on which the data can shed light that we have not had time to analyse as yet. It is hoped that these can be reported in later publications. Among these are questions concerning major differences in the characteristics of the self-ratings depending on individual factors such as sex, grade level, and the like. Preliminary analyses indicate, for example that women are likely to rate themselves uniformly lower than men. Another significant question is whether the standard deviation of answers to individual questions can be employed to further improve the final set of answers to a Delphi exercise. In particular, it should be worth examining whether the standard deviation can be used to select a set of questions on the first round which are already sufficiently accurate (on the average) so that additional processing is more likely to make the responses less accurate.

There are several interesting questions concerning the micro-structure of a Delphi exercise that are raised by the present results that can be answered only by additional experiments. For example, a reasonable hypothesis would be that feeding back the medians of just the higher self-ratings subgroups for those questions where adequate

self-rating subgroups exist would lead to improvement. There is an accompanying question of whether the second round answers of just the higher self-rating subgroups should be selected as the group response.

The findings of this study form a substantial improvement in the state of the art of Delphi procedures. Above all, the fact that systematic results from previous experiments were able to furnish a guide to designing specific rules for processing the answers is an indication that the subject is beginning to take on more of the properties of a rational technology.

REFERENCES

1. Dalkey, N. C., The Delphi Method: An Experimental Study of Group Opinion, The RAND Corporation, RM-5888-PR, June, 1967.

2. Brown, B., S. Cochran, N. Dalkey. The Delphi Method II: Structure of Experiments, The RAND Corporation RM-5957-PR, June, 1969.

3. Brown, B. and O. Helmer. Improving the Reliability of Estimates Obtained from a Consensus of Experts, The RAND Corporation, P-2986, September 1964.

4. Campbell, R. M., "A Methodological Study of the Utilization of Experts in Business Forecasting," Unpublished PhD Dissertation, UCLA, 1966.