III.B. 4. A Delphi Evaluation of Agreement between Organizations

CHESTER G. JONES

Introduction

Delphi [1] is often used to combine and refine the opinions of a heterogeneous group of experts in order to establish a judgment based on a merging of the information collectively available to the experts. However, in this process it is possible to submerge differences of opinion and thus suppress the existence of uncertainty. In many situations it might be advisable to run separate Delphis using more homogeneous groups of experts in order to highlight areas of disagreement. This paper will report on an activity that did just this and point out several areas in which the types of responses obtained were fundamentally very different. In some cases these differences were quite unpredictable, and so, a highlighting of the variations greatly increased the information obtained. Running one Delphi using a subset of the experts from each group would probably not have illuminated some of the differences in opinion. The mere weight of pressure to move toward the median response [2] would have caused a joint Delphi to converge toward a middle position. In addition, the presence of disagreement is much more significant when large groups share similar positions. The traditional approach to Delphi generally results in the using of a small number of experts from any one area.

One concern that is often raised about the credibility of Delphi results is that individual experts may bias their responses so that they are overly favorable toward areas of personal interest. This is of particular concern when experts are asked to evaluate areas in which they are presently working and when the final Delphi results could impact the importance attached to these areas. In this paper results will be presented that indicate that no such bias occurred in the Delphis reported on. It appears that the particular groups of experts used were able to rise above the desire to protect personal interests.

Background

The United States Air Force presently maintains an official list of System Concept Options (SCOs) in order to indicate to the Air Force Laboratories potential future technology needs. This activity is primarily- a means of communicating to the laboratory planners the thinking of Air Force System planners. However the number of potentially worthwhile systems possibilities, and thus the number of technology needs, exceed the resources available to fulfill all the possibilities and needs. Clearly the Air Force Laboratories needed a means of establishing priorities for the System Concept Options. Thus it was decided to undertake a program of Delphi evaluation. This program was run by the Deputy for Development Planning, Aeronautical Systems Division, and was limited to considerations of those SCOs that fell under the Deputy's jurisdiction. Thirty SCOs were evaluated. They covered a rather large spread in need for technological support as well as proposed mission use. Some concepts represented a rather straightforward extrapolation of present technology, while others would require substantial technology development programs. The missions represented included most of the areas of interest to the Air Force including many strategic and tactical possibilities as well as systems intended to meet support and training requirements.

It was decided to conduct separate Delphis utilizing personnel from various Air Force organizations, in order to determine how closely the organizational opinions agreed. In this way it was believed that not only would a basis for prioritizing the systems be obtained, but in addition, the results would help to indicate areas of communication problems between organizations. If organization viewpoints in a particular area differed greatly, there would appear to be a need for increased communication about the area.

Delphis were conducted within the following four USAF organizations: Deputy for Development Planning, Aeronautical Systems Division (ASD/ XR); Air Force Avionics Laboratory (ANAL); Air Force Aero Propulsion Laboratory (AFAPL); Air Force Flight Dynamic Laboratory (AFFDL). The experts chosen were senior managerial and technical personnel (both civilian and military), and were selected so that representation of most if not all of the major departments within the organizations was present. A total of sixty-one experts took part in the evaluations which involved three rounds of questioning.

The above organizations are of two different types. The Deputy for Development Planning is a systems planning *organization* having responsibility for identifying promising aerodynamic system concepts and defining them to the point where development decisions can be made. It has no direct responsibility for research activities. The three laboratories are responsible for developing technologies in their assigned areas which will improve system capabilities. The Avionics Laboratory is concerned with electronic systems, the Aero Propulsion Laboratory with atmospheric engines, fuel, etc., and the Flight Dynamics Laboratory with aircraft structures, controls, aerodynamics, etc. Thus the four groups that were asked to evaluate the list of SCOs are quite different in their areas of expertise. In particular it should be emphasized that the laboratory groups were being asked to compare SCOs some of which required considerable support from their particular laboratory, others of which required little or no support. All of the participants were, however, senior Air Force personnel and were thus knowledgeable of activities at other Air Force Laboratories.

Results obtained for three of the questions used will be discussed in this paper.

Question 1. Please rank-order the SCO list of systems on the basis of where current Air Force Laboratory Programs will make the greatest contribution toward success of the system.

Question 4. Given that each system becomes a technological success, rank order the SCO list in terms of importance of each system to National Defense.

Question 5. Considering technology, timing, and system importance, rankorder the SCO list according to where you think the Air Force Laboratories can make the greatest contribution to National Defense.

Each of these questions involved a complete ranking of thirty items, which proved to be a trying but not impossible task. It should be noted that succeeding round changes in answers often required a large restructuring of the list. That is, a change in the answer or rank of one system generally changes the rank of other systems (however, the participants were allowed to use a limited number of ties if necessary and thus a few participants avoided this problem). This interrelation of answers tends to make convergence difficult, since dis agreement in one area impacts other areas.

Convergence

One indication of the effect of a Delphi experiment is the amount of convergence caused by the iteration process, where convergence is a measure of how much more agreement is achieved on succeeding rounds as opposed to the first-round response. In this effort, one measure of convergence was the change in the spread between the lower and upper quartile values for a given question and a given SCO. In all of the Delphi experiments the spread between the lower and upper quartile values generally showed considerable reduction during the course of the efforts. However, as indicated in Fig. 1, the average amount of convergence varied considerably from group to group for some questions.

All of the groups achieved basically the same degree of convergence for Question 1. However, the convergence on Questions 4 and 5 follow significantly different patterns. In particular, the ASD/XR group achieved less convergence on Questions 4 and 5 than that achieved by the laboratory groups. The ASD/XR group was the only group primarily composed of system planners and so this group's failure to converge as well as the other groups on Question 4 would seem to be quite important. The ASD/XR group should be the best suited to serve as experts concerning Question 4. Thus, for Question 4 the greatest uncertainty is associated with the most expert group. If one Delphi had been run combining experts from each of those groups, it appears possible that the greater convergence between the laboratory experts might have caused a considerably better overall convergence than that shown in the ASR/XR result. Thus the relatively less expert participants might have caused the creation of a false sense of expert agreement.

Correlation between Questions

In reviewing the results, it was obvious that some groups tended to give SCOS similar rankings for different questions, while other groups changed many of the SCOs rankings drastically from question to question. Table 1 shows the pearman rank correlation coefficient for each Delphi for each combination of questions.

Clearly the ASD/XR answers suggest a greater change in laboratory emphasis (as shown by the low correlation between Questions 1 and 4, and between Questions 1 and 5) than that indicated by the other three groups. The system planners thus indicated a greater need for laboratory redirection than the laboratory personnel. Again we have an

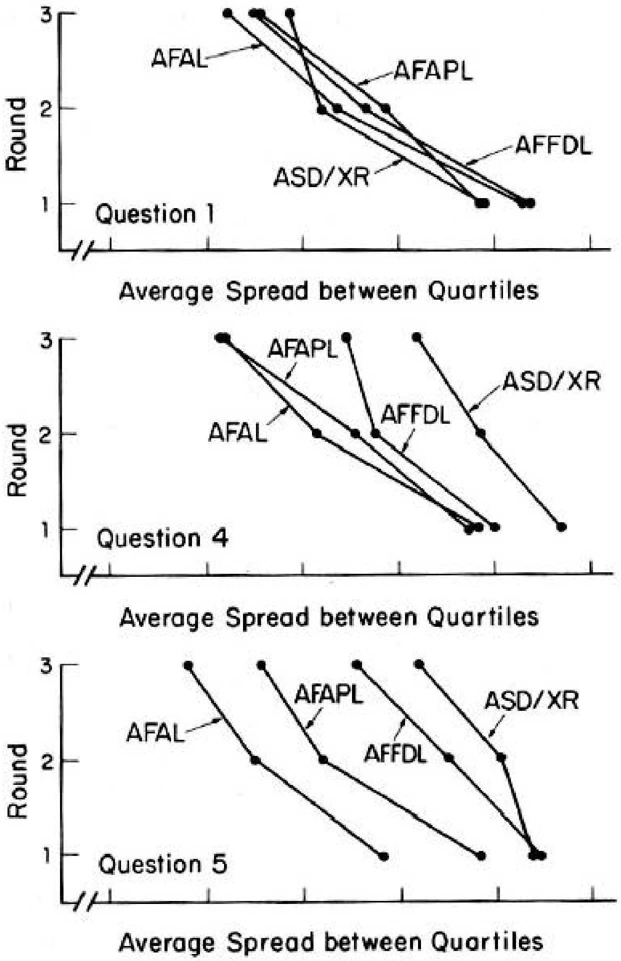

Fig. 1. Average interquartile spreads for questions 1, 4, 5.

area of disagreement that might be camouflaged had one combined Delphi been utilized.

It is interesting that the AFAPI. results indicate the least correlation between Questions 4 and 5. Although it might seem that the answers to these questions should correlate closely, there are several possible reasons to explain lack of correlation:

(1) A system may be important but not need substantial laboratory support.
(2) The necessary laboratory support might best be supplied by non-Air Force Laboratories.
(3) A system might be important if technologically feasible, but the necessary technological developments might not be considered likely in the near future.

Table 1

Spearman Rank Correlation Coefficient for Each Question Combination

Questions	ASD/XR	AFAL	AFAPL	AFFDL
Q1-Q4	+.295	+.788	+.571	+.448
Q1-Q5	+.315	+.863	+.904	+.698
Q4-Q5	+.746	+.925	+.579	+.844

Thus there might be a logical explanation for this lack of correlation. However, the data are surprising enough to indicate the desirability of a more detailed review of the AFAPL results. A subsequent review of the AFAPL answers indicated that many of the comments used to justify the apparently inconsistent results did involve considerations such as those listed above. However this example shows the value of looking for correlation between answers, and then, highlighting comments that justify departures from expected correlation.

Bias by Time Period

Figure 2 shows the average evaluations for Question 5 when the SCOs are grouped according to date of estimated technological feasibility. Obviously the system planners (ASD/XR) with their more futuristic interests attach greater importance to the far-term, more advanced systems. This might be a result of the planners' greater awareness of the possible benefits these futuristic systems offer. However, a possible reason for the laboratory viewpoint might be a greater appreciation of the difficulty associated with solving the technological problems.

Again the results suggest the possibility of a communications gap. Both groups should benefit from an exposure to the reasoning that led to such diverse results. This type of exposure might best go beyond a Delphi-type exchange (which is generally limited in the amount of information transferred). Such a transfer of information is essential if the potential value of the SCO list is to be achieved. It is often not enough to establish priorities, unless all parties concerned accept and understand the logic that led to the priorities.

Laboratory Bias

There was some concern before the laboratory efforts were started that the results might tend to be biased. Although the laboratory participants were instructed to rank the laboratory efforts by the total efforts from all the Air Force Laboratories, it was hypothesized that the participants' greater knowledge about their own laboratory programs and the natural tendency to promote one's personal interests would lead to a

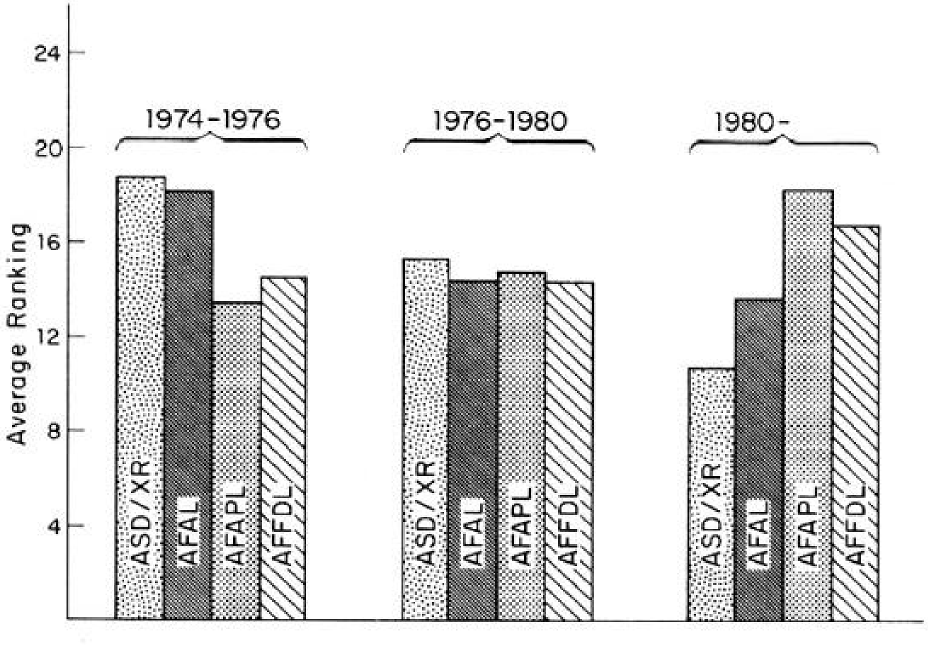

Fig. 2. Average responses for Question 5 when SCOs are grouped by year in which engineering development could start.

bias in favor of their laboratory's efforts. In order to test this hypothesis, the rankings obtained on Question 5 for the SCOs that received crucial support from each of the laboratories were compared.

In mid-1972, each of the laboratories published reports that reviewed their Technology Planning Objectives (TPOs) and the relevance of each TPO to each SCO. The top relevancy category indicated a TPO that the laboratory felt was essential to a given SCO. Table 2 shows the average ranking given for Question 5 to the groups of SCOs having a top relevancy match with the various laboratory TPOs, respectively. The lowest number in each column indicates the organization placing the greatest emphasis on that laboratory's progress. Therefore, bias would be indicated if the lowest number in a given laboratory's column was on the row corresponding to that laboratory's Delphi. The Delphi conducted in Laboratory B gave poorer (larger numerically) rankings to SCOs that were felt to be essentially related to one or more of their TPOs than any of the other groups, while the Delphi conducted in Laboratory A gave neither the poorest nor the best rankings to SCOs that were felt to be essentially related to one or more of their TPOs. Although the Delphi conducted in Laboratory C did give the best (numerically lowest) ranking to SCOs considered to be essentially related to their TPOs, the average ranking is not too different from those obtained in the other Delphis.

Table 2

Average Answer to Question 5 for SCOs That Are Related to Programs of Particular Laboratories

DELPHI	Relevancy Match with		
	Laboratory A	Laboratory B	Laboratory C
ASD/XR	12.7	15.1	15.6
Laboratory A	12.0	17.6	15.9
Laboratory B	15.1	22.3	15.8
Laboratory C	11.6	20.0	14.9

Thus, the hypothesis that a given laboratory Delphi tends to indicate biased rankings for SCOs that receive crucial support from that laboratory's effort does not appear to be valid. The answers obtained from Question 5 do not indicate the presence of laboratory bias.

Summary

The results discussed in this paper indicate information that was obtained by comparing several Delphi experiments utilizing experts from different organizations that probably would not have been obtained had one Delphi been run utilizing a subgroup from each of the groups of experts. Clearly differing organization viewpoints were identified, despite the fact that all of the groups involved very senior Air Force personnel who shared access to a considerable common information base. That is, all of the organizations had detailed knowledge of many of the same programs.

A noticeable difference in the amount of convergence was observed where in one case the apparently more expert group showed the poorest convergence. Disagreement was also apparent concerning the question of whether or not the laboratory programs should be redirected (as well as the related question of whether laboratory efforts should be directed toward near-term or more futuristic technology needs).

Comparisons of results were also made to determine if the laboratory group gave answers that were biased to support their own program. This investigation failed to show the presence of any real bias. This finding is very encouraging, for it suggests that at least these groups of technical experts were able to place their professional ethics above the common desire to promote personal gain. Had this not been true, the worth of this activity would be greatly reduced.

References

1. Olaf Helmer, and Nicholas Rescher, *"On the Epistemology of the Inexact Sciences," Management Science, 6, No. 1 (October 1959).*
2. Norman C. Dalkey, *The Delphi Method: .9n Experimental Study of Group* Opinion, The Rand Corporation, RM-5888-PR; *1969.*