CHAPTER 3

Documents and Questions

To provide the necessary basis for the test, we required a collection of documents, a set of search questions, and a complete assessment to determine the documents relevant to each question. These aims were accomplished in three main stages:

Stage 1. A letter was sent to authors of research papers, requesting search questions and a relevance assessment of the papers they cited.

Stage 2. Using the collection of documents and a set of questions made up from the replies to stage one, technically competent people examined every document in relation to every question to find any relevant documents in addition to the authors' cited documents.

Stage 3. The additional documents judged relevant in stage two were submitted to the authors, requesting their final assessment of relevance.

First will be given details of the methods used in these three stages, and the response made by the authors. Then will follow a more detailed examination of the question-document assessment of relevance, and finally a brief analysis of the questions.

Methodology and authors' responses

271 recent papers on the subject of high speed aerodynamics and aircraft structures were obtained. Although high speed aerodynamics had been chosen as the main subject for the test, a small set of documents dealing with aircraft structures was introduced in order to examine the effect of including two dissimilar subjects in one collection. These papers were referred to as base documents, and in order to be accepted for the test a base document had to satisfy certain criteria; it had to be a paper published in the English language containing at least two references in a bibliography, these references being in English, dated 1954 or later and likely to be readily obtainable. Since aerodynamic papers contain on average about twelve references, neither this nor any of the other requirements caused the rejection of many papers. Most of the selected papers were published during 1962, and the first half of 1963; the articles from one prominent journal predominated, but some research reports were included. A list of the different sources of those which were finally used is given in Table 3.1. 76.9% of the papers are American publications, 22.5% British and 0.6% Swedish.

To the author of each of these papers was sent a form, quoting the title and reference of his own paper, and also listing up to ten of the papers which had been included as references. The authors were asked to do two things.

1. To state the basic problem, in the form of a search question, which was the reason for the research being undertaken leading to the paper, and also to give not more than three supplementary questions that arose in the course of the work, and which were, or might have been, put to an information service.

	Total
U.S.A.	
Journal of the Aerospace Sciences (later A.I.A.A. Journal)	102
National Aeronautics and Space Administration Technical Notes	38
Great Britain	
Royal Aircraft Establishment Reports and Notes	22
Aeronautical Research Council Papers	6
National Physical Laboratory Reports	3
National Gas Turbine Establishment Reports	1
Southampton University Reports	1
College of Aeronautics Reports	3
The Aeronautical Quarterly	3
Journal of the Royal Aeronautical Society	2
Sweden	
Aeronautical Research Institute Reports	1

TABLE 3.1 BIBLIOGRAPHICAL ORIGIN OF BASE DOCUMENTS USED IN THE TEST

	Totals
Australia	1
France	1
Great Britain	49
India	1
Israel	2
Japan	3
Sweden	1
Switzerland	1
United States	123

TABLE 3.2 COUNTRY OF RESIDENCE OF AUTHORS OF BASE PAPERS

	U.S.A.	G.B.	Other
AUTHORS' COUNTRY	67.6%	26.0%	5.5%
	(123)	(49)	(10)
COUNTRY OF PUBLICATION	76.9%	22.5%	0.6%
	(140)	(41)	(1)

TABLE 3.3. COMPARISON OF AUTHORS' COUNTRY OF RESIDENCE AND COUNTRY OF PUBLICATION

2. To assess the relevance of each of the submitted list of papers which had been cited as references, in relation to each of the questions given. The assessment was to be based on the following scale of five definitions:

(i) References which are a complete answer to the question. Presumably this would only apply for supplementary questions, since if they applied to the main question there would have been no necessity for the research to be done.

(ii) References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.

(iii) References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.

(iv) References of minimum interest, for example, those that have been included from an historical viewpoint.

(v) References of no interest.

An example of a completed sheet was included with each letter; this, the covering letter and example of material sent, are shown as Appendix 3 A.

It was originally expected that half the authors would complete the form to our requirements, and that there would be an average of two questions with each reply. During early March, 1963, 82 letters were sent out and by the end of that month 47 replies had been received with an average of 3½ questions. Further letters were despatched up to the middle of July, and then later one chase letter was sent to those who had not replied. By the end of September we had received the excellent response of 182 completed forms of the 271 sent (67.2%). Some authors wrote to say that they could not spare the time; many other letters were returned because change of address prevented delivery. The authors continued to supply an average of 3½ questions, and the total of those received was 641.

Most of these authors, 67.6% lived in the U.S.A., with 26.9% in Great Britain and 5.5% in other countries. Table 3.2 shows the figures from each country, based on the 182 authors with whom we corresponded. A complete list of the authors is given in Appendix 3F. It is an interesting sidelight on publishing habits to notice that eight of the British authors published in American sources, and nine out of ten of the other foreign authors did the same, but all the authors residing in the U.S.A. published there. Figures are given in Table 3.3. Some of the authors had changed their country of residence by the time of the test, and the figures are based on the country of residence in which their particular research paper was written.

As the forms were being received, the document collection was being made up, and 1,018 unique documents resulted from the cited papers. The base documents themselves were also included in the collection, adding 173 more documents (9 were already included as cited papers), but in order to avoid any possible bias in the results, these base documents are always completely deleted from the results when the questions to which they gave rise are being tested, 209 further documents, taken from similar sources, brough the whole collection to its final 1,400 documents. For the indexing, which was proceeding during this time, single xerox copies of the documents were made. Full bibliographical information concerning the document collection is given in Appendix 3C.

To prepare for the next stage, 361 of the 640 questions were selected for use in the test. The basis for this selection was questions that had two or more documents assessed as relevance grade 1, 2 or 3, and questions that were grammatically complete were selected first. Some questions were received abbreviated, although the missing idea was quite clear from another of the author's questions. For example,

Q. 247 when received was worded "Can the hypersonic similarity results be applied to the technique". By examination of the other supplementary and basic questions, (Q.13, Q.12) it is seen that the technique under investigation is methods for predicting surface pressures of an ogive forebody at angle of attack, so question 247 was rewritten to include this. When, as in the example, the meaning was quite obvious, we inserted the missing words, and the re-submission of the question to the authors in stage three revealed no disagreement with the amendments.

The next task was to find whether there were in the collection documents other than those which had been in the list of citations, which were also relevant to any of the questions. This was done by examining every document in relation to every question, noting any new documents that were judged as possibly relevant, and then submitting these documents to the original authors for their final assessment of relevance.

The task was performed by students, with a knowledge of aerodynamics, who were engaged in post-graduate study at the College of Aeronautics. Over 1,500 man-hours of effort during the 1963 summer vacation were put in by five people. The job involved in theory over half a million individual judgements, and was an extremely onerous task. The questions were supplied on individual slips, with space given for recording the file number of any document judged as relevant. Access was also given to the original forms giving all the questions supplied by the author, the source document, and the authors' relevance assessment of the cited papers. Details of the document collection were supplied in the form of typed sheets, listing the documents in file order, and giving authors, titles and bibliographical details. Complete copies of all the documents were readily available to the students.

The ultimate procedure adopted was to work on sections of the document collection, ranging from 100 to 400 documents, depending on the number of people working at the same time. The questions were first sorted into broad subject groups, and small batches of very similar questions were done together. Thus some of the prominent features and subject areas of sections of the documents were soon committed to memory, to assist fairly rapid scanning of the document lists. The document titles were examined first, and any documents that could remotely contain material connected with the question were recorded on the question slip, so that at the end of a 'scan' of the titles, the documents themselves could be examined. The students were instructed to be quite liberal in their judgements, and to include documents that they considered were only possibly relevant. An initial attempt was made to grade their decisions for relevance, but this was found to be too difficult to do consistently, and so was given up.

The task was tedious, particularly for people of intellectual capability, but 361 questions were finally completed. Those who carried out the task would not claim to have found every possibly relevant document, since question interpretation would not always agree completely with the authors' real need, and since human error was inevitable. Some figures giving information on the number of relevant documents missed by the students is given later in this chapter. Documents judged as relevant, which really were not, did not cause any difficulty, since the original author of the question was taken as the final arbiter of relevance. For 86 of the 361 questions, no other documents were considered to be relevant; for the other 275 questions, there was found at least one document judged as possibly relevant, with an average of 3.3 per question.

When submitting these documents to the authors, it was decided to add some extra documents which had been suggested as a result of a test of the questions by the technique known as bibliographic coupling. A description of the processing of the cited papers in the documents of the collection, which resulted in a citation index and bibliographic coupling groups, is given in chapter 7. In the theory of bibliographic coupling, as worked out by Dr. M. M. Kessler, (Ref. 20) it is shown that, as the coupling strength increases, so also does the probability of the document being relevant to the question. It was therefore decided to include all documents retrieved by bibliographic coupling at a coupling strength of 7 or more (i.e. documents that had seven or more references in common with one of the author's cited relevant papers of grade (1), (2) or (3)). Of the 213 documents produced in this way, only the unexpectedly small number of 15 had already been assessed as possibly relevant by the students. The balance of 198 were submitted, along with the student assessed documents, in the second communication to the authors. This time the authors were requested to do three things; for reasons considered later.

1. To make a relevance assessment of the new documents submitted, in relation to their search questions, using the same relevance scale as before.

2. To examine the selected questions (which they themselves had originally asked), and to indicate the relative importance of each term or concept in the question by marking with a 'weight' from the following scale:-

(i) A paper that did not cover this term would be of no use.
(ii) It is desirable that this term should be covered by the document.
(iii) This is a term which is not absolutely essential to the enquiry.

3. To list any alternative terms or concepts that might be used in a search programme for the questions and, if necessary, to include a completely rephrased version of the question.

A xerox copy of the questions as he originally wrote them was sent to each author, together with a list of the new documents submitted, giving authors, titles and bibliographical references. Against each such document submitted was indicated the question to which the document was thought to be relevant, and to assist the relevance assessment a xerox copy of each document abstract was included. Each of the questions was re-submitted on a separate sheet, with space provided for alternative words to be added, either against each single term, or the concepts of the questions. Examples of the above are given in Appendix 3B.

Most authors received a total of at least eleven sheets for examination, which together with the abstracts of the documents submitted, made a somewhat daunting package. In spite of this, 144 out of 182 authors (79.1%) returned completed forms, with yet others being unable to help and some having changed addresses as before. Our main concern was to obtain the relevance assessments, which were needed for 283 of the questions and the authors' responses provided assessments for 201 of these. 78 questions had not been resubmitted to the authors because no possible relevant documents had been noted; adding these to the 201 questions where the relevance assessments had been completed meant that there was a total of 279 questions which could be used. This fell slightly short of the 300 questions originally planned; as will be considered later, we were by this time beginning to suspect that the test would provide more data than could be handled or would be required, and therefore no effort was made to bring the total number of questions back to 300.

Most authors included the weighting of the questions in their reply, over half of the questions had some alternative terms added, and 28 of the questions were submitted in rephrased form. (See Appendix 3B).

A summary of the position regarding the questions is as follows:-

1.	Total of questions received	641
2.	Questions discarded for various reasons	280
3.	Questions matched against complete document collection for relevance ((1) - (2))	361
4.	Questions having no additional relevant references	78
5.	Questions resubmitted to authors for relevance decisions ..	283
6.	Questions returned by authors from stage (5)	201
7.	Questions available for test ((4) + (6))	279

The relevance assessments

The basic data on the authors' relevance assessments is given in Tables 3.4, 3.5, 3.6 and 3.7. These tables highlight various aspects of the relevance assessments, and the figures given are taken from the 279 usable questions obtained. In each table, the documents that were submitted to the authors are split into three categories:-

1. Those cited in the author's own original paper;
2. Those the students found and judged as being relevant;
3. Those retrieved by bibliographic coupling at a strength of 7 plus, and which were additional to the two categories above.

Each table also gives a figure for the total of all categories, the four divisions being shown as the left hand parameter in each table. The relevance assessments made are given in the body of the tables, these being split into several categories:-

1. Documents submitted (Tables 3.4 and 3.6)
2. Documents assessed as relevant, i.e. accepted:-
    (a) Totals (Tables 3.4, 3.5, 3.6 and 3.7)
    (b) Details of the four grades of Relevance (Tables 3.5 and 3.7)
3. Documents assessed as not relevant, i.e. rejected (Tables 3.4 and 3.6)
4. Total documents assessed as relevant expressed as a percentage of documents submitted. (Tables 3.4 and 3.6).

The figures given are in two forms in each table:-

1. Grand totals of documents, resulting from the whole set of questions involved.
2. Figures for one average question, calculated by the arithmetic mean. These averages are correct to one decimal place, but in a few cases a slight adjustment has been made to preserve the correct totals.

Tables 3.4 and 3.5 giving the figures for the whole set of 279 questions will be examined first. The bottom section of Table 3.4 shows that 3,087 documents were submitted to the authors of which 1,126 were rejected as not relevant, and 1,961 (i.e. 63.5%) were accepted as relevant. Table 3.5 gives a breakdown of the 1,961 documents accepted, showing that 171 were graded relevance (1), 461 were relevance (2),

Origin of documents	Submitted to the authors for assessment	Accepted as relevant	Rejected as non-relevant	% accepted as relevant
1. Cited in authors' papers	1972 (7.1)	1250 (4.5)	722 (2.6)	63.4%
2. Additional documents selected by students	917 (3.3)	592 (3.1)	325 (1.2)	64.6%
3. Additional documents by bibliographic coupling	198 (0.7)	119 (0.4)	79 (0.3)	60.1%
4. Complete total	3087 (11.1)	1961 (7.0)	1126 (4.1)	63.5%

TABLE 3.4 RELEVANCE ASSESSMENTS OF DOCUMENTS AS DECIDED BY AUTHORS IN RELATION TO THEIR SEARCH QUESTIONS

The total for all 279 questions is shown, with the average for each question in brackets.

		Grades of Relevance			
Origin of documents	Relevant documents	1	2	3	4
1. Cited in authors' papers	1250 (4.5)	158 (0.6)	348 (1.2)	492 (1.8)	252 (0.9)
2. Additional documents selected by students	592 (2.1)	12	97 (0.4)	344 (1.2)	139 (0.5)
3. Additional documents by bibliographic coupling	119 (0.4)	1	16 (0.1)	66 (0.2)	36 (0.1)
4. Complete total	1961 (7.0)	171 (0.6)	461 (1.7)	902 (3.2)	427 (1.5)

TABLE 3.5 GRADES OF RELEVANCE AS DECIDED BY THE AUTHORS

The total for all 279 questions is shown, with the average for each question in brackets. It will be noted that this table represents a breakdown of the figures as given in the second column of Table 3.4.

Origin of documents	Total submitted to the authors for assessment		Accepted as relevant		Rejected as not relevant		% accepted as relevant	
	Basic	Supp.	Basic	Supp.	Basic	Supp.	Basic	Supp.
1. Cited in authors' papers	820 (7.0)	1152 (7.2)	589 (5.0)	661 (4.1)	231 (2.0)	491 (3.1)	72.0%	57.3%
2. Additional documents selected by students	351 (3.0)	566 (3.5)	258 (2.2)	334 (2.1)	93 (0.8)	232 (1.4)	73.5%	59.0%
3. Additional documents by bibliographic coupling	85 (0.7)	113 (0.7)	59 (0.5)	60 (0.4)	26 (0.2)	53 (0.3)	69.4%	53.1%
4. Complete total	1256 (10.7)	1831 (11.4)	906 (7.7)	1055 (6.6)	350 (3.0)	776 (4.8)	72.2%	57.6%

TABLE 3.6 THE RELEVANCE ASSESSMENTS GIVING A COMPARISON OF BASIC AND SUPPLEMENTARY QUESTIONS

This table gives the same data as Table 3.4 except that the 279 questions are divided into the two groups of 118 basic questions and 161 supplementary questions, the figures in brackets representing the average for each question.

Origin of documents		Relevant Total	Grades of Relevance			
			1	2	3	4
1. Cited in authors' papers	(B)	589 (5.0)	12 (0.1)	159 (1.3)	273 (2.4)	145 (1.2)
	(S)	661 (4.1)	146 (0.9)	189 (1.2)	219 (1.4)	107 (0.6)
2. Additional documents selected by students	(B)	258 (2.2)	1	53 (0.5)	144 (1.2)	60 (0.5)
	(S)	334 (2.1)	11 (0.1)	44 (0.3)	200 (1.2)	79 (0.5)
3. Additional documents by bibliographic coupling	(B)	59 (0.5)	0	6 (0.1)	30 (0.2)	23 (0.2)
	(S)	60 (0.4)	1	10 (0.1)	36 (0.2)	13 (0.1)
4. Complete total	(B)	906 (7.7)	13 (0.1)	218 (1.9)	447 (3.8)	228 (1.9)
	(S)	1055 (6.6)	158 (1.0)	243 (1.6)	455 (2.8)	199 (1.2)

TABLE 3.7 RELEVANCE ASSESSMENTS GIVING A COMPARISON OF BASIC AND SUPPLEMENTARY QUESTIONS FOR ALL GRADES OF RELEVANCE

This table gives the same data as Table 3.5 except that the 279 questions are divided into the two groups of 118 basic questions and 116 supplementary questions, with the average for each question in brackets.
(B) = Basic question (S) = Supplementary question.

902 were relevance (3), and 427 were relevance (4). In terms of an average question, one can read off the figures as 11.1 submitted, 4.1 rejected, 7.0 accepted, and so on.

Examining the different origins of the documents in turn, the cited papers are seen to exceed all the other categories in size. From this group 4.5 documents per question were assessed as relevant; the additional groups of documents added another 2.5, making an average of seven relevant documents for each question. 63.4% of the cited documents submitted were accepted as relevant, and this seems satisfactory when it is remembered that all the references cited would not be relevant to all the questions given. In many cases some references are relevant to one of the questions only, and not relevant to the other questions at all. Table 3.5 shows that 14% of the relevant documents were graded as relevance (1), and some more details concerning this will be given when considering Table 3.7.

The additional papers that the students judged as relevant totalled 917. These are not, of course, 917 unique documents, as one document might be relevant to several questions. The acceptance rate was 64.6%, and this may be taken as a clue to the success of this difficult task, but further details are given when Tables 3.6 and 3.7 are examined, and when comment is made on the success of the students' judgements. Of the 592 accepted, only 12 (2%) were graded at relevance (1), so in most cases the authors considered these additional papers submitted were not as relevant as the cited ones about which they already knew.

The additional bibliographic coupling documents, submitted because they had seven or more of their references in common with the cited papers of relevance (1), (2) or (3), were only those which had not already been selected by the students as possibly relevant (see chapter 7). Table 3.8 shows that of the 312 documents retrieved by bibliographic coupling, 87 were cited papers and 12 were base documents; of the remainder only 15*had been selected by the students as possibly relevant, leaving a balance of 198 further documents to be submitted to the authors. The acceptance rate of these was 60.1%, a little lower than the acceptance of the students' documents, and only a single document of the 110 accepted was graded relevance (1).

In assessing all the additional relevant documents submitted, the authors did not know which had been selected by the students and which were retrieved by bibliographic coupling. The small variations in the acceptance rate (see final column of Table 3.4) by the authors for the different categories are so slight that they are not statistically significant. However there is significant difference in the proportion of documents put into the various relevance grades. From Table 3.5, it can be seen that with the cited papers 41% were included in grades (1) and (2); of the additional relevant papers found by the students only 18% were put in those grades and 15% of these revealed by citation indexing. The fact that so many of these additional references were placed in relevance grades (3) and (4), may be due to the fact that the authors did in fact know of the existence of many of those additional papers, but had selected the cited ones as being the most relevant to include in this paper.

So far the figures have been derived from the total set of 279 questions, but, as previously stated the questions fall into two groups. The authors had been asked to give the one basic question that gave rise to their work, and then to give any supplementary questions that came up during the progress of the work. Of the 279 questions, 118 are basic, and 161 supplementary. In order to discover whether the authors' assessments of their basic questions were in any way different to the supplementary questions, the same figures from Tables 3.4 and 3.5 are set out again in Tables 3.6 and 3.7, now divided into the two categories of questions.

*The 15 documents which were both selected by the students and retrieved by bibliographic coupling might be expected to have a higher acceptance rate by the authors, but

Table 3.6 shows a considerable difference between the basic and supplementary questions. 72.2% of all documents submitted to the basic questions were accepted as relevant, but for the supplementary questions acceptance was 57.6%. Such a difference might be expected in the case of the cited documents, since more of the references in an author's paper are likely to be included as relevant to the basic question, but the difference in acceptance shows the same proportional difference in all the additional documents submitted as well, (see Table 3.6). A possible explanation of this is the probably different attitude of the authors regarding the basic and supplementary problems. In the case of the basic problem no one complete answer would be available, and any document that shed some light on the problem, even if only remotely, would be likely to be accepted. The supplementary problem had more often been solved satisfactorily some time previously, and the author would therefore want to accept only those documents which dealt with the problem in a way that met his particular requirements.

Individual relevance assessments, done by 182 different people, and with no personal interaction with the project staff, cannot be entirely consistent. However the assessments were made by experts in their subject, and represent the individual and personal needs of the people concerned - the situation in which every information retrieval system has to operate. The evidence appears to show that the assessments were carefully done, although the task was sometimes difficult; as one author said:-

> "Relevance assessment is not easy, but I have done the best I can. In the case of this subject matter, the literature is so extensive that the chances of a relative newcomer picking out what mattered would be very poor; much of what are, in this connection, significant details have not been published anyway; even more important perhaps is that only long association with such a subject, both academically and experimentally, can enable one to appreciate what is useful and to judge what is misleading, unreliable or definitely faulty."

The use of four relevance grades might appear to be too precise a distinction to be able to make in practice, but quite a number of the authors indicated '½' grades, i.e. (1-2), etc. For the testing stage we accepted these documents at the lower grade. The definitions of the grades was a problem to one author:-

> "Actually ... none of your definitions (1), (2), (3), (4), (5) fits my attitude toward the references. All of the references were of considerable interest to me because they showed me what people had done so far, how recently, and by what methods. None was useful in suggesting methods of tackling the problem. I already knew all of the mathematical procedures that had been used in the papers, and several that had not been employed. To a large extent, it was interesting to find how little had been done, and in some cases, how inadequately."

Another author suggested that papers containing new or original answers to a problem should have a separate grade, and several authors indicated that a given document was a complete answer to their question, but an incorrect one. One new idea for assessing relevance was suggested:-

> "... the 'assessment of relevance' categories seemed particularly difficult to interpret in relation to most of these additional documents. I believe that I have 'scored' the documents roughly in proportion to the degree of irritation I should feel if a librarian produced them in response to my original query. Whether this is a proper basis for measurement of relevance may be arguable!"

The relevance assessments that the authors made of their own cited papers reveal some information on the citation habits of authors, but any observations can only be made within the limits of this situation, in which in most cases only a selection of the cited papers was used.

A few of the authors assessed all their cited papers as not relevant to the basic questions, and one explicitly stated that he did not find any relevant at all. An analysis of 174 of the basic questions, more than was ultimately used, shows that 36% of the cited papers submitted were assessed as not relevant, and if marginally relevant papers graded (4) are included, the figure is 52%. The results from the 118 basic questions in Table 3.6 give results of 28% and 46% respectively. It may be concluded that about half the references in an author's paper are not included in connection with the main problem of the paper, a fact which may assist examination of the possibilities, and limitations, of bibliographic coupling and citation indexing.

There were some cases where a cited document was not strictly relevant to any of the search questions at all, as one author honestly explained:-

> "I have had some difficulty in classifying some of my references into the required categories: chiefly those which occur at the beginning of the report when I attempt to relate this report to my own previous work. It is difficult to know whether they should be categorised as 3, 4, or 5: from the librarian's point of view they should probably be in category 5, but it is not easy to admit that several of one's references are, strictly, irrelevant to all the questions discussed."

Another good explanation for this case was:-

> "In the particular paper of mine a number of references are included, not to give information on the basic search question, nor do they arise from any subsidiary questions; rather they are included to amplify certain details in the text. For example the first three references of my paper are included purely to save time and words in the report, as I felt it completely unnecessary to describe experimental equipment which had been described fully elsewhere. Thus the first three references merit a 'five' rating."

One author supplied us with his reasons for inclusion of six of his references.

> "My assessments of reference 3, 6 and 9 refer really to many papers of which these are typical examples; No. 8 was not located - it just happened to turn up at the right time; No. 4 did not come to hand until after the work was completed and the report nearly so; No. 11 was included merely in order to satisfy anyone who wanted a long list."

A separate investigation, to extract similar information more thoroughly, might be of value, particularly if the author supplied reasons why each paper was, or was not, relevant. Comments on relevance itself, in the match between the questions and the documents, is made later.

The authors' assessments of the additionally submitted documents might be expected to have suffered a little in reliability, due to the time lag between the first letter and the second, and due to the additional documents being supplied as abstracts only. However some authors would be expected to have been aware of some of the additional documents, and, having the full bibliographical details, could examine the full text if they wanted to. Of the 201 questions for which additional documents were submitted, 39 were returned with all the additional documents assessed as not relevant, leaving 162 questions which had one or more of the documents relevant. Several authors indicated a continuing interest in the problem of their own paper, and the quick response to the second questionaire may indicate that the time lag was not a problem.

The large and difficult task undertaken by the students must next be examined. Some error would be expected of any job like this, and two pieces of evidence may indicate the magnitude of the documents missed.

1. Of the 198 documents found only by Bibliographic Coupling, 119 were assessed as relevant by the authors, (see Tables 3.4 and 3.5). There was only one graded as relevance (1), and the majority were graded (3).

2. In cases where an author had given more than one question that we were using, and also where we submitted additional relevant documents in relation to more than one of the questions, all documents submitted were listed together on a sheet with an indication given against each document of the question to which that document was judged to be relevant (see Appendix 3.2). However, there were cases when an author considered that a document which had been submitted in relation to one of the questions only was also relevant to another of his questions. This occurred in 32 questions, and involved a total of 75 documents.

This last fact means that the figures in Table 3.4 referring to the additional student assessed papers include these 75 documents, and the corrected figure for documents selected by the students is 842. Of these, 517 were accepted as relevant, giving an acceptance rate of 61.4% as against the previous figure of 64.6%.

Together with the Bibliographic Coupling documents that were accepted, a total of 194 relevant documents were missed by the students, which means that they found 517 of the 711 that were assessed as relevant, i.e. 73%. Reasons for failing to find the known loss of 27% may be:-

1. The students' interpretation of the question was more strict than that of the author, resulting in the students rejecting what the authors may have accepted.

2. The enormity of the task and inevitable occurrence of human error.

We may hypothesise that if the students' interpretation of the question had been more liberal, a large number of possibly relevant documents would have been selected, resulting in a difficult task of assessment for the authors, and thereby perhaps resulting in a much lower acceptance rate. Had more been submitted, more would probably have been accepted, but absolute perfection could not be achieved unless each author examined every document in the collection himself. The relevance assessments in relation to each question are given in Appendix 3G.

The questions

The authors apparently found no difficulty in preparing the search questions, and the number received was greater than expected, with each author supplying an average of $3\frac{1}{2}$ questions. Space was provided on the form for four questions, and of the 182 authors who replied, 120 supplied four questions. 40 supplied three questions,18 supplied two questions, and 4 authors only submitted the basic question. The high average, together with the fact that two-thirds of the authors supplied four questions, suggests that some authors could have written more questions, if space had been provided. However, since in practically every case all of the cited papers submitted were assessed as relevant to one of the questions given, so implying that none of the references was included specifically to answer a question which they had not supplied because of lack of space, it is reasonable to assume that four questions represented a near maximum for these authors.

The requested distinction between basic and supplementary questions clearly fitted the authors' view of their different problems, and only in one case did an author indicate that two of his questions were equally concerned with his basic problem. The 279 questions finally available for testing comprised 118 basic and 161 supplementary questions. There appeared to be no fundamental difference between the basic and supplementary questions. The set of questions is given in Appendix 3D.

The subject areas of the base documents were high speed aerodynamics and aircraft structures. The questions mainly fall into these two areas, but some of the supplementary questions in particular concerned subjects away from the centre of the two subject fields chosen. In aerodynamics, some questions dealt with chemistry of gases, sonic boom, flow in compressors, stability and control, spaceflight re-entry, and heat conduction. The structures questions mostly involved thermal and mechanical deformation and loading, with a few on vibration, effects of noise, and material properties. Some questions involved both subject areas, namely on aeroelasticity and flutter, while there were also some purely mathematical requests.

The generality of search questions is largely a matter of degree, but we would say, in the context of an aeronautical research organization, that most of the questions are reasonably precise, asking for a clearly defined part of the subject. There are a few broader questions (e.g. Q.41 "What progress has been made in research on unsteady aerodynamics"): there was one question which was not used in the tests because we considered it might have a hundred relevant documents, and would probably have retrieved the whole collection.

As previously stated, 279 questions were available for searching. Of these, 58 were really two or more questions stated in one, since they had a logical sum relationship. (e.g. Q.129 "What experimental measurements exist of spanwise and chordwise loadings on swept wings at low subsonic speeds and small incidence") For this reason, most of the tests were made with the remaining 221 questions, although at later stages in the tests,various subsets of thirty to forty questions were used for various purposes. The composition of these various groups of questions is given in Appendix 3E). Questions varied in length; the search terms ranged from 2 to 15 and the average number of individual search terms in the 221 most used questions was 7.6, median 7.9, and the mode was 7. These figures were obtained at the stage when the search programmes included every possible word, and a more conventional library search would be made on fewer terms than this, an average of probably 4 to 5.

It is always difficult to prove that any set of questions is really typical, or average in some way, but since each of these questions is a statement of a real need for information that arose in the course of some 180 research projects, they are probably as typical a set as can be obtained outside a real life situation. Many of the questions may have been put to an information service at some stage.

Without the facility to cross-examine the questioner, interpretation of the meaning gave less trouble than expected. A deep knowledge of the subjects would probably have revealed some facts and connections not appreciated, but many replies to the second questionaire included additional search terms suggested by the authors, and in some cases alternative rephrased questions. An example of the intricacies of the subject is seen in the following comment, made by an author to explain why one of the additionally submitted documents was not relevant to his question:-

> "It might seem strange that the paper by Kuchemann and Kettle would be of no use at all in answering my question. This is due to the fact that the influence of end plates is different for streamlined and unstreamlined bodies. In the first case they modify the vortices shed from the tips whereas in the second case they prevent spanwise flow brought about by the blockage of the body. There is no connection between these two effects."

The test design has produced a set of documents which have been assessed as relevant to a set of questions. Since this has not been done in a real life situation, can it be argued that the questions are artificial and the match with the documents unreal?

Considerable discussion and argument on these points has taken place in connection with the questions used in Cranfield I and the Western Reserve University test. Although the present question-gathering method did involve a base or 'source' document, it has not been used in the same way as in the previous tests. Previously the questions were framed so that the source document would be a complete answer to the question, but in the present test the question is the real need or research problem that gave rise to the 'source' document being written. Although the 'source' documents are included in the collection, it is only the cited documents from each 'source' document that are assessed and counted as relevant, with the addition of the extra relevant documents found. The 'source' document for each question is removed from the collection when that question is being tested and does not appear in any of the results at all. There is therefore, no reason for continuing to argue about the unreality of tests based on source document questions, or to continue to imply that the 'Cranfield test method' necessarily involves the use of such questions. However, we have stated a belief that source document questions 'can still be used satisfactorily in situations where time and cost are important considerations, as might be the case in an evaluation of a small operational information retrieval system'.

This comment was given in a reply to an article by D. R. Swanson, on 'The Evidence Underlying the Cranfield Results' (Ref. 4), in which he emphasised what he called 'the artificial' or 'biased' nature of the relationship of the question to the source document', in Cranfield I and the W.R.U. Tests. Swanson, in a sample taken from the first project, demonstrated that this biased relationship was shown by an unusually close match between the words of the question and the titles of the relevant documents. In his paper, Swanson gives the result of an analysis of the terms used in a set of 100 questions and the titles of their accompanying source documents. This was done by the Cranfield group and discussed at some length on pages 27-32 of Ref. 2, although Swanson does not comment on this work. Instead he prepared an admittedly more exact method, which would give, according to his view, retrieval of the source document and the number of irrelevant documents also retrieved would be small. To do this, he took the 100 document titles given in Appendix 4B, and made a list of all the terms which did not occur more than once. From this he argued that, if such a term also occurs in the matching question, then the document would be retrieved, with an average of 60 other documents also being retrieved. This statement is incorrect, in that Swanson bases it on the view that there were only 6,000 documents in the index searched, whereas there were 18,000, so a search of the nature proposed might be expected to retrieve an average of 180 documents.

However, using this method, Swanson finds a close correlation between the result of his 100 searches and the actual search results, and goes on to imply that the use of questions based on source documents will give predictable results.

To find whether these results could be repeated, we carried out the same procedure with the 114 questions and source documents of the W.R.U. test, as given in Appendices 2a and 2b of Ref. 3. This procedure gave 232 terms, of which 132 occurred only once. The result of this analysis was to show that 38 documents would have been retrieved by the use of a key term occurring not more than once, this representing a recall ratio of 33%, as against the 85% recall achieved by the Cranfield facet index. On the other hand, assuming that each key term occurring once in 114 documents would occur on an average of nine times in the whole collection, this method would have given a maximum precision ratio of 11% as against 16% achieved by Cranfield. Such a precision ratio of 11% could, of course, only be achieved by the hindsight of selecting the correct term and no other. For instance, Q.107 'Effects of increasing molybdenum content by carburising steels' is counted as a success by the fact that 'carburising' occurs in both question and document title. However, 'molybdenum' meets the single-use requirement, so would have retrieved the source document for Q.21, which would have been completely non-relevant. This effect would probably reduce the relevance ratio to less than 5%, but even so, the performance obtained by this method is vastly inferior to the performance obtained by the Cranfield index, and appears to make untenable the criticisms of Swanson.

There would appear to be three possible reasons for the difference in results of the similar tests done by Swanson and at Cranfield. Firstly, the W.R.U. collection was narrower in subject coverage than the collection of the first Aslib-Cranfield project. For instance, one key word given by Swanson is 'Titanium'. Since only some 300 documents in the whole collection dealt with metallurgical subjects, such a term is clearly unlikely to occur more than once in a hundred documents, whereas in the W.R.U. count it occurred on eight occasions. (This is an aspect of the generality ratio discussed later)

A second reason could be a significant difference in the quality of titles. Many documents in the first Aslib-Cranfield test were research reports, with titles which were fuller than usually occur in commercial journals, from which many documents were taken for the W.R.U. test.

Total documents retrieved by bibliographic coupling at strength of 7 or more		312
Documents which had already been assessed for relevance by being references.	87	
Base documents	12	
Documents which had been located by students	15	114
Submitted to authors for relevance assessment		198

Table 3.8 BREAKDOWN OF 312 DOCUMENTS RETRIEVED BY BIBLIOGRAPHIC COUPLING AT STRENGTH OF 7 OR MORE.

QUESTION 145

Has anyone investigated the unsteady lift distributions on finite wings in subsonic flow

RELEVANT DOCUMENTS

1698. The unsteady lift of a wing of finite aspect ratio, (STRONG MATCH)

1705. On the kernel function of the integral equation relating the lift and downwash distributions of oscillating finite wings in subsonic flow. (STRONG MATCH)

1704. A systematic kernel function procedure for determining aerodynamic forces on oscillating or steady finite wings at subsonic speeds. (WEAK MATCH, because 'finite wings' and 'subsonic' are commonly used terms in this collection)

1700. Two and three dimensional unsteady lift problems in high speed flight. (WEAK MATCH)

1703. General airfoil theory. (NO MATCH)

1792. Some low speed problems of high speed aircraft. (NO MATCH)

TABLE 3.9 EXAMPLES OF QUESTION/TITLE MATCHES FOR RELEVANT DOCUMENTS

Terms underlined in the document titles are those matching the required terms in the question.

The third, and probably most significant reason was the greater care taken with the questions for the W.R.U. test. There appears to be no reason to apologise for the fact that it was not possible to exercise such close control over the question compilers when we had to obtain some 1,600 questions for Cranfield I, but by the time of the W.R.U. test, the importance of the matter had been accepted, and the question compilers were personally selected and more adequately instructed.

In the W.R.U. test, an analysis was made of all documents in the collection against each question and, as given in Appendix 3C of Ref. 3, 42 other documents were assessed as equally relevant as the source documents. As a further check on source document questions, the titles of these documents have also been matched against the appropriate questions, using the list of terms generated with the original 114 source documents. Fourteen documents had a single term match with the questions, so again the recall ratio was 33%, the same as with the source documents. This appears to show fairly conclusively that, in the W.R.U. test, there was no unnatural relationship between the terminology of questions and source document titles, and lends support to the strongly-held view of the Aslib-Cranfield staff that questions based on source documents can still be considered as being, in the right circumstances, a convenient and economic device for testing I.R. systems.

Some unnatural relationship was clearly present in Cranfield I, but it is wrong to conclude from this that whenever there is a substantial match between question and title, then the relationship is necessarily unnatural. Some proportion of questions in a real life situation are bound to have some relevant documents with a close question title match, and if this is not the case then all Permuted Title or K.W.I.C. indexes are useless. However, although as explained earlier, source-document questions are not used in the present test, Swanson still expresses doubt and comments on the present test method:- 'This is some improvement (since the title-question correlation is probably diminished); but it is still dubious in principle - a 'biased' or 'special' relationship between questions and relevant articles persists' (ref. 4). Although no evidence is presented to justify this statement, an examination of some of the questions and their relevant documents has been made, to find out the extent, if it exists, of the bias of the suggested relationship.

Using 35 of the questions*, and their associated 287 relevant documents, we first examined the correlation between the questions and document titles. The words and phrases of the questions were examined for a 'match' with the words and phrases in the titles, and generally an identical word or phrase only was considered as a match, except that synonymous word ending variants were accepted. In terms of the whole question, two levels of matching were distinguished:-

Level A Strong Match Two or more concepts, or important subject words were demanded. A single concept was only accepted if it was one of the vital ones in the question, and in a few cases a single word was accepted as a vital or 'key' term provided it was used less than twenty times in indexing.

Level B Weak Match These rules accepted any match down to a single word, provided it was a subject content word. The general descriptive words such as Problem, System, Solution, Parameters, High, Large, etc. were not accepted.

*These questions are the 7 search-term questions and appear as Question Set 1 in the Appendices.

Strength of match	Relevance grades				
	(1)	(2)	(3)	(4)	Totals, all relevant
Strong match	12	17	40	20	89
Weak match	3	20	39	25	87
No match	4	24	54	29	111
(Total)	19	61	133	74	287
Percent strong match	63.2%	27.9%	30.1%	27.0%	31.0%
Percent strong and weak combined match	78.9%	60.7%	59.4%	60.8%	61.3%

TABLE 3.10. RELEVANCE GRADES OF DOCUMENTS
WITH SPECIFIED QUESTION-TITLE MATCH

Strength of match	Cited documents	Additional documents	All documents
Strong match	44	45	89
Weak match	38	49	87
No match	67	44	111
(Total)	149	138	287
Percent strong match	29.5%	32.6%	31.0%
Percent strong and weak match combined	55.0%	68.1%	61.3%

TABLE 3.11. COMPARISON OF THE CITED AND ADDITIONAL
DOCUMENTS WITH SPECIFIED QUESTION-TITLE MATCH

Some examples of a strong match are: Chemical, Kinetic (Question 9); Viscous, flat plate (Q.82); and Slip (frequency of 19, Q.87). Some examples of a weak match are: High Speed (Q.2); Aircraft (Q.2); Hypersonic (Q.9); and Structural (Q.49). Further examples can be seen by reference to Table 3.9.

Out of the 287 documents examined against the 35 questions, 89 (31%) showed a strong match; an additional 87 had a weak match, and the total of 176 represents 61.3% matching. 28 of the questions had one or more documents with a strong match, and 32 had one or more with a weak match.

This shows that nearly one-third do have a strong question-title match, but since the assessment of relevance has been done in four grades, we may expect that those documents with a strong match will be graded as more relevant than those with a weak match. Table 3.10 divides the results into the four relevance grades, and shows that the probability of a relevance (1) document being strongly matched is more than twice that of the relevance (2), (3) or (4) documents. That the relevance (2),(3) and (4) documents show the same probability may be accounted for by the difficulty of consistently doing such a refined grading of relevance, but the relevance (1) documents seem to indicate a strong trend.

Whether it is taken that these figures show an unusual question-title match or not, the presence of an unnatural question-document relationship cannot be proved or disproved by this. One would have expected a certain strength of title match in this subject, where titles are usually fairly long and a good indication of the subject of the document. The documents examined were the total of those relevant to each question, and included both the original documents cited in the authors' base document, and also the additional documents discovered in the collection. It is obvious that these additional relevant documents, discovered by the students' examination of the collection and by bibliographic coupling, were discovered and assessed as relevant in a situation equivalent to a real life one, and therefore it would be quite absurd to suggest that an unnatural or biased relationship could possibly exist in their case. So a comparison of the question-title match between the 'cited' relevant documents and the 'additional' relevant documents will provide some evidence of any unnatural differences in the question-document relationships.

The 287 relevant documents comprised 149 cited and 138 additional, and the matching scores were calculated for each group. Table 3.11 presents the results, and it is shown that the additional relevant documents had a slightly stronger question-title match than the cited ones, 32.6% to 29.5% for the strong matches, and 68.1% to 55.0% for the weak matches. Ten of the 35 questions had no additional documents at all, and the cited document for these questions have been included in the results; deleting these ten questions would reduce the matches for cited documents to 27.6% and 51.4%.

These results might alter over the whole set of questions, but there is no reason to expect that they would change significantly. On the basis of the question-title match anyway, no real difference exists between the cited and additional documents. We suggest that this indicates that there is no justification for any implication that there is a biased or unnatural question-document relationship, and that the relevance assessments and relevant documents found are not really different from that which might happen in a real life situation. Further evidence can be obtained from some of the test results themselves, where the retrieval performance in recall of the cited documents can be compared with the additional documents.

Co-ordination Level	Cited Documents		Additional Documents	
	Total Recalled	Recall Ratio	Total Recalled	Recall Ratio
1	99	94.3%	128	92.8%
2	80	76.2%	101	73.2%
3	59	56.2%	75	54.3%
4	40	38.1%	46	33.3%
5	17	16.2%	25	18.1%
6	9	8.6%	10	7.2%
7	2	1.9%	3	2.2%
Total Relevant	105		138	

TABLE 3.12 COMPARISON OF RECALL PERFORMANCE OF RELEVANT 'CITED' AND ADDITIONAL DOCUMENTS IN RELATION TO 25 QUESTIONS.

All questions had seven starting terms; the table shows the effect on recall of increasing the search requirements from any one term to all seven terms.

Using the same set as considered in the previous paragraphs, the 25 questions which had some additional relevant documents were used, comparing 105 cited with 138 additional documents. Here again the difference between the two groups is not significant, (see Table 3.12). For instance, at a coordination level of 2, the recall ratios are 76% and 73% for cited and additional documents; at a coordination level of 5, the figures are 16% and 18%. These results (which are, of course, only a small sample of what will be presented in a later report) should have revealed any unnatural question-document bias, whether conspicuous in the title or not, had any bias been present at all. We are confident that there is no measurable unnatural match between the questions and the documents themselves. Questions obtained from a real life situation and tested on an existing collection might give different results in some way, but until such a test is done, and a comparison is made of different test methodologies, it is not possible to state in what ways, and by how much, the present test method falls short of the ideal in this respect.