Chapter 2

TEST DESIGN

There has been a considerable amount of comment during the past few years about test design in general and the test design for Cranfield I in particular. That much of this has been, unfortunately, misinformed has been due both to a failure to appreciate the basic problems and purposes of an evaluation test, and also to a failure to distinguish between two main types of testing.

The first type of testing is that which is concerned with the evaluation of an operational information retrieval system, a sub-system of an operational system or a system or sub-system proposed for an operational system. In all such cases, there is no basic intention of advancing knowledge concerning information retrieval systems in general, although in the present state of fragmentary knowledge, this may well be a by-product. Basically such a test is designed to provide data for an analysis to be made which will show how the system can work more efficiently either in regard to operational or economic factors, in supplying the particular requirements of a given body of users. Such a test was that performed by Lancaster on the index of the Bureau of Ships (reference 5). Well designed on the basic Cranfield test procedure, with defined limited objectives, it produced, economically and quickly, data which enabled decisions to be taken on the optimum methods for the information retrieval system at the Bureau of Ships. As a 'research' pay-off, it revealed yet another situation where the use of roles was economically inefficient and operationally of doubtful value, and added to the growing body of data on the problems created by the use of roles of the type proposed by the Engineers Joint Council, in the Thesaurus of Engineering Terms.

There are many different variations of this type of test situation. One can, for instance, devise a new system or sub-system and test it while it is still comparatively small as effectively as one can test the performance of a long-established operational system, but the characteristic of all such tests is that they are made with a given situation in mind, their parameters are fixed by the pre-determined environment of the system being evaluated.

The second type of test - the type with which this report is concerned - is where one is dealing with an experimental situation. In such a case, the purpose of the test is to advance knowledge in some aspect of information retrieval without any particular operational requirement in mind. For this to be done, it is necessary to advance from a firm foundation of what is known. To make such an advance may require the use of unproved techniques, and, since the attempt is being made to investigate the unknown, there is always the possibility that, however meticulously the test has been designed, some unexpected factor will interfere with the objective of the test. If such a factor can be recognised early enough, it may be possible to adjust the design to take account of the new situation, but the risk has to be accepted that the weakness may only become apparent towards the end of the test.

A classical example of such a situation was the test carried out by Documentation Inc. Inc., where the objective was to compare the performance of a Uniterm index and the alphabetical subject catalogue compiled by the Armed Services Technical Information Agency. The first stage of the test involved the indexing of 15,000 documents by the Uniterm system, at the same time as they were also being indexed by the ASTIA staff. The second stage was for the two groups to carry out searches in their indexes for some ninety odd questions and then for each group to analyse the output of their searches to find which documents were relevant. Up to this point, everything appears to have gone according to plan. The final stage was intended to be a comparison of the output of the two sets of searches, in order to find which system had been successful in obtaining more relevant documents.

The problem which arose at this final stage was that neither group was willing to accept the relevance assessments of the other group; rumour has it that at the end of the second day of discussion, the two groups were still arguing about the meaning of the first search question. No real blame can be fixed on those who organised the test; in 1952 it was not unreasonable to think that two groups of intelligent people would, without serious difficulty, be able to come to an amicable agreement as to which documents were relevant to a particular question. If any fault can be found, it only lies in the failure to make generally available either of the two reports which are said to have been prepared by the two groups taking part in the test. The only published account was a brief paper by Gull which appeared some years later in American Documentation (reference 10), and which dealt mainly with the results of the searches. Gull does, however, make the following very apt comment: "When one considers that a fairly thorough search of the literature indicates that this comparison of two reference systems is the first undertaken so far, it is not surprising that the results revealed clerical errors and an incomplete design of the test."

With the exception of a small test done in 1953 by Cleverdon and Thorne (ref. 11), this had been the only test of an I. R. system carried out before the test design for Cranfield I was prepared in 1956. While access to the complete reports of the ASTIA-Uniterm test might have revealed some more information, the only positive fact known in 1956 concerning test design of I. R. systems was that failure to have a firm agreement on question-document relevance could result in complete failure to realise the test objectives. Concerning information retrieval systems, however, nothing was known for certain. For any belief categorically stated by one expert, it was possible to find the exact opposite stated by another expert. Those were, in fact, the halycon days when one could argue all night without producing a shred of evidence for one's views, when Metcalfe, for instance, could write a fascinating book (ref. 12) proving in three hundred pages that an alphabetical subject catalogue was vastly superior to a classified catalogue without having to, or being able to, present one piece of experimental data to support any of his many assertions.

The field of investigation for Cranfield I was therefore wide open, in the sense that it would prove or disprove some conflicting beliefs. Since it was uncertain as to what was of major importance, the decision was deliberately taken to plan the test over a wide range of aspects. Not only index languages but qualifications of indexers, indexing time, categories of documents, search tactics and search capability, optimistically (over-optimistically some might argue) all were incorporated in the test design. Any knowledge would be new knowledge and there was practically no limit to what could be attempted, although there were certainly definite but unknown limits as to what could be achieved. From a personal viewpoint, however, one limitation was essential in the design; actual questions could not be used if these involved relevance assessments by other people than the questioners. This restriction had to be accepted, and the result was the adoption of the technique of using prepared questions based on source documents. Although this technique has been strongly attacked in many papers, no-one has suggested any other method which would have permitted so much reliable data to be obtained so economically.* However, by the time the design of the present project was being considered, the position had changed radically. The conclusions coming from Cranfield I, supported by other smaller investigations, had deliniated more sharply the problem areas for investigation; equally important was the realization that progress would be dependent on the use of more refined test methodology.

As outlined in the previous chapters, the new project was to deal with index language devices; the first objective was the precise measurement of recall and precision ratios. The essential prerequisite to obtaining these measures (in an experimental situation) is the determination of the sets of documents which are and are not relevant to each of a set of test questions. Before proceeding to discuss the various ways of determining this matter, it may be helpful to consider a recent paper by the late Dr. Taube 'The pseudo-mathematics of relevance' (ref.13), which is being widely quoted as discrediting the results of the Cranfield investigations.

Any paper by Dr. Taube merited serious consideration, and in particular any paper dealing with the question of relevance, since this was the critical problem in the original test carried out by Documentation Inc. While the paper presents what at first sight appears to be a plausible argument, it is, in fact, based upon a confusion and distortion of meaning of two uses of the term 'relevance'. First there is the use of the term on its own where it denotes, in a true life situation, the subjective assessment of an individual in relation to a document or a set of documents which he receives in answer to a search question, so that he says "these documents are relevant to my questions, those other documents are not relevant". The second use of the term is in 'relevance ratio', which is the manner of expressing the proportion of relevant documents retrieved to the total of documents retrieved in a search. As such, 'relevance ratio' has nothing to do with the determination of relevance, but merely involves a numerical calculation of those documents which have been previously allocated to one of the two sets of relevant and not relevant.

At a meeting in Washington in 1964 of a group of some thirty people concerned, to a greater or lesser degree, with evaluation of I.R. systems, the paper in question, (which was originally written in March 1964) was amongst the documents circulated. Since it was clear from the discussion that Dr. Taube was still confusing the two meanings, Cleverdon agreed that in future we would cease to use the term 'relevance ratio' and substitute another term. Possible alternatives were 'acceptance rate' or 'precision ratio', both of which were being used by other groups with the same meaning as 'relevance ratio'. As stated earlier, 'precision ratio' was selected, and if one substitutes this term in those cases where Taube

*In these days when large grants are common for small investigations, it is of interest to recall that the five years' work of Cranfield I, including the test of the Metallurgical Index of Western Reserve University, was covered by two grants from the National Science Foundation, totalling $44,000.

Bibliography #: 453 Title: USE OF FLUORINE FOR ROCKET PROPULSION

A. Terms/Hits

a. Total Search Terms 37

b. Maximum Hits Possible 906

c. Anticipated Hits 500

B. Most Heavily Posted Terms:

	Terms	Postings
1.	PROPELLANT	2830
2.	FUEL	1765
3.	OXIDIZER	384
4.	FLUORINE	300
5.	FLUORIDE	259

C. Type of Logical Equation Specified:

____ a. Loose. High output. Irrelevant material expected.
____ b. Moderately loose. Some irrelevant material.
X c. Moderately tight. Very little irrelevant material.
____ d. Tight. No irrelevant material expected.
____ e. Analog. Analog measure: ____

D. Initial Search Results:

a. Hits (Total Output = T) 441
b. Accepted Hits After Editing (Accepted Accessions = A) 379
c. Acceptance Ratio, A/T x 100 = 86.5 %

E. Auxiliary Search Results:

a. Hits (T') 10
b. Accepted Hits (A') 10

F. Reject Analysis:

a. Rejects on Initial Search 63
b. Rejects Attributed to Type of Equation, i.e., out-of-scope or marginal upon examination 62
c. Rejects Attributed to "Noise", "False Drops", etc. ____
d. Other Rejects, e.g., Indexing Errors 1
e. Total Rejects Considered
Excessively high ___ High ___ Average ___ Low X

G. Miss Analysis:

a. Misses detected, overall 2
b. How were misses detected? STAR CUMULATIVE INDEXES — PUBLISHED SUBJECT INDEXING

H. Analyst's Comments and Recommendations (i.e., search strategy, reject analysis, new terms, delete and transfer suggestions, indexing errors, etc.):

1. THE 62 REJECTS ALL DEALT WITH FLUORINE COMPOUNDS AND FLUOROCARBONS, BUT WITHOUT SUFFICIENT RELATIONSHIP TO THEIR USE IN PROPELLANTS.
2. AUXILIARY SEARCH: CHLORINE TRIFLUORIDE AND OXYGEN DIFLUORIDE (OXYFLUORIDE) WERE SEARCHED INITIALLY UNDER THEIR "PRE-COORDINATED" NAMES. THE AUXILIARY SEARCH UTILIZED (1) CHLORINE INTERSECT FLUORINE UNION FLUORIDE, AND (2) OXYGEN INTERSECT FLUORINE UNION FLUORIDE. INDEXERS SHOULD BE ENCOURAGED TO USE "PRE-COORDINATED" PROPELLANT NAMES.
3. TERM "OXYGEN DIFLUORIDE" SHOULD BE USED IN LIEU OF "OXYFLUORIDE".
4. N62-10209: DELETE POSTING UNDER "FUEL"
5. N63-10406, N64-11223: POST UNDER "PROPELLANT".

FIGURE 2.1 NASA SEARCH SYSTEM ANALYSIS SHEET

used the term 'relevance' with this meaning, it is immediately apparent that the whole argument is defective. The argument in the paper starts with a quotation from a Cranfield paper written before this decision to change to the term 'precision ratio' had been taken. Substituting this term, but not in any way changing the original meaning, we would now have written, "With the aid of the set of documents and the set of questions [for which the document/question relevance assessments have been previously made by the questioner] it will be possible to test each index language device in turn and so get precise figures for the effect on recall and precision ratios."

Taube's comment on this was "some way or another a vague or hardly recognisable and admittedly difficult notion [i.e. relevance] has turned out to be precisely measurable". It is not, of course, relevance which is being measured, but the decisions regarding relevance which have already been taken. As Salton says (Ref. 14), "once acceptable relevance judgements are available for all documents with respect to all search requests, the calculation of recall and precision becomes perfectly straightforward and unambiguous."

It is interesting to find, in the issue of American Documentation for April 1965, that there is a brief note (ref. 15) by two members of the staff of Documentation Inc., in which they discuss a NASA Search System Analysis Sheet. The example which they presented has been reproduced on page 12, and from this it can be seen that these members of the staff of Documentation Inc. have been able to derive, for this particular search, an acceptance rate (i.e. precision ratio or relevance ratio) of 86.5%.* It is interesting to note that, on the Analysis Sheet, the phrase used is 'accepted hits after editing'. This implies that the determination of the relevance of the document to the question has been by a member of the staff of Documentation Inc., and his standard for relevance might be very different from that of the questioner. This leads us back to the point we had reached before the diversion to consider briefly the matter of relevance. As we argued earlier, there were sound, compelling reasons for the use of source-document questions in Cranfield I, because they gave, simply and economically unequivocal relevance assessments. More particularly, it still remains probably the most effective and economical method of establishing the general recall ratio in many test situations. By 1961, however, it was quite unacceptable for an experimental investigation of the type we had in mind. What were the alternatives? These can most simply be tabulated under various aspects as follows.

Types of search questions

1. An actual question that is put to an information retrieval system and searched at the time it is required.

2. An actual question that has been put to an I.R. system. In other words, one obtains questions that have been used previously, either with the system being tested or some other system.

*To save misunderstanding, we would point out that an error has been made in calculating this figure. It should, of course, be 85.9%.

3. A prepared question, that is a question which has been composed specifically for the purpose of the test and is not a question which meets an actual need of the questioner. Such prepared questions may or may not be based on a particular document or documents.

Method of Relevance Assessment

I By the questioner
II By the consensus of opinion of a group of people
III By an individual, not the questioner
IV By matching the indexing with the search programme.

Type of Individual(s) Involved

A User of a system
B Scientific or technical staff, not users of the system
C Librarians or other information staff.

If we now chart Type of Question against Method of Relevance Assessment, the various possibilities can be shown

Method of Relevance Assessment

	-	I	II	III	IV
Type of Question	1	A A	A BC	A BC	A -
	2	A A	A BC	A BC	A -
	3	ABC ABC	ABC ABC	ABC ABC	ABC -

In the chart, the upper half of each box represents the type of person asking the question, the lower half represents the type of person making the relevance assessment.

An additional variable concerns the type of document on which the reference decision is based, for this can be either

α The complete text
β An abstract
γ The title.

It can be seen that the Documentation Inc. example discussed above was, presumably, the use of an actual question (1A) where the relevance assessment was made by an individual, not the questioner (III) who was a member of the information staff (C), probably basing his decisions on document titles,making up the code (1A)(IIICγ). For Cranfield I the code would have been (3B)(IBα), which is to say that prepared questions were used (3), based on complete documents (α), this resulted in the relevance being determined by the questioner (I) and the individuals involved were technical staff not concerned with the system (B).

The theoretical ideal is (1A)(1Aα) that is the use of actual questions with a relevance assessment made at the time by the questioner from complete texts. This cannot be achieved in an experimental situation since there is no body of users who can ask questions, nor would the experimental collection normally be of sufficient size to justify actual searches. For this project, it was considered that the nearest to the ideal would be the combination (2B)(IBβ+), that is questions which had been asked, with a relevance assessment being made by the questioner who would be a scientist. (β+) implies that nothing less than abstracts would be used; the expectation would be that full texts would also be used. The wisdom and implications of this choice will be considered in relation to the test results. What can be stated here is that the operational performance characteristics of the system being tested will almost certainly change depending on the combination of questioner and relevance assessor used, and care should be taken in interpreting figures which do not define how they have been obtained in this respect. A few illustrations of what can happen may help to clear up this point. In the Documentation Inc. example previously quoted, the precision ratio of 86.5% is very high. A probable reasin is that it is based on the relevance assessment of a member of the information staff; when the set of documents is sent to the questioner, his relevance standards may be such that he will grade the large majority as non-relevant, so the relevance ratio would then drop considerably.

As another example, in a report of the evaluation of the EURATOM information retrieval system (ref. 16), a precision ratio of 65% is given. The key to this high figure is in the following sentence taken from the text of the paper. "Finally, the computer's answers have to be checked, since it would be unreasonable to expect them to be 100% complete and correct".

What has happened in this case is something rather different. The precision ratio is not being calculated on the actual search output but on the search output after technical information staff have rejected the documents which they considered non-relevant. A somewhat similar reason was the cause of some confusion at the NATO Advanced Study Institute on evaluation of information retrieval systems, when Altmann, in presenting the results of a test on the information retrieval system of the Harry Diamond Research Laboratories (ref. 17) gave figures of 80% for precision ratio. In this case, it appeared that the procedure was for the questioners, who were also making the searches, to eliminate documents which, from title or abstract, appeared to be non-relevant; this maybe gives interesting information about the ability of users to eliminate non-relevant information on the basis of the title but, as with the EURATOM test, gives no information at all on the performance of the system in regard to precision.

The discussion so far has been dealing with precision ratios; while there is still considerable doubt as to the most useful way, in an experimental situation, of obtaining relevance assessments, once that assessment has been made the determination of precision ratio is a straightforward matter. The same is not, however, true of recall ratio, because this is dependent on the number of relevant documents which have not been retrieved. This problem was effectively side-tracked in Cranfield I by the use of source-document questions; since this method had been ruled out for the present test, there was only one apparent alternative, namely to look at every document in relation to every question. This decision automatically placed a restriction on the size of the test collection and the number of questions to be searched. This was not considered a serious handicap, since the W.R.U. test had shown that a collection of only one thousand documents was sufficient to provide a considerable amount of data for analysis. There seemed to be some advantage in having a larger number of questions in relation to the number of documents in the collection than had previously been used, and the decision was to aim at 1,200 documents with 300 questions.

There was no readily available collection of questions which had actually been used on some previous occasion. Even if there had been, it would not have been possible to have the originators of the questions check the documents for relevance. The method adopted, therefore, to obtain the documents and the questions was to select a number of recently published research papers, mainly dealing with high speed aerodynamics, but about 20% of which covered aircraft structures. The author of each paper was to be requested to provide the basic problem, in the form of a search question, which had been the reason for the research being undertaken, and also to give some additional problems which had arisen in the course of his work. At the same time he would be asked to state which papers in his list of references were relevant to the various questions he had provided. It was intended that the document collection would be made up of the papers that had been included as references.

'Relevance' is obviously a matter of degree. The problem in arranging for relevance assessments to be made is to decide how many degrees of relevance can be consistently recognised. In the test of the index of Western Reserve University, two levels of relevance were used; previously , Swanson (ref. 18) had attempted ten levels. The decision in this test was to use four levels of relevance; details of this and the whole procedure of obtaining the questions and document collection are given in chapter 3.

The references in any given paper might be expected to give a high proportion of relevant documents to any question arising in connection with that paper, but at the same time there was the probability that other documents in the test collection would also be relevant. The author might have known about these documents but have decided not to use them. Alternatively, he might not have been aware of their existence; possibly they might have been published after he had finished his work. While it was essential that ther should be a complete cross-check of every document and of every question, it was impracticable to send 1,200 documents to each of 200 or so authors for them to make the assessments individually, so a screening process was first necessary. This was to be done by recruiting a number of postgraduate students who would (hopefully) be able to eliminate most of the non-relevant documents for each question. Then it would only be necessary to send to each author those papers which had a reasonable possibility of being relevant, for each author to make a final decision concerning relevance.

We would forestall criticism of the method outlined above, by admitting immediately that it includes nothing which overcomes the basic problems of the meaning and determination of relevance. No-one is more aware that relevance is a shifting notion, certainly between individuals and often for the same individuals at different times. Is there, then, justificiation for the comments by Taube that any attempt to measure system performance is useless, since such measurement must be based on relevance decisions. We would strongly argue against this, for it is the very situation which an information retrieval system has to face. Users do ask questions and then accept or reject the search output in what might seem an arbitrary manner. The objective of the methods used in this test was to get as near as is possible in an experimental test to a true life situation in relation to relevance decisions. While they certainly represented an advance on the methods in Cranfield I, it is not intended to suggest that the design was perfect; again it is necessary to go back to the time when the test was designed, and say that in 1961 it appeared to be the best techni que that could be adopted for the particular requirements. The experience of this test has shown not only its advantages, but also some disadvantages, and these are briefly discussed in Chapter 8.

So far the discussion on the test design has been entirely concerned with the methods to be used in obtaining a set of test documents and questions, and establishing the relationship between the documents and the questions. All such activity was an essential preliminary to the investigation itself, the general background of which was considered in the previous chapter. To summarize this briefly, we started from the belief that all index languages are amalgams of different kinds of devices. Such devices fall into the two groups of those which are intended to improve the recall ratio and those which are intended to improve the precision ratio. In other words, there are some devices which will always enlarge the class and thereby retrieve more documents, with the probable result that more relevant documents will be retrieved. On the other hand, the precision devices will always act in the reverse manner by narrowing the class, thereby retrieving fewer documents, with the probable result that some relevant documents will be eliminated. The purpose of the test was to investigate the effect which each of these devices, alone or in any possible combination, would have on recall and precision.

To enable this to be done, it was essential that it should be possible to hold everything constant except the one variable being investigated. The organization of the file, with its completed matrix of document/question relevance assessments, was the first step towards this. The next stage was to determine and fix, once and for all, the concept-indexing of the documents and the relationships of the concepts. By concept-indexing is meant the decision as to which concept and groups of concepts are significant from the viewpoint of retrieval. Such concept indexing can only be in the terminology of the document. As soon as there is any 'translation' of the document terminology to any kind of formalized language, then one of the index language devices must have been brought into use. Therefore the decision was to concept-index, at a high level of exhaustivity, the documents in the collection so that they might be translated into any type of index language which it was desired to test. Details as to how this was done are given in chapter 4.

The original proposal to the National Science Foundation contained the following statement. "At this stage it should be possible to decide which technique appeared to have the most satisfactory characteristic for adaptation to automatic indexing. Dr. J. O'Connor has explained the techniques which can be used to investigate methods of automatic indexing without actually using computers. (ref. 19). Our approach would be partly to investigate new techniques, but might as usefully be concerned with testing methods proposed by others and measuring the performance of such methods against the results from human indexing."

The possibility and the hope that the test collection could be used by other groups and provide direct comparison with the Cranfield results was partly responsible for the decisions concerning the indexing technique and also the searching method. This permitted starting from the absolute basic point of matching any actual word in the question with any term used in the concept-indexing and then to introduce all the devices by stages. It is agreeable to be able to record that it will be possible to compare the results of the Cranfield tests with two experiments using computers. In England, at the Cambridge Language Research Unit, the complete set of Cranfield indexing is being processed on the Atlas computer, and it will be possible to measure and compare the effectiveness of the 'clumping' process which Dr. Needham has been investigating(ref.29) In the United States, at the Harvard Computation Laboratory, a sub-set of the indexing has been processed by a number of the options of the SMART programme which Professor G. Salton has designed. There is particular interest in this work, in that, in addition to the searches based on the Cranfield indexing, searches have also been made on the abstracts taken from the documents. This work is discussed in more detail in Chapter 7.

It was, of course, known that decisions would have to be made concerning the physical methods which would be used for carrying out the searches. Fortunately, no firm decisions were taken on this point; the methods ultimately used are discussed in Chapter 6.

Finally, there would be the necessity to present the results in a meaningful manner. The recall/precision ratio figures and curves of Cranfield I have undoubtedly taken a hammering over the past few years, and there are many who have sought the elixir to change them into the pure gold of a single figure. Far from being able to do this, it was by 1961 clear to us that, if there was to be any comparison of experimental results, it was necessary first to investigate the effect on performance of the generality ratio, namely the relationship between the number of relevant documents and the size of the collection. The first tentative ideas on this had been put forward on page 101 of Ref. 2; in this project it was planned to attempt to measure the effect of this factor on recall and precision.