CHAPTER 7

TESTING OF EXISTING SYSTEMS

Whereas the main investigation had attempted to simulate as far as possible normal operating conditions, yet it had artificialities, some of which were inevitable, others which were imposed by the requirements of the programme. Even if in the indexing it had been practical to simulate exactly a normal operation, the results of the work would inevitably be subject to the limitations that are inherent in any single system, such as the operators of the system, the subject field, the purpose of the system, etc. There would, therefore, be at the conclusion of the investigation the dual requirements of attempting to find to what extent the results of the investigation had been influenced by the inherent artificialities and also to what extent the results could be considered applicable to other operating conditions of the type mentioned above.

From the commencement it had been hoped that it would eventually be found practicable to do these two separate tasks with only a small amount of the effort and expenditure involved in the present project. When the first set of analyses had been made (as discussed in Chapter 5), it seemed that it would be possible to make a test which, in addition to providing information on these points, would also be of some value to the system being tested.

At the request of the Director, the Library of the English Electric Co. Ltd., at Whetstone, kindly agreed to allow the Project staff to carry out a test on their catalogue. From our viewpoint this was to be an experiment designed to show whether such a technique of testing could produce valid results, and at the same time it was hoped that the co-operation of the library staff would to some extent be repaid by providing data which would assist in the operation of the catalogue.

On the completion of this test in June, 1961, a report was prepared for private circulation amongst the libraries of English Electric Group of companies, and the following account is largely taken from this report, which was prepared by Miss Warburton of the Project staff and Mrs. J. Aitchison, Librarian at Whetstone.

The English Electric Library at Whetstone, now the Central Library Service for the Group, developed a faceted classification scheme in 1958, the

third edition of which was issued in March, 1961. The system is used at the Whetstone and Bradford libraries of The English Electric Company and has since been adopted by newly established libraries at the Stafford and Liverpool Works. The facet scheme is also used in certain library publications, the most important of which is the monthly Reports Abstract Bulletin indexing English Electric reports. The catalogue at Whetstone, which was tested between January and April 1961, contained cards for approximately 36,000 documents.

The schedules were constructed according to facet principles as understood from the writings of members of the Classification Research Group. The subject field, the whole of engineering, was divided into the basic categories, plant and machines, components, materials, physical phenomena, operations and instruments, and within these categories terms were further analysed and grouped according to clearly defined characteristics. The schedules, like those of the Cranfield Facet Schedules, are short compared with enumerative schemes since only basic terms are listed, from which the classifier must himself synthesise class numbers for complex subjects.

The orthodox method of preferred order and chain index technique is used in the Whetstone Catalogues, but care is taken to ensure that the scatter of subsidiary terms inherent in preferred order does not render the catalogue unworkable as a speedy retrieval tool. The number of concepts used in combination are kept within limits and it is recognised that there should be consistency in omission and addition of terms during synthesis. In certain parts of the schedules there are rules disallowing certain combinations of terms so that scatter is reduced. Also, when adding new material, the classifiers check both chain index and classified sequence so that like documents are kept together and not separated by slight inconsistencies in combination of terms. As a result, the average number of terms used in combination is three as compared with the five used in the Facet catalogue at Cranfield. At the time of the test the catalogue contained 76,000 cards of which 48,000 cards (63%) were in the class sequence and 28,000 cards (37%) in the chain index. The average number of cards in the catalogue for each document was 2.05.

The first requirement in a test of this nature is to have relevant information concerning the indexing and it was fortunate that this was already available in that, as a matter of routine, records were maintained of the indexing of each

document. 200 references were chosen at random from those documents which had been indexed, and the selected documents were divided out amongst a number of technical staff at Whetstone. They were asked to compile questions of a kind which they might expect to put to the library in the normal course of their work. These questions were to be such that one of the documents which they had been given would provide a satisfactory answer to a question. These questions were collected and transferred to search cards.

Altogether there were 186 questions, and Miss Warburton, with some small assistance of a former member of the project staff, made searches for all these questions. In addition, members of the English Electric Library staff repeated 68 searches. As in the main project the searches were continued until the source document had been located or until no further search programmes could be devised, and the searches were considered on this basis as being either successes or failures. The results are as shown in Table 7.1.

	Searches	Success	Failure	% Success
Project staff	186	144	42	77.4
E.E. Library staff	68	50	18	73.5

TABLE 7.1

SEARCH RESULTS IN ENGLISH ELECTRIC CATALOGUE

The average time for carrying out a complete search was $5\frac{2}{3}$ minutes; for failures only, it was nearly 11 minutes, while for successes it was only 4 minutes.

The first point of importance arising from this test was to find that the recall efficiency fell within the range of the systems tested in the Cranfield Project. As explained earlier, the English Electric staff normally used less elements in notation than had been Cranfield practice, this resulting in a lower number of cards for the chain index. In addition, the indexer was in the habit of checking previous indexing of similar documents, and this achieved greater consistency in notation. One result of these factors appeared to be a decrease in retrieval time compared with the Cranfield facet index.

An interesting point arose in the searching by Miss Warburton. It had been thought that there might be difficulty for her in searching a system where the classification scheme was unfamiliar and where many of the questions dealt with

a subject field (nuclear engineering) in which she had no practical experience. In fact, her first fifteen searches were all successful and this, combined with the final success rate, shows that for an experienced librarian neither an unfamiliar indexing system nor a relatively strange subject field present too serious problems.

	E.E. Test	Facet Cranfield Test
Poor Question	12%	4%
Indexing	43%	47%
Searching	32%	33%
System	13%	16%

TABLE 7.2

REASONS FOR FAILURE IN ENGLISH ELECTRIC CATALOGUE

The comparison of failures (Table 7.2) showed remarkable consistency. For instance, assessed indexing failures were 47% against 43%. Within this, the figure for insufficient indexing was 22% as against 20%. The errors due to the system were in both cases mainly concerned with difficulties caused by the chain index.

As has been discussed earlier, the preferred order with a chain index had appeared to be a major weakness of the project facet catalogye, so it was interesting to have this confirmation with the English Electric catalogue. It was this which made us decide to retest the facet catalogue as discussed on page 59 , and the results of this test were sufficiently conclusive for the new systems subsequently set up at two other English Electric plants to be without preferred order.

It had been the hope that it would be possible to make an assessment of the relevance of other documents retrieved in the searches, but due to staff changes, this proved impracticable. Although, in this respect, the test cannot be considered as complete, it still served to show a number of interesting points. Most important, it showed that this method of testing was comparatively straightforward and provided a considerable amount of data for analysis, some examples of which are given in Appendix 7A. Further it appeared to confirm

in a real life situation, in a somewhat different subject field, many of the results obtained in the project. This latter point might be argued, the contention being that the similarity of results merely confirms that this level of results will be achieved when this type of test is made. This is not an argument which we accept, but it is considered in detail in the final chapter.

During a visit by the Director to America in February 1961, the National Science Foundation asked if it would be possible for the Cranfield team to carry out a test on the index to metallurgical literature prepared for the American Society of Metals by the Center for Documentation at Western Reserve University. This proposal had the support of Dean Jesse Shera and Professor Allen Kent of the Center, and this was an opportunity not to be missed, even though there were obvious difficulties in carrying out a test at such long range. During two days spent by the Director at Cleveland, discussions were held with the staff at W.R.U. who would be concerned with the work and the method of operation was established.

The metallurgical index at W.R.U. is widely known, and well documented (see Refs. 4 and 5). Of particular interest was that W.R.U. were shortly taking delivery of a G.E. 225 computer, and searching would be done on this machine.

At the time of preparing this report, the investigation on the W.R.U. index is not completed, so the account which follows does not attempt to make any final assessment of the efficiency of the W.R.U. index. Rather it is intended to illustrate the principles of the testing technique, the development from the previous test, and further developments possible.

The collection of documents to be tested was restricted to some 1,300 documents which were indexed from journals or patents issued during the first three months of 1961. Restricting the collection to this size was mainly for practical reasons associated with searching in the computer, but it had the added attraction that it was possible for this relatively small collection of documents to be indexed at Cranfield, thus permitting a comparison to be made between two separate systems.

W.R.U. sent to Cranfield the set of references in the form of marked title pages of all the items indexed. These would have already been indexed in the usual way in America and a special tape was prepared containing the indexing entries for this set of documents. The indexing at Cranfield was done entirely by Mrs. J. Aitchison, using the English Electric facet classification which has been described earlier in this chapter. Since metallurgy had been a subsidiary subject

at the Whetstone library, it was found necessary to expand and make additions to the schedules in order to cope with the concentration of documents in the comparatively narrow subject field. This in itself was an interesting and useful exercise and will be considered in detail in the report on this work; here it is sufficient to say that little difficulty was found in accommodating the new concepts.

Instructions to the indexer at Cranfield were that the indexing should be as thorough as was reasonable, and should for preference err on the side of over-indexing rather than under-indexing. As can be seen from the sample indexing card, (Fig. 4), the indexing was recorded in three stages. First the concept indexing was done; this is to say that the indexer wrote down those concepts in the document which were considered to be sufficiently important to merit indexing. The second stage was to translate the concept indexing into the notational elements of the descriptor language and the final stage was to decide on the suitable combinations of the separate notational elements which would serve as index entries. This last sentence shows that the use of the chain index was abandoned, a decision which appeared to be justified in the light of the results obtained in the project and in the English Electric test.

It was desired to simulate as far as possible the conditions under which the W.R.U. index would be tested. In a conventional card catalogue, the searcher has, at the time and in the course of the search, the advantage of being able to read the titles, the references and possibly the abstracts of the documents retrieved, and this might well be of assistance in devising further search strategies, or the eye might alight, by chance, on an entry under a heading that might not otherwise be searched. With a computer, however, the searcher will not have any such assistance. Therefore in the card catalogue prepared for the Cranfield indexing, the cards contained nothing more than the notational heading and a code number which identified the individual article.

The assistance of a group of twelve metallurgists in ten different organisations was enlisted for the preparation of the questions to be used for the test searches. Each person was sent a number of the marked title pages and allowed to make a personal selection of articles dealing with subjects of his particular interest. These articles were then used as the basis for questions of a kind which he might expect to make in the course of his normal duties. A total of 137 questions were obtained in this way, and were sent to W.R.U. There the questions were programmed, searches made in the computer, and the references retrieved in

the course of these searches sent to Cranfield. Each search was checked to ascertain whether the source document had been retrieved; if so, the search was considered successful and W.R.U. were advised which had been the source document. When the search was unsuccessful in this respect, W.R.U. were not given the reference to the source document, and so were able to make a fresh search.

The questions were also given to members of the project staff to search in the Cranfield facet catalogue. The difference between these searches and those done in the main project test was that in this case the searcher did not know when or whether the source document had been retrieved. This meant the search was continued until no further reasonable programmes could be devised, and since, as described above, the catalogue cards only contained the notation and a document code number, the searcher had no information to assist in deciding whether a satisfactory reference had been found.

Following the searches, a sample was taken of the other documents which had been retrieved in the searches, and these were sent to the individual who had originally prepared the questions. They were asked to make an assessment of those other documents in relation to the questions in the scale of 1 to 5, these representing:

1. A document more useful than the source document
2. A document as useful as the source document
3. A document of some interest
4. A document of no interest
5. A false drop.

In practice it was found difficult for different persons to be consistent in their assessments between 4 and 5 and in the statistical analysis it was decided to group these assessments together.

The results of the tests on the two indexes are given in Table 7.3. There is no sinister reason for the fact that not all the figures are given; it is merely that they are not at present available. Some preliminary discussion of these results and analysis of some of the W.R.U. failures is contained in papers by J.Melton and A.Rees (Refs.6 and 7), and further comment can await the final report on this test. It is, however, already apparent that the method of organising this test has made considerable advances beyond those adopted in the test of the

English Electric catalogue. This is partly due to the fact that in this case there was a control provided by the indexing done at Cranfield, which was made possible because the test collection had been restricted to 1,300 documents at the request of W.R.U. who would have had major problems in searching by the computer through their whole collection. Originally a collection of only 1,300 documents might have seemed too small to give valid results but now that the main project has given so much basic data concerning the performance of retrieval systems, it is practical to make further subsidiary tests with relatively small effort.

It was frankly a surprise to find that the W.R.U. system was not particularly effective either from the viewpoint of recall or relevance. The use of roll indicators should in theory have improved the relevance by eliminating irrelevant documents and the very detailed indexing as practiced by W.R.U. would be expected to result in a high recall. In fact neither with recall or relevance did it equal the performance of the Cranfield index and various hypotheses can be formulated as to why this should be.

If there have been clerical errors with the computer, these will be easily revealed by the analysis. More likely is that the fault will lie with the intellectual processes of concept indexing or in the formulation of the search programmes. To check these factors and to find the exact effect of the descriptor languages, it is intended that there shall be an exchange between W.R.U. and Cranfield of these two sets of data. W.R.U. will use the Cranfield programmes for searching in their index, while we will use the W.R.U. programmes for the facet catalogue. Then we shall pass to W.R.U. the concept indexing done at Cranfield to be translated into their system and retested both with W.R.U. and Cranfield programmes, while we shall do the same with the W.R.U. concept indexing. When this is done, it is certain that a considerable amount of additional information will be available on the strengths and weaknesses of the two systems and still more on the methods in which they have been used.

Further information has already been obtained from the Cranfield indexing, where the instructions were that every reasonable concept should be indexed and there should not be any inhibitions concerning the number of entries through concern regarding the size of the resulting catalogue. The intention was to do "exhaustive" indexing, so that it would be possible to find the effect of reducing the number of entries. Originally, as will be seen from Table 7.3, on an average

	W.R.U.	Cranfield
No. of documents indexed	1300	1150
Average No. of elements per document		7
Average No. of entries in catalogue		$12\frac{1}{2}$
Average indexing time		
Concept indexing		4.4 mins.
Notational indexing		4.5 mins.
No. of questions searched	125	129
No. of source documents retrieved	91	116
Percentage success	79%*	90%
Average No. of documents retrieved per search	15	7.4
Average search time		7.3 mins.
No. of source documents found in:		
1st programme	79	47
2nd programme	12	38
3rd programme		16
4th programme		8
5th programme		1
6th programme		6

TABLE 7.3

RESULTS OF SEARCHES IN WESTERN RESERVE UNIVERSITY AND CRANFIELD INDEXES

* After the searches had been made, it was found that due to clerical errors, 9 source documents had been omitted from the W.R.U. search file. This percentage figure has therefore been adjusted accordingly.

7 notational elements were used for each document with $12\frac{1}{2}$ entries in the catalogue.

No. of entries	% Source Documents Retrieved
$12\frac{1}{2}$	90%
8	89%
5	82%
3	74%

TABLE 7.4

RECALL FIGURE FOR SOURCE DOCUMENTS AS AFFECTED BY NUMBER OF ENTRIES

This resulted in a recall figure of 90%, The entries were then re-assessed and approximately one third were eliminated, so that an average of only 8 cards were in the catalogue for each document, (Table 7.4). On re-searching with the original search programmes on this basis, it was found that the recall efficiency remained at 89%, which shows that the eliminated entries nearly all represented redundant indexing, at any rate in so far as the search questions used in the test were concerned. The next stage was to reduce the entries to an average of 5 per document; this had the effect of bringing the recall figure down to 82%. The final stage was to bring the entries to an average of 3, resulting in a recall figure of 74%.

Further analysis of the search results is based on the 120 questions which were searched both at W.R.U. and at Cranfield.

In order to obtain a complete picture of the position regarding recall and relevance, an assessment was made not only of the documents retrieved, but of all the documents in the collection in relation to every question. For discussion of this, it is necessary to define the meaning given to the two terms, 'recall ratio' and 'relevance ratio'.

'Recall ratio' equals $100\,\frac{R}{C}$ where C equals the total number of documents in the collection which have an agreed standard of relevance to a given question, while R equals the number of those relevant documents retrieved in a single search. On the other hand, 'relevance ratio' equals $100\,\frac{R}{L}$ where L equals the total number of documents retrieved in a single search. As an illustration, presume that in a given collection of documents, ten are known to have an agreed satisfactory standard of relevance. In a single search, six of these documents are retrieved, plus another twelve documents which were irrelevant. In this situation recall ratio equals

$100 \times \frac{6}{10} = 60\%$ while the relevance ratio would be $100 \times \frac{6}{18} = 33\%$.

The results of the analysis showed that a rating of relevance 2 applied to 175 documents, made up of 120 source documents and 55 non-source documents. In addition there were 307 documents of relevance 3, that is documents of some interest. Table 7.5 illustrates the effect which decreasing the number of entries has on both recall and relevance. It will be noted that the recall figures for relevance 2 documents are lower than those given in Table 7.4. The latter referred only to retrieval of source documents, whereas the figures in Table 7.5 include the 55 non-source documents that were assessed as relevance 2, and the recall figure for these was lower than for source documents.

It can be seen from Table 7.5 and the accompanying graph (Table 7.6) that, as recall falls quite sharply, there is a slight improvement in relevance. Within the range in which we were operating, it appears that 1% improvement in relevance results in a 3% drop in recall, this applying both to relevance 2 documents and to all documents of relevance 2 or 3. It should be mentioned that this graph represents performance as influenced by indexing, and not by searching. By this it is meant that in all cases the same level of search requirement was maintained, namely that references would only be accepted if the notation contained a minimum of one less concept than originally demanded (see page 14 and Appendix 2B). It will be noted that increasing the entries per document from 3 to 5 and then from 5 to 8 resulted in an increase in recall ratio of 9% and 7%, but by increasing from 8 to $12\frac{1}{2}$ entries only gave an increase in recall of 2.4%. From this it is assumed that no further increase of indexing entries is likely to result in any material changes in the recall ratio, in other words we have probably come quite close to the maximum recall figure that can be obtained by indexing. If, however, the search rules were relaxed, then it would certainly be the case that the recall ratio would improve, and the upper dotted line in Table 7.6 is a probable extension of the performance curve. Alternatively if the search programme had been made stricter, so that no reference would be accepted unless it fulfilled the requirement of containing all the concepts of the question, the recall ratio would decrease as shown in the lower continuation of the dotted line.

Another family of curves could also be produced by varying the search conditions as given in the preceding paragraph but holding steady the indexing with regard to number of entries. This is an investigation which still has to be done.

No. of entries	No. of documents retrieved	Relevance 2		Relevance 2 & 3	
		Recall Ratio	Relevance Ratio	Recall Ratio	Relevance Ratio
$12\frac{1}{2}$	891	83%	16.3%	64.1%	34.6%
8	824	80.6%	17.1%	60.6%	35.5%
5	643	73.1%	19.9%	50.7%	38%
3	491	64.6%	23%	42.1%	41.3%

TABLE 7.5

RECALL AND RELEVANCE RATIO IN FACET CATALOGUE OF W.R.U. TEST AT VARYING INDEX ENTRIES

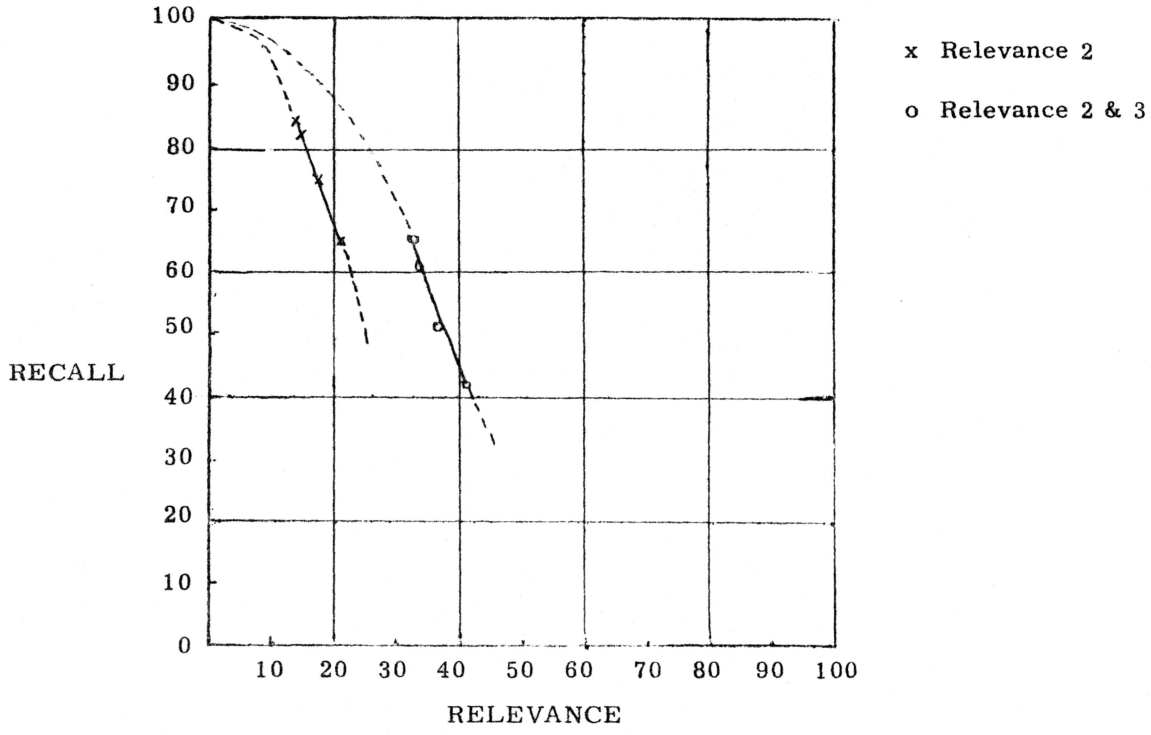

TABLE 7.6

PERFORMANCE OF FACET CATALOGUE OF W.R.U. TEST

An analysis of failures has been made for the Cranfield indexing and some of the more interesting are included in Appendix 7B. Summarised, the reason for 13 failures are shown to be as in Table 7.7.

Question	Too broad	1
Indexing	Insufficient indexing (concepts)	6
	Incorrect indexing (concepts)	2
	Omission of index entry into *term in schedules*	1
	Failure to translate concept indexing into class number	4
	Lack of permutation elements	6
Searching	Failure to search systematically	1
	Insufficient searching	3
System	More than one place in schedules for single subject	2
	Inability to permute in parts of schedules	1

TABLE 7.7

REASONS FOR FAILURE IN FACET CATALOGUE OF W.R.U. TEST

These reasons in total exceed the number of failures because the assessments often showed more than one aspect as responsible for the failure. Breaking down the analysis in another way gives the following reasons for failures:-

Indexing alone	6
Indexing and question	1
Indexing and searching	3
Indexing, searching and system	2

One aspect of this work which has been the cause of comment concerns the questions which are used for searching. There is a feeling that these questions are not typical and are in some way different from "true" questions. The problem has been well put by Alan Rees in his paper dealing with the W.R.U. test, and the following paragraphs are extracted from his report: (Ref. 7)

"The Synthetic Nature of the Questions.

"It is difficult to state explicitly how these differ from natural questions. Their artificial nature is obvious at once to all of our programmers. A synthetic question is based upon an association of concepts often peculiar to the document described and is therefore derived from an artificial environment. A real question, on the other hand, is derived from a wide background of associated

ideas and represents concepts which a questioner feels are likely to be found in a document or documents. Such questions can be formulated in many ways due to the difficulty of individuals in real question asking situations in expressing their information needs.

"I have two points which I wish to make in relation to the synthetic nature of the questions-

"First, since a synthetic question is based upon one document,the formulation of the question reflects the individual, subjective interpretation of the questioner of the significance of the document. It is, therefore, analogous to the problem of indexing a document which is dependent upon point-of-view, subject background and insight of the indexer.

"I raise this point here since a preliminary check reveals that although some questions are based upon the title or author abstract of the source document, at least one question asks for concepts of minor importance in the text and, more seriously, another asks for "Processes available for producing sound, contamination free welds in titanium and its alloys" (Q21) while examination of the text of the source document by several members of our staff reveals nothing beyond an oblique mention of titanium in an eight page article. In the text there is no indication that contamination free welds have been achieved for titanium.

"The subjective judgement of our indexers is, therefore, to be balanced against the subjective judgement of Mr. Cleverdon's question compilers. Or, to put it in another way, the language of the question and the language of the indexing are not equidistant from the document.

"Second, since in many instances the association of concepts in the question is peculiar to the document from which it is derived, being perhaps an inverted statement drawn from the text, the relevance of any items identified, other than to source document, is artificial and it is to be questioned whether we are discussing the same kind of relevance as found in searches based on real questions.

"Third, the synthetic nature of the questions naturally precluded any negotiation with the question asker as to their precise meaning. Our experience in the analysis and machine searching of some six hundred questions now permits us to identify those questions where further clarification is needed prior to final definition of search strategy.

"In at least ten cases in the present series of test searches we were able to predict failure prior to searching. If the same requests were received from real question askers a concerted effort would have been made to clarify them. We have found that the degree of success achieved in searching is directly proportional to the amount of insight possessed into the nature, purpose and environmental origin of the requests. It is particularly difficult to plan an effective search on the face value of a written request."

While we would not necessarily agree with all these comments, at present there is little point in arguing the matter, for there is not sufficient evidence either way. It is reasonable to assume that the test figures are affected in some way or another by the use of special questions, but it does not necessarily follow that this affects the validity of the test. More work requires to be done on this specific point to decide exactly what correction factors should be built into the results, and we await with interest more detailed analysis of the questions from this viewpoint, which the experienced staff at W. R. U. will be making. In addition it is the intention to search the two catalogues with a number of questions that are put to W. R. U. in the course of their normal service to The American Society of Metals and this should help to a better appreciation of this problem.

It would be unwise to be over-confident on the general applicability and acceptance of this test technique, but the results so far are sufficiently encouraging to justify making a series of such tests in other subject fields. It is capable of extension by any organisation which is seriously interested in, for instance, achieving maximum economic efficiency. A short series of tests covering documents indexed under varying time allowances would permit the optimum to be selected. As has been done with the Cranfield-W. R. U. index, the effects of exhaustive indexing can be investigated, and the decision as to whether to include more or less cards in a catalogue or whether to use 10 or 50 uniterms can be based on something more logical than hunches.