1965 d

APPLICATION FOR GRANT TO
THE NATIONAL SCIENCE FOUNDATION
WASHINGTON

An investigation into the methodology
of evaluation techniques based on
a test of the MEDLARS system of the
National Library of Medicine

Director
Cyril W. Cleverdon

Submitted by
Association for Special Libraries and Information Bureaux (Aslib)
London, England
June 1965

AN INVESTIGATION INTO THE METHODOLOGY OF EVALUATION TECHNIQUES BASED ON A TEST OF THE MEDLARS SYSTEM OF THE NATIONAL LIBRARY OF MEDICINE

PROPOSAL

A grant of $55,832 is requested by Aslib from the National Science Foundation over a period of two and a half years, for the purpose of the design and direction of an investigation into the methodology of evaluation of information retrieval systems, based on an evaluation test of the MEDLARS system of the National Library of Medicine. The work involved in carrying out the test will be the financial responsibility of the Library, and this application for grant is therefore limited to those activities which will be done in England.

Some general problems of evaluation methodology are considered in the paper included as Appendix 4, a paper which was stimulated by the discussion at the meeting on evaluation of document searching procedures, organized by the National Science Foundation and held in Washington in October 1964. In the account of this meeting published in Scientific Information Notes, Vol. 6, No. 6, it is stated

"One prominent suggestion, which was offered by a number of the participants, has found wholehearted acceptance among the Foundation staff; namely that tests of procedures now in use be carried out by employing several possible alternative test methods, such tests being carefully designated to determine the merits of the methods used, as well as to produce data on the performance of the procedures being tested. The Foundation, therefore, invites enquiries and proposals from organizations interested in designing and conducting such tests with the primary purpose of advancing our knowledge of testing methodology for document searching procedures."

The Director of the National Library of Medicine advised the Foundation that he would be willing to assist in any such development of the methodology of evaluation by making available the MEDLARS facility. It is as a result of the discussion with the Director and the staff of the Library that the present proposal is made; its practical implementation is entirely dependent upon the active co-operation of the staff of the Library, and it is only right to acknowledge their generous and enthusiastic response to the proposal, the valuable suggestions relating to the test design, and the offer to carry out the activities necessary to meet the test design.

The objectives of the project can be stated as follows:-

1. The development of a controlled study for the testing of the methodologies for evaluating information retrieval systems.

2. The measurement of the present operational performance of MEDLARS, using methodologies developed in (1) above.

3. The establishment of the factors responsible for specific system failures.

4. Development of a methodology for continuous quality control of MEDLARS.

It is necessary that the test should be so designed that it will provide the data required to answer the questions raised by management. Therefore, the first stage was for the National Library of Medicine to determine their requirements, and these appear in Appendix 1. The general test design is given in Appendix 2; a discussion of the design is given in Appendix 3. This test has been designed specifically with the requirements of the Library in mind; it also, however, contains the necessary elements to permit comparison to be made between different evaluation methodologies. It is intended that the detail design should be prepared during the first six months of the test, in which time experimental testing will be carried out for the purpose of providing data to assist with this detail design. It might be said that the general design, as given in Appendix 2, goes into some detail in that, for instance, it specifies the number of searches to be made. These figures should not be taken as fixed, but, where included, they are intended to represent the upper limit of the effort which the National Library of Medicine will be asked to undertake in this investigation.

ORGANIZATION

The project will be under the administrative control of Mr. L. Wilson, Director of Aslib. The Project Director will be Mr. Cyril Cleverdon of the College of Aeronautics at Cranfield, who will be responsible for all technical aspects. The National Library of Medicine will appoint to their staff a senior analyst who will be responsible to the Project Director for the implementation of the test design, and who will devote his full time to the project. He will have the assistance of a clerical clerk. Mr. Charles Austin, of the National Library of Medicine will be responsible for seeing that the in-house duties associated with the test are carried out.

The Director of the National Library of Medicine will invite a number of workers in the field to form an advisory committee to consider the detail design and make appropriate recommendations.

The Project Director will appoint a number of consultants, who will be available to advise on various aspects of the project.

In the later phases of the work, much of the analysis will be carried out in England. Apart from this, it is envisaged that there will be meetings of the project staff either in Cranfield or Washington at intervals of not longer than three months.

In adition to the Project Director, Mr. Michael Keen, who has been on the staff of the present Aslib-Cranfield project, will be employed full-time. Mr. Keen will be concerned with the analysis and the evaluation of the results; however, in that there will be some delay before this stage of the work can commence, Mr. Keen will, for the first year, be mainly engaged on theoretical and experimental investigations into the methods that are or can be used in the measurement and presentation of the results of information retrieval tests, as detailed in Appendix 6. This will not be directly related to the evaluation of MEDLARS, and will mainly make use of the data obtained in the present Aslib-Cranfield project, but it is of major importance in the general methodology of evaluation, for agreement on measures is a fundamental prerequisite to comparative evaluation.

FINANCE

	First year	Second year	Third year (6 months)
	$	$	$
Salaries			
Technical staff (full-time including insurance, etc.)	5,750	6,000	3,000
Technical staff (part-time)	3,000	5,000	2,000
Clerical staff (part-time)	2,000	3,000	1,500
Consultants	2,000	2,000	500
Supplies	2,000	3,000	1,500
Travel (including visits to U.S.A.)	2,200	2,200	1,100
Publication			800
	16,950	21,200	10,400
Overheads 15% of above	2,542	3,180	1,560
	$19,492	$24,380	$11,960

Total for two and a half years $55,832

APPENDIX 1

TEST REQUIREMENTS OF THE NATIONAL LIBRARY OF MEDICINE

Indexing

1. Are there significant variations in indexers?
2. What is the effect of exhaustivity of indexing (I.M. and non-I.M. headings)?

Index Language

1. Are the terms sufficiently specific?
2. Are there significant variations in specificity of terms in different areas?
3. Are pre-coordinate type terms, (including sub-headings) which have been included to meet the requirements of Index Medicus, hindering the efficiency of retrieval by MEDLARS?

Searching

1. What are the requirements of the users regarding recall and precision?
2. Can search strategies be devised to meet requirements for high recall or high precision?
3. Should the output be screened by library staff?
4. Should greater effort be requested of the user?
    a. By having the information staff at his locality question him?
    b. By having NLM staff question him?
    c. By presenting him with the output of a search on a small sub-set of the whole collection, and then, if necessary, rephrasing the question?
5. What is the time of preparation of search strategies and formulations?
6. What is the average response time of the system to a request?

APPENDIX 2

GENERAL TEST DESIGN FOR EVALUATION OF MEDLARS

INTRODUCTION

This section outlines the general pattern of the methods that will be used in the evaluation of MEDLARS. It does not purport to go into details, since it is planned that the first six months of the programme will be spent in carrying out experimental tests and preparing detailed design.

SUMMARY

1. PREPARATION FOR TEST

10 - 20 research groups to be contacted.

2. TESTS

1st Series 400 actual questions.

a. 100 questions to be searched as received.
b. 100 questions to be resubmitted after the output from the search of part of the file has been seen by the questioner.
c. 100 questions to be searched after local librarian has had discussions with questioner.
d. 100 questions to be searched after NLM staff have had discussions with questioner.
e. 50 questions of (c) and (d) also to be searched as originally submitted.
f. Precision percentage will be determined by having each questioner assess 25 documents for relevance.
g. Recall percentage will be determined by having questioners, before the search is made, list known relevant documents.
h. NLM staff will attempt to assess sample of output for relevance.

2nd Series 150 (approx.) questions will also be searched in the specialized indexes at five research centres, and the results compared with MEDLARS.

3rd Series Records will be kept of a minimum of 200 searches in Index Medicus in various research centres. The main purpose will be to establish methods and effectiveness of use.

4th Series Back-up tests to be made if further data are required in any particular area. May require use of prepared questions.

3. ANALYSIS OF SYSTEM FAILURES

4. ASSESSMENT OF OPERATIONAL EFFICIENCY

5. EVALUATION METHODOLOGY

1. PREPARATION FOR TEST

The co-operation will be solicited of some ten to twenty medical research groups. These groups will be selected in such a way that they are reasonably representative of the complete user group of MEDLARS in relation to (in order of importance)

a. Various specialist subjects.
b. Scientists and clinicians.
c. Government, academic or industrial groups.
d. Availability and quality of local medical library services.

In addition it is required that some of these groups should have indexes in their own speciality which are more comprehensive in relation to coverage and/or more detailed in respect of indexing than MEDLARS. For practical reasons, it is recommended that if there is a choice between two or more groups, that group located nearest to the National Library of Medicine should be selected.

The mechanics of obtaining the necessary co-operation of these groups might differ in various cases, but would presumably include a communication from the Director of the Library to the Director of the group, followed by a personal visit from the evaluator to the group for discussion with, at least, the senior member of the library. The type of assistance required of the members of the test user groups will become apparent in the course of the design.

2. THE TESTS

There will be three or possibly four series of tests, intended to measure the performance of MEDLARS by different methods and also to provide the necessary data for analysis. In the first series, a maximum of 400 questions will be searched. These questions will be those which have originated from individuals within the test groups and will not have been specifically requested for the purpose of the test.

The originators of all questions will be asked to submit, with their questions, answers to the following points.

a. Can you give any indication of the number of papers relevant to your question which you consider are likely to have been published since January 1965.

1 - 5 □, 6 - 20 □, 21 - 50 □, 51 - 200 □,

201 - 500 □, 501 - 1,000 □, 1,001 - 5,000 □,

5,001 + □.

b. Do you wish a search which will present you with
(i) All relevant references
(ii) A selection of the relevant references.

c. If you only require a selection of the relevant references, please indicate approximate number you wish to receive.

d. If you already know of any papers published since January 1964 which are relevant to your question, please list up to five such references.

Different actions will be taken in respect of the searches for the four hundred questions as follows.

I One hundred questions will be searched as received.

II One hundred questions will be searched as received over a sub-set of the file (e.g. some twenty-five thousand references). The output from this search will be sent to the questioner for his assessment, and he will be given the opportunity to rephrase his question in the light of the type of document he has received. The rephrased question will then be searched on the complete file of MEDLARS for 1965.

III One hundred questions will be considered by the search staff. This will result in one of two situations.

a. The staff will require elucidation or further specification of the question.
b. The staff will be satisfied that they understand the question and they will prepare search strategies.

Whichever of these situations arise, the question will be referred back to the librarian of the group concerned, either with a request for information or with the proposed search strategy. The local librarian will discuss the matter with the questioner, and the redefined question or the amended search strategy will then represent the formal search question.

IV The final set of one hundred questions will be tested in a similar manner to the preceding set, except that the conversation with the questioner will be held by the evaluator, either by correspondence, telephone, or if necessary, by a personal meeting. In particular the evaluator will try to elicit whether any categories of documents are to be excluded.

V Fifty questions in each of the sets III and IV will be searched in the form in which they were originally presented, in addition to the complete search for the revised question.

It is to be understood here, and throughout all other aspects of the test, that the necessary clerical records regarding search strategies and search output will be maintained.

Following a search, the questioner will be sent the usual list of references representing the complete output of his search. For each of twenty-five of the references (or the complete output if less than twenty-five items), he will be sent a relevance-assessment form for completion. The relevance of each of these references is to be determined from the actual documents, and the questioner will be supplied with copies of all papers which he cannot readily obtain. Either in personal interview or in an accompanying letter, the questioner will be advised as to the standards he should apply in making his decisions concerning, in an agreed scale, the relevance or non-relevance of the selected documents. In particular, he will be asked to specify the reasons for the rejection of the non-relevant papers.

This test is intended to establish the performance of MEDLARS in regard to recall and precision. Its ability to do so is based on certain assumptions, the first of which is that the majority of questioners will be able to identify one or more relevant documents before the search is made. If this is the case, then it will be possible to obtain the recall ratio of the system by ascertaining how many of these known relevant documents are retrieved. Care will have to be taken not to have a bias built in to any such figure by the fact that, for certain types of questions where there are likely to be few relevant documents, questioners do not know of any relevant documents. For this and other reasons, further action will be taken to check the recall ratio.

The second assumption is that MEDLARS is operating at a precision ratio of at least 15%. If this is the case, then the expectation will be that on an average there will be three or four relevant documents in the twenty-five documents assessed for relevance by the questioner. This will enable the actual precision ratio to be obtained by measuring the number of relevant documents in the total of these analysed. Although there will be little difficulty in determining this on an overall basis, there may be some disparity between the 'firm' and 'mushy' language areas, and careful analysis will be required to ensure that the data is reliable.

Before being sent to the questioner, the information staff of the Library or the evaluator will be checking the documents selected for assessment, and it may be found that they are able to eliminate correctly a significant proportion of the non-relevant documents. If this should be the case, it would be possible to send the questioner a set of twenty-five references which were more likely to contain relevant documents, and thereby increase our knowledge of the set without involving the questioner in additional work.

The second series of tests involves a comparison with other indexes. The requirements were that up to five of the co-operating organizations should be those having indexes more comprehensive and/or more detailed than MEDLARS. These organizations will be asked to duplicate the searches for questions originating within their own organizations.

This series of tests will create serious problems in detail design and organization and it will be difficult to ensure that no bias is built in to the results. Firstly, it might be asked why questions should be submitted to MEDLARS if the organization has an index which, in its speciality, is assumed to be efficient. It might, in fact, prove to be more useful to have MEDLARS duplicate searches which the individual organizations have carried out. It will be desirable that a duplicate search by MEDLARS and another organization should be carried out within a short period of time, so that the combined output can be submitted to the questioner for analysis. Care will have to be taken in selecting the output to make it representative of the five parts into which it could be divided, namely

1. Documents retrieved by MEDLARS and Index A
2. Documents indexed by both groups but retrieved only by MEDLARS
3. Documents indexed by both groups but retrieved only by Index A
4. Documents not included in Index A and retrieved by MEDLARS
5. Documents not included in MEDLARS and retrieved by Index A.

Finally, assuming that the above difficulties are satisfactorily met, there remains the problem that the number of searches for each organization may be too small to give reliable comparison between MEDLARS and any single index, even though collectively (i.e. by adding together the figures for the five specialized indexes) a reasonable comparison can be made. There are various ways of dealing with this, and the most satisfactory procedure in the light of the test results will be adopted.

A practical matter is whether the same search programme should be used both for MEDLARS and for the specialized index. If this were done, it has the attraction that it eliminates the variable of search strategy, and therefore one is left with a comparison of the indexing and the index languages. However, experience with the test of the metallurgical index at Western Reserve University has shown that, by analysis, it is possible to sort out the effect of the various factors involved. The present evidence appears to indicate that search strategies play a most important part in the efficiency of a system, and it is therefore recommended that the MEDLARS group and the specialized organization group should each decide their own strategies. However, it will be important to ensure that, with each question, the two groups have the same interaction with the questioner.

It could be argued that, in the light of the difficulties involved, it would be a waste of effort to undertake this particular test series. The potential value of the results, however, not only in relation to MEDLARS but as a comparison of different methods, is such that it is considered justifiable to make the attempt.

The third series of tests involves a comparison of MEDLARS with Index Medicus, combined with a study of user problems in making searches in Index Medicus. The straight comparison between MEDLARS and Index Medicus can (it is understood) be made by analysis of the computer print out, since I.M. terms can be distinguished from non-I.M. terms. For the user study, the co-operation of questioners and librarians will be required. When making searches in Index Medicus, they will be asked to record their original problem, the decisions in relation to the heading searched, whether going from one heading to another was a personal decision or whether influenced by cross-reference instructions, the results of the search in relation to the relevant documents retrieved, the time taken, and a statement concerning satisfaction in respect of the user criteria, together with any comments.

This user study will not be primarily concerned with establishing performance figures so much as investigating how people use Index Medicus. Many of the questions will, on analysis, show that they were not of a nature which would justify a search being made in MEDLARS. However, where dissatisfaction with the Index Medicus search is expressed by the user, the question could be re-searched by NLM, either in Index Medicus or MEDLARS as seems appropriate.

The fourth series of tests will be carried out if it is found that insufficient information has been obtained from the first test series to give reliable data on performance in regard to recall. Although four hundred searches are to be made, it is possible that there will be requirements for additional data in certain subject areas or with certain types of questions. If this could be remedied by obtaining a further supply of actual questions, then this would be the obvious course to adopt. Failing that, it would be necessary to solicit prepared questions, these either being based on documents known to be in the collection or alternatively questions for which there are known relevant documents. The number of such searches would not exceed one hundred.

In addition to the complete clerical records that will be made to assist with the analysis, records will be maintained regarding the times taken for all significant intellectual and clerical operations involved in carrying out the tests described earlier. These would include such matters as

a. Time involved in preparing search strategies
b. Time involved in computer programmes
c. Time taken by information staff to eliminate non-relevant references
d. Time taken in discussions with users on definitions of questions
e. Search times as specialist organizations.

An operational problem in the test will be created by the fact that 60% of the documents indexed are in foreign languages. A questioner may, of course, as part of the normal MEDLARS service specify that he shall only be sent references to papers which are in languages which he can understand, but in any request for a comprehensive search to be made, it can be anticipated that a significant proportion of the documents to be assessed will be in an unintelligible language. This is a difficulty which will have to be faced; whether it is capable of relatively simple solution can only be known after some preliminary testing has been done.

3. ANALYSIS OF SYSTEM FAILURES

The major analysis will be carried out on the results obtained in the first series of tests, and will be directed towards determining the reasons for failure to retrieve known relevant documents. It is not possible at this stage to be precise as to the number of relevant documents which will not be retrieved, but if earlier assumptions are correct, it may be assumed that from 200 to 300 documents will come in this category. This would meet the minimum requirements for the analysis.

A random sample will be taken of 400 of the non-relevant documents that were retrieved in the searches, and an analysis made of these documents. Some additional analysis of a similar nature, in regard to both relevant and non-relevant documents, will be required in connection with the second series of tests, partly in order to obtain a comparison with the indexes of the specialist organizations, but also to obtain information on the value of the varying evaluation methodologies.

The methods used in this analysis will be similar to those used in earlier Aslib-Cranfield evaluations.

4. EVALUATION OF MEDLARS

From the results of the tests, it will be possible to present figures showing the operational efficiency, in regard to coverage, recall and precision of MEDLARS, either as a system or as a series of specialist sub-systems in those areas where significant differences occur.

As in the original Cranfield project, a consultant will, using appropriate statistical techniques, prepare an analysis of the test results to ensure their validity.

From the analysis discussed in Section 3, an assessment will be made of the factors which prevent optimum performance, and recommendations made as to the methods which could be used to effect required improvements. More particularly, from the test results and the analysis will come the answers to the management questions set out in Appendix 1.

5. EVALUATION METHODOLOGY

In addition to clerical records of the time involved in carrying out searches etc., records will also be kept of the time and effort involved in the evaluation itself. An assessment will be made of the techniques which have been used in the various test series, in relation to their general effectiveness and the amount of effort they entail. This test will, in itself, considerably enhance our knowledge of test techniques; more particularly, however, it will establish a "laboratory" in which further experiment and development can take place.

APPENDIX 3

DISCUSSION ON TEST DESIGN

The philosophy of this test design is based on two assumptions. The first, which can hardly be argued, is that an information retrieval system is established to meet the requirements of a user group. The second is that an evaluation of an information retrieval system or any part of it can only be carried out in relation to one or more of the requirements of the user group. Many criteria have been suggested as suitable for evaluation of I. R. systems, but the argument is advanced in Appendix 4 that the criteria of interest to the users were restricted to the following six matters.

1. Coverage; the extent to which the system includes all documents relevant to the user group.
2. Recall; the ability of the system to present all relevant documents which it contains.
3. Precision; the ability of the system to withhold non-relevant documents.
4. Time; the interval between the demand being made and the answer being given.
5. Presentation; the physical form of the output.
6. Effort; the effort, intellectual or physical, demanded of the user.

The test specification as presented by the National Library of Medicine (Appendix 1) is comprehensive in that it requires an assessment to be given in relation to each of the six user requirements. While management must be basically concerned with the ability of the system to meet these user requirements, they also need to know how each sub-system, or part of a sub-system is operating. The major sub-systems can be considered as

1. Acquisition
2. Indexing
3. Index language
4. Searching
5. Clerical routines
6. Physical store.

It will be noted that the user requirements do not match the sub-systems. 'Recall' is concerned with acquisition, indexing, index language and searching, while 'user effort' is certainly concerned with searching and the store, and may be related also to index language and indexing. It is this fact which makes management decisions so difficult, in that before improving performance in relation to any of the user requirements, the effect of the interaction in a number of sub-systems must be taken into account.

Each of the six major sub-systems can also be broken down into a number of minor sub-systems. The specification does not, in fact, require that all these minor sub-systems should be investigated, either because in some cases they do not apply to the system, or because the decision has been taken that they are not sufficiently important or critical as to merit evaluation. The questions included or implicit in the specification are listed below linked to their main sub-systems.

1. ACQUISITION

a. Is MEDLARS indexing the most useful set of journals.
b. Is the delay between the receipt of a journal and its appearance in the indexing system significantly affecting performance.

2. INDEXING

a. Are there significant variations in indexers.
b. What is the effect of experience in indexing.
c. Do the indexers recognise the specific concepts that are of interest to the user group.
d. What is the effect of exhaustivity of indexing (I. M. and non-I. M. headings).
e. Could more use be made of indexing that is published in the journals.

3. INDEX LANGUAGE

a. Are the terms sufficiently specific.
b. Are there significant variations in specificity of terms in different areas.
c. Are pre-coordinate type terms, which have been included to meet the requirements of Index Medicus, hindering the efficiency of retrieval by MEDLARS.
d. Should C. I. M. cross-references be included in each monthly issue. Would this improve the ability of the user to make successful searches.
e. Is the quality of term association in MESH satisfactory.
f. Should topical sub-headings be reintroduced.
e. Would weighting of terms improve precision.

4. SEARCHING

a. What are the requirements of the users regarding recall and precision.
b. Can search strategies be devised to meet requirements for high recall or high precision.
c. Should the output be screened by library staff.
d. Should greater effort be requested of the user
   (i) by having the information staff at his locality question him
   (ii) by having N. L. M. staff question him
   (iii) by presenting him with the output of a search on a small sub-set of the whole collection, and then, if necessary, rephrasing the question.
e. Can users search Index Medicus as effectively as library staff.
f. What searches justify use of MEDLARS rather than Index Medicus.
g. What is the time of preparation of search strategies and programmes.
h. What is the response time of the system to a request.

5. CLERICAL ROUTINES

a. Do input procedures result in significant number of errors.

6. STORE

a. Are computer programmes flexible enough to obtain desired performance level.
b. Do computer rejects and 'bugs' delay system response time.

On the practical level the test design is also based on two assumptions. The first is that there is a limit to the amount of additional work that can be put on to the MEDLARS staff. The second is that one obtains diminishing returns by trying to have the questioner do too much analysis.

If the effort requested of the Library staff can, as is planned, be spread over a time of sixteen months, it will represent something in the range of 5% to 10% additional to their present normal output. This is understood to be acceptable.

With regard to the users, it is estimated that they can carry out the requirements in a period of two to three hours, this being additional to the time one might expect them to spend for their own purposes. It has been found in other Cranfield projects that two scientists out of three are willing to co-operate if told that this is the time limit. More important, we find that, as far as can be judged, the work is done conscientiously, which is essential for an investigation such as this, and it is better to have twenty-five documents assessed carefully than twice as many where the decisions are taken without due thought.

It is essential to bear in mind that it is unrealistic to expect the co-operation of a large number of users if too much is demanded of them, and there is no part of the test which requires more care and control than the decisions by the users regarding relevance. As knowledge of and experience with evaluation techniques has advanced, in successive tests by Cranfield,greater efforts have been made to ensure that the decisions relating to question/document relevance are made as objectively as possible. In this design, by the use of genuine search questions, by the relevance assessment of the output by the questioner at the time he wants the information, by the control that will be established by having statements on the reasons for his relevance decisions, we are going further than has elsewhere been attempted in obtaining reliable decisions. That this will involve a greater effort is appreciated, but it is considered to be entirely justified bearing in mind the objectives of this test.

Although, as previously stated, there is need for care not to demand too much of those voluntarily co-operating, experience has shown that there are a number of scientists, possibly one in every twenty, who become very interested in a documentation research project, and who can be prevailed on to undertake additional tasks. If this should be the case in this evaluation, there will be the possibility of investigating other techniques for establishing recall, such as saturation searches or sampling. This is a method which is superficially very attractive, but which has not been included in the design because of the great effort it demands of those making relevance assessments. However, if willing co-operators are available, it would be interesting to try this technique.

APPENDIX 4

IDENTIFICATION OF CRITERIA FOR EVALUATION OF OPERATIONAL INFORMATION RETRIEVAL SYSTEMS

by

Cyril W. Cleverdon

Many criteria have been proposed as being suitable for measurement in the evaluation of I. R. systems as, for example, those listed by Bourne in Ref. 1. I suggest that for all I. R. systems, such criteria fall into two distinct classes, namely those which either are or are not of interest to the user of the system. Criteria of interest to the user are limited to the following:-

1. The extent to which the system includes all relevant literature, (i. e. coverage)
2. The ability of the system to present all relevant documents, (i. e. recall)
3. The ability of the system to withhold non-relevant documents, (i. e. precision)
4. The interval between the demand being made and the answer being given, (i. e. time)
5. The physical form of the output, (i. e. presentation)
6. The effort, intellectual or physical, demanded by the user, (i. e. effort).

All other criteria are the sole concern of the system managers, which is a general term to include all those who decide the policy, finance the system or are in any way responsible for or participate in the actual operation of the system. These management criteria are only important in relation to their influence on the user criteria. Since it is a reasonable assumption that an I. R system basically exists for the purpose of meeting the requirements of the user group, it follows that for management the most important criteria should be those of direct concern to the user group. Therefore it is shown that an evaluation of an I. R. system must basically concern itself with the set of user criteria.

The five user criteria listed above fall into three categories. Recall and precision represent software, time and presentation represent hardware, and effort falls into an intermediate category that is influenced both by the software and the hardware.

The evaluation of the hardware criteria from the user's viewpoint is quite straightforward. To find the time factor it is only necessary to record the time lapse between the request and the receipt of the output for a statistically valid number of cases. To evaluate the presentation, one has merely to observe whether the user receives a list of document numbers, a list of bibliographical references, a list of titles, a set of abstracts or a set of complete documents, either readable text or microform. To evaluate the effort demanded by the user is only slightly more complex because of the possibility, in certain systems, that the effort can vary from the minimum of expressing the query in natural language to the maximum of conducting the complete search unaided. However, in any single system, evaluation of this point appears to require only a straightforward observation of a number of cases. 'Coverage' is normally a management decision, and it is a relatively straightforward task to calculate its extent.

In Ref. 1, Bourne states and asks:

"It is not clear why so much attention has been given to recall and relevancy. Should these be regarded as better criteria than any of the others proposed?"

By a process of elimination, the comment and question have been answered, for it is shown that there are only six fundamental or user criteria, and of these the only two that demand any serious intellectual effort in their measurement are recall and precision. This is not to deny that there is considerable interplay between the six user criteria, nor to suggest that the management criteria are un-important, but I do emphasize that the importance of the management criteria is in relation to the effect which they have on meeting, economically and efficiently, the user requirements.

The requirements of the users will vary enormously. For recall the demand may be anything from 1% to 100%. The time requirement may be three minutes, three hours, three days, three weeks, or even longer. Each member of the user group will have his own views - probably conditioned by what he is accustomed to - as regards presentation and amount of effort demanded. Regarding precision, the tolerance ratio is less predictable, and may well change with the growth of the system, but the realistic user will probably know that if he demands 100% precision he will inevitably have to accept a low recall rate.

An evaluation test is a management tool which enables management either to improve the performance of the system in relation to the user criteria, or alternatively (and hopefully concurrently) to enable decisions to be taken which will permit management operations to be carried out more economically. Bearing this in mind, it appears reasonable to argue that the complexity of an evaluation of an operational I.R. system has been sometimes over-emphasized. Although there are the complications of the interplay between the conflicting requirements of the user group, and between the user group and the system, yet the constraints that an operational system imposes on itself greatly reduce the variables that have to be considered, as opposed, for instance to the freedom from constraints that is characteristic in the evaluation of experimental systems.

The first objective of testing an operational I. R. system must be to measure the performance of the user criteria. Hopefully, sometime in the future, we may be able to evaluate, in the sense of compare, performance figures obtained by one system with those obtained by other systems, but until a great deal more data has been accumulated, researchers should confine themselves to measuring the performance of a system as an entity. In making the measurements of the performance in regard to user criteria, the data that are obtained should enable an evaluation to be made of the management criteria, and I would argue that it is only by this means that management criteria can be assessed. 'Index Medicus' uses an average of two entries per document, 'STAR' averages five entries, 'MEDLARS' averages nine entries, "TAB' averages fifteen, N.A.S.A. computer system averages twenty-two, American Society of Metals computer system averages thirty. No-one can evaluate the management decisions which resulted in these variations without first knowing how these decisions affect the performance of the user criteria. It is, of course, true that some decisions in regard to these indexes have been taken primarily for economic reasons, but again, until one knows for certain how the performance has been affected, it is impossible to argue the logic of the economic decisions.

The dependence of the management operational criteria on the user criteria is set out in Table I. There is, of course, a certain amount of overlap, but it is shown that recall and precision are in no way influenced by the type of store, and that conversely time is not affected by the type of index language, although it is possible that a decision to multiply access points for a given document would affect search time, in, for instance, a card catalogue or printed index.

Table I

Management Operational Criteria in Relation to User Criteria

RECALL and PRECISION	Indexer's ability to analyse document content Indexing speed Exhaustivity of indexing Index language recall and precision devices Index language structure Index language susceptibility to error Ability of users to express requests correctly Programmer's ability to devise search strategies.
TIME	Type of store Type of query Location(s) of store Demand rate Input to store Size of collection
REPRESENTATION	Type of store
USER EFFORT	Type of store Location(s) of store Availability of management staff Ease of translation from natural to index language

In addition to the operational criteria shown in the table, there will be for management a further set of criteria, all of which have the common characteristic that they involve financial considerations. As such, it can be argued that the financial criteria are, in the final count, the most important of all, but I would argue that in the evaluation of an operational system, one should work back to cost criteria by way of user criteria and management operational criteria.

The foregoing is an outline of what can be considered as of general applicability in the evaluation of any operational I.R. system. The paper will now develop the argument to show what is to be or can be done in the evaluation when applied to a single system.

First to define a user group. This I would do by saying that, in our context, it consists of a number of people who are working towards a common objective. This objective may be acquisition of further knowledge on, or the application of, a particular discipline (e.g. statisticians), it may be the design, manufacture and sale of a particular object (e.g. automobiles), it may be the defence of the nation, the maintenance of civil law and order, or the health either of a nation or of all human beings. In many cases the members of the user group will be of different subject disciplines; in other cases the users will be a single discipline sub-group within the major group.

From these various categories, I take the information services of the National Library of Medicine as an example for discussion. This is an organization which accepts as its mission, inter alia, the task of acquiring and indexing all recorded medical information as a step towards:

1. The publication and wide distribution of monthly and annual indexes to medical information.

2. The provision of a search service for specific requests, this facility to be available at a number of different locations.

To enable it to carry out these services, neither of which are intended to be financially self-supporting, it is granted an annual budget of $x. The basic problems for investigation can, as with all information services, be considered as being:

1. What are the requirements of the user group?

2. Are the requirements of the user group being met?

3. If the answer to (2) is yes, can these requirements be met for a lower expenditure of money?

4. If the answer to (2) is no, what improvements can be made which will permit the requirements of the user group to be met?

This immediately raises a difficulty, which is to define the requirements of the user group. Although little serious research work has been done on this matter, every librarian has a considerable amount of accumulated knowledge on how far his customers are willing to accept less than the optimum. However, so great are the variations in this respect, that it appears the most straightforward and logical course to assume that optimum performance in each of the five user criteria is required, that is to say 100% recall, 100% precision, immediate service of complete documents with minimum user efforts. It is true that it is rarely that these desired objections will be simultaneously obtainable, nor is it likely that the large majority of users will demand or expect their simultaneous attainment. However, within the user group there will be individuals who will, at some time or other, require and expect optimum performance in at least one of the criteria.

In relation to the N. L. M. there are obvious management decisions which affect the practicability of approaching the optimum performance of the various criteria, and one might set out their decisions in relation to user criteria as follows:

1. Coverage

In regard to coverage, the decision is to attempt to make possible optimum performance, including in the system the widest possible coverage of medical literature.

2. Recall

There has been a double decision. Searches in MEDLARS are intended, by virtue of increased exhaustivity of indexing, to give maximum recall; something less than maximum is accepted for Index Medicus, where a lower level of exhaustivity is practised.*

* It is appreciated that Index Medicus serves another purpose as a current awareness bulletin, but at present we are considering it only in relation to its role as an index that is used for retrieval.

3. Precision

Here again there is a division between the two types of indexes. In Index Medicus, precision is limited by the subject headings used: in MEDLARS, post-coordinate searching is intended to make a higher level of precision possible.

4. Time

Management decisions based on the problems concerned with Index Medicus have affected the form of the store for MEDLARS, and thereby influenced the time factor for a search in the latter. The view has presumably been taken (as a personal opinion, quite logically and correctly) that if a quick search is essential, it must be accepted as being possibly incomplete, and for such searches Index Medicus is available. Conversely, if a search has to be comprehensive, then time penalties must be accepted, so the relative slowness of a computer search can be tolerated.

5. Presentation

The decision here is that output shall be in the form of titles and references.

6. Effort

Index Medicus can involve anything from minimum to maximum user effort; MEDLARS requires only the exact formulation of the question by the user.

The constraints of this particular I.R. system are now obvious, and there is no purpose in attempting to evaluate, in regard to N.L.M., any aspects which do not come within the boundaries of the operation. The following, then, are the questions which the system managers should ask in relation to searches by MEDLARS.

1. To what extent is the complete literature of possible interest to the user group being entered in the system?

2. Of the documents in the system, on an average what proportion is being retrieved of those which are relevant to an enquiry?

3. In answering an enquiry, on an average what proportion of the retrieved documents are not relevant?

In the carrying out of a test, the managers can expect that the test should, in addition to finding the performance of the system in relation to the above matters, also provide data on the following points which are relevant to operational criteria.

1. To what extent was the failure to retrieve relevant documents due to

   a. Indexing, whether human errors or policy decisions fixing too low a level of exhaustivity.

   b. Searching, whether human errors or too specific searching.

   c. Clerical errors within the system.

   d. Index language, whether faults of the structure or lack of recall devices.

   e. Original questions being framed too precisely.

   f. Their non-inclusion in the system.

2. To what extent was the retrieval of non-relevant documents due to
    a. Indexing, whether human error or too high a level of exhaustivity.
    b. Searching, whether human error or too general searching.
    c. Clerical errors within the system.
    d. Index language, whether faults of the structure or lack of precision devices.
    e. Original questions being framed without sufficient precision.

Further, the test should be carried out in such a way that data are available to permit the system managers to assess the effect of changes in the operations, such as:

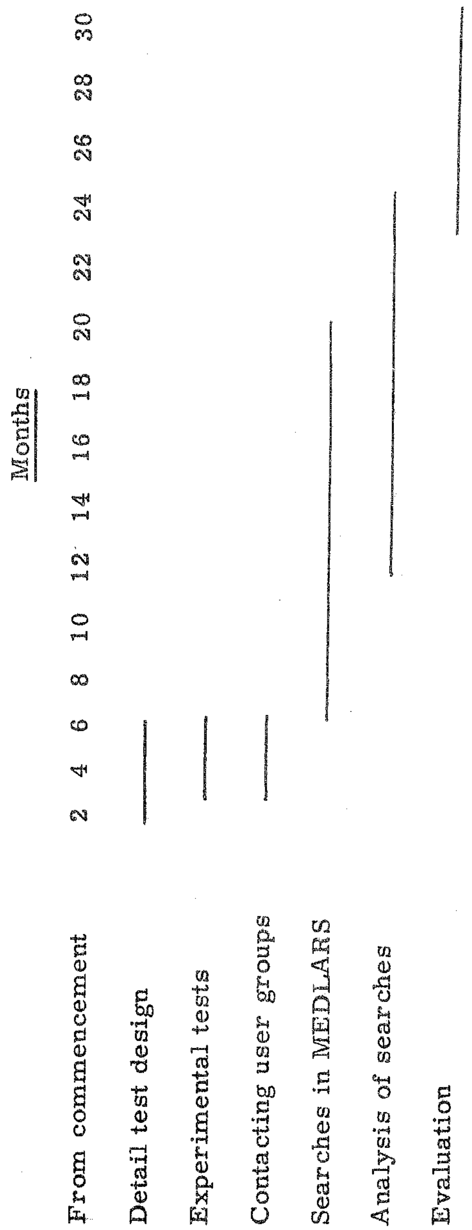
1. Varying the level of exhaustivity of indexing
2. Using indexing staff of higher or lower intellectual ability
3. Varying the average indexing time
4. Adding or deleting index language recall or precision devices
5. Using searching staff of higher or lower intellectual ability
6. Greater control of clerical operations
7. Increasing or decreasing user effort.

Ultimately the stage is reached when the system managers can take their decisions with full knowledge of the effect such decisions will have, not only in regard to user criteria but also in respect to the the economics of the system.

Ref. 1. Review of the criteria and techniques used or suggested for the evaluation of reference retrieval systems, by C. P. Bourne, Stanford Research Institute. September, 1964.

N.B. This paper was issued in November 1964 before there was an intention to carry out the present proposal, and it was quite fortuitous that the MEDLARS system was taken as an example of the argument.

APPENDIX 5


Months
From commencement
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Detail test design
Experimental tests
Contacting user groups
Searches in MEDLARS
Analysis of searches
Evaluation

APPENDIX 6

MEASUREMENT OF RETRIEVAL PERFORMANCE

The position has now been reached where a number of different groups have carried out or are carrying out evaluations of experimental information retrieval systems. In general, it can be said that most of these groups are agreed that the essential factors to be measured are recall and precision, but a number of variants have been introduced into the measures as originally presented in the Aslib-Cranfield work. There is no doubt but that in the present state of development this diversity of approach has been useful; equally there is no question but that ultimately agreement must be reached concerning performance measures if confusion is to be avoided and if progress is to be made in improvement of systems. An example of the confusion that exists is exemplified by a recently published statement that the experimental results obtained at Harvard Computation Laboratory disprove the Cranfield hypothesis concerning the inverse relationship of recall and precision. This is based on the misunderstanding that the 'normalised recall' and 'normalised precision' of Harvard are the same as 'recall' and 'precision' of Cranfield. To go back to the basic equations, the statement in fact implies that

$$\frac{100R}{C} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

and that

$$\frac{100R}{L} = \frac{\sum_{i=1}^{n} \ln i}{\sum_{i=1}^{n} \ln r_i}$$

which is manifestly incorrect, and it is therefore useless to attempt, on the basis of the original published figures, any comparison of the two tests. It is, however, essential that comparison should be made, and it has, in fact, been possible to reproduce the Cranfield and Harvard results in such a way that they are comparable.

Whereas so far each group has tended to develop different measures which serve to emphasise various facets of the performance, no attempt has been made to apply them to a common set of data, although a paper giving a theoretical analysis was published a few years back. The objective of the proposed investigation is to extend this earlier analysis by an examination of the various measures which have been used or suggested, and then to apply them to sets of experimental data. The results of this will be analysed to determine how efficiently the measures show the effect of various 'software' or 'environment' variables, and, based on this work, proposals may be made for new measures. While it is optimistic to hope that there will result immediate general agreement on the measures to be adopted, the final report will show clearly the relationship of their various measures and their effectiveness in presentation, so that discussions can take place and decisions be made with a reasonable basis of experimental evidence.

I INTRODUCTION

1. Definitions, Scope and Application

Definitions of technical terms used.
Scope of the work - to include retrieval performance only, excluding measures of cost, time, effort, etc.
Application - to the two situations of evaluating operational systems, and experimental tests. An ideal aim is the encouragement of standardisation in reporting of tests.

2. Measures of Retrieval Performance

All measures include, or are combinations of, the figures in the 2 x 2 contingency table:-

A

		Relevant	Not Relevant		
	Retrieved	a	b	a + b	
X					Y
	Not Retrieved	c	d	c + d	
		a + c	b + d	Total Collection	

B

3. A Survey of Measures in use

This would bring Swets paper (Science, 19th July, 1963) up to date, and include Fairthorne's measures, and those used by Western Reserve, etc.

II VARIABLES OCCURRING IN RETRIEVAL TESTS

The aim is to isolate all the possible variables, and to show how the effect of each of them can be displayed by the measures. The variables can be considered in two groups:-

1. System 'Software' Variables i.e. the parts of a system which affect the storage and retrieval of items,
e.g. (a) Exhaustivity and specificity of the indexing
(b) Index language type: the recall and precision devices incorporated
(c) Search rules.
These parts can not only be varied by type, but also by the amount of 'intelligence' used in them.

2. System 'Environment' Variables i.e. the conditions in which a system operates, and which affect the retrieval of items,
e.g. (a) Size of the collection
(b) Assessment of relevance made by the users
(c) Subject area, preciseness of the language.
These conditions can vary enormously, but some evidence is now becoming available on their effect.

(N.B. Variables in section 1 particularly affect $\frac{a}{a+c}$, and the position of the cutoff line X-Y.
Variables in sections 2 (a) (b) affect line A-B, while 2(c) affects both A-B and X-Y.

III THE CONVERSION OF RETRIEVAL FIGURES INTO PERFORMANCE MEASURES

1. Averaging results of sets of questions

Two averaging methods are in use:-
(a) Averaging totals
(b) Averaging averages
The claims for these, and what each represents will be examined, with examples. The need for a minimum number of sets of results for statistical validity will also be examined.

2. Format of Results and influence of Search Rules

The central problem is the CUTOFF - the point at which the distinction is made between retrieved and not retrieved items (line X-Y). The problems are the actual choice of this point, and the averaging of results where each question produces a different set of points. The case without a cutoff (e.g. the SMART system) will also be considered, as will problems involving 'abnormal' cases of performance, e.g. when $a = 0$ and $b \geqslant 1$, Fairthorne's case of $a + c = 0$, etc.

IV AN INVESTIGATION INTO METHODS OF TOTALLING SETS OF RESULTS

Different ideas to meet the aims of section II, and the problems of section III, can be suggested and tried out, using the enormous volume of figures now being obtained in the present Aslib-Cranfield Project. The opportunity is unique, since no other test has been so carefully designed and controlled. The best methods discovered will be used in publishing the report of the project's work, but fuller details can be given in this report and the scope of methods and examples broadened to include other situations and tests (e.g. The Western Reserve test figures, others done in the U.S.A. etc.).

V THE PRESENTATION OF PERFORMANCE MEASURES

Both physical display (using tables, graphs, charts, etc.) and intellectual interpretation of the measures will be examined, aiming particularly to show clearly the effect of the vcriables on system performance.