8

# Laboratory tests of manual systems

**E. Michael Keen**

## 8.1 Introduction

The essence of manual retrieval systems is that all operations of storage and retrieval are carried out by humans directly, with no more aid than the time-honoured record sheets, index cards, printed pages, and so on. Tests of in-house systems have covered several kinds of library catalogue, pre-coordinate index, and post-coordinate index. Tests of published systems have compared many styles of printed subject index. In many cases the computer is now used in the construction of both in-house and published systems, but the characteristics of the final product and its use have been affected very little, with the indexing and searching processes still manual in character. Even in systems where a computer is used in matching a formulated search against indexed documents, just as much human skill will be involved, and interactive searching requires human judgement of the highest quality. Though this chapter will concentrate on fully manual systems, some work in these semi-automated areas will be referred to.

Laboratory evaluation testing of manual systems began with the first Cranfield project[1,2]. With the special library catalogue or index in mind, the traditional index languages (Universal Decimal Classification and Alphabetical Subject Headings) were being challenged by Faceted Classification and the Uniterm system of post-coordinate indexing. So a four-way comparison was mounted. Using a realistically large document collection four test indexes were constructed along practical lines, then laboratory controls were introduced to generate the search requests, identify relevant documents, conduct the test searches, and score the search results. It can be seen now that this approach represents a mid-point in laboratory test techniques between a deeply artificial test of highly controlled subsystems and the testing of real-world systems under conditions of controlled laboratory searching. The deep laboratory approach was soon to be exemplified in the second Cranfield project[3,4], where the index language variants tested covered many linguistic forms, and where the searching was so rigorously controlled as to be machine-like or 'unintelligent'. The other extreme of test had begun already with a comparison of two systems by D. R. Swanson[5], in which one system in particular was analogous to real subject indexes and was subject to

little special control in construction. The clearest use of this technique is the Aberystwyth 'Off-shelf' test[6] where six printed indexes were taken off the library shelves and searched in a laboratory.

The existence of these rather different test techniques, yet all conducted within laboratory environments, suggests that the problem of testing manual systems is that of controlling the rioting variables without distorting them or so torturing them that they become skeletons of reality. Of course the answer to control with realism will never be found, at least to the satisfaction of the perfectionist or sceptic, but the researcher is always striving for improved ways of steering between this Scylla and Charybdis. Are manual systems more difficult than automated ones in this respect? Do manual systems contain more human and behavioural variables? It is doubtful if logical analysis would distinguish any fundamental differences, but in the area of searching in particular the sheer abundance of possibilities for variability and human error is surely greater in the evaluation testing scene of manual systems than of automated ones.

**Evaluation test validity**

What constitutes a valid evaluation test? Why can we regard everything done before Cranfield 1 (and some later work) as inadequate to answer questions of the merit of retrieval systems? Cyril Cleverdon himself regards some tests as incomplete[7], and says that as a result they do not advance the state of knowledge about information retrieval. The three things he requires of a test are that there is a collection of documents under test, a set of search requests, and some relevance decisions that identify documents relevant to the requests. These requirements need not be met in a 'real' manner: even a set of simulated documents, requests and relevance decisions could be used. Thus a valid test must involve the total environment of information retrieval even if only one small subsystem is under investigation, such as varieties in term order in printed index entries.

In addition to these three desiderata, a further three seem necessary. First, to be acceptable as a test the measures and performance criteria measured must be adequate. In particular, a measure or valid estimate of system recall needs to be made, as well as the more easily obtainable precision performance (or a substitute for it). These are necessary whether efficiency measurements are also being made or not, since measures of efficiency or cost cannot be interpreted without knowing what quality is being provided for the expenditure incurred. Another aspect of this is that a sufficiently comprehensive set of performance criteria be measured. For operational testing this means all the criteria covering effectiveness and efficiency, but in a laboratory situation a criterion such as cost would not be appropriate unless an accurate simulation was involved. System coverage and currency would also hardly apply. But 'off-shelf' testing needs a wide range of criteria, and the Aberystwyth test covered hits (recall), waste (precision), search time, search effort (page turns), presentation clarity (relevance prediction), and usability (subjective preferences of searchers). Cranfield 1 concentrated on recall, and also systematically varied indexing time, but the sample precision ratios included in the final report reveal the growing understanding there was of adequacy of effectiveness measurement. The deep laboratory test may well

only require measures of hits and waste, as Cranfield 2, so this criterion for test validity is to be related to the objectives of a given test.

A second desideratum applies to tests in which performance is being compared under different circumstances, and is that the comparison be so controlled that the cause of any performance difference may be determined. This may not always require no more than a single variable to be altered at any one time, but in the present state of information retrieval theory this is usually the safest procedure. Another aspect of this requirement is that where possible the test environment factors should be held constant; for example, a comparison of manual boolean logic search results with those of a search path from a ranked search output would be best made on one test collection using the same set of search requests and relevance decisions. It is recognized that some comparisons cannot be made without some change in environment factors, such as the comparison of general and specific requests requiring different request sets, but then there would be an advantage in keeping the document collection unchanged.

A third criterion for acceptability is the practical matter of the availability of a full report describing the test in sufficient detail. There have been suggested minimum lists of matters to be included in reporting evaluation tests but nothing has gained acceptance. If the method used for some vital part of an experiment cannot be determined, then its results are really as suspect as those from tests known to be inadequate.

## 8.2 Test types

The history of laboratory manual testing seems to consist of only a few large studies, each one looking at a number of the basic parameters that govern the behaviour of information retrieval systems. Few hypotheses have been clearly formulated, but these tests constitute a host of quite tight experiments that have given us most of the light we have on index languages, indexing and searching. Examples of tests will now be given, categorized for convenience into index language comparisons, indexing and searching experiments and printed index comparisons. The writer's own work will often be used as the main illustrations of these distinctive test types, so other studies would need to be added for a comprehensive picture. Some of the findings and conclusions of this testing activity will be given in the next section.

### Index language comparisons

Cranfield 1 remains a classic set of experiments in objectives, details and procedures[1,2]. It provided all the necessary and sufficient circumstances for testing. All subsequent tests have, knowingly or unknowingly, faced the same problems, but rarely with the common sense and ingenuity of Cyril Cleverdon. What was tested has already been briefly described. Overlapping it in time was a test of a faceted classification used manually, versus a complex semantic code and role operator system with machine searching, known as the Western Reserve University test[8]. Here Cyril Cleverdon and Jean Aitchison showed that a small test collection in laboratory search

conditions can provide much valuable data even in the realm of retrieval failure analysis. They also rode the storm caused by the unexpected and unwelcome outcome of the comparison, and made people face the possibility that complexity and intelligence at input may not result in a superior result at retrieval.

Though Cranfield 2 used machine-like search procedures in testing 29 index language devices[3,4], we may say that the findings about the effectiveness of natural language (either indexing or titles) apply to manual systems as well, unless practical considerations of file or vocabulary size inhibit. The 63 requests and 200 documents subset of the Cranfield 2 aerodynamics collection have probably become the most heavily used test collection. The ISILT experiments[9,10], as Cranfield 1, took the untested debates of the day (minimum vocabulary post-coordinate systems for example) and once again tried to provide measured results to replace unmeasured opinion.

Two large index language comparisons that utilized manual indexing and search formulation, but machine searching, were the Case–Western Reserve University test[11] and Tom Aitchison's INSPEC work[12]. Many small scale tests were carried out on the need for syntactic devices (e.g. links and roles) in index languages, and these culminated in tests of the relational indexing system carried out by Jason Farradane[13] and in ISILT.

## Indexing and searching experiments

Index language testing has dominated the main thrust of laboratory investigations in spite of the evidence of Cranfield 1 that it is the operations of indexing and searching that matter most. No large-scale laboratory experiments have tackled these two processes as primary variables, though many tests have experimented with them as secondary variables: all the large index language tests mentioned did so. Cranfield 1 is a classic in this respect. The 18 000 documents, which were indexed by four languages, were built up from batches of carefully selected components. There were different types of document (articles, reports, book sections, etc.), general or specialist subject areas, five time limits allowed for indexing, and individual performance in indexing and searching was related to level of experience and the use of subject specialists versus librarians.

Clearly defined parameters of exhaustivity and specificity as they affect both indexing and searching were explored in Cranfield 2 and ISILT. The comparison between pre- and post-coordinate search files was systematically tackled in ISILT, and the phenomenon of 'preserving the context' by multiple specific pre-coordinate entry was carried over from ISILT to the printed index experiments known as EPSILON[14–16]. There have also been numbers of smaller projects in which just one of the two processes has been studied, but most such studies on indexing quality or consistency have not reached the status of valid evaluation testing.

Turning to tests of the search process, many laboratory experiments have employed very strict controls. It is true that in operational tests manual search formulation and strategy can vary dramatically from person to person, as was clearly seen in the Medusa current awareness work[17]. One experimental method is to obtain results by progressively broadening the

searches so that a path or search curve can be plotted. This has been done not only by the use of machine-like levels of term matching, but in freely pursued manual searches by varying the stopping points the different recall needs of users can be simulated. Thus in ISILT and the printed index tests the minimum point was the need for a single highly relevant document, proceeding through intermediate points to the need for all available documents of any relevance strength.

As a final example of indexing and searching experiments we ask the question as to whether these operations perform most effectively when done manually or when they are automated in some way. Remarkably few tests of this have been done because it is such a very difficult comparison: like comparing apples and pears. Three attempts are now illustrated.

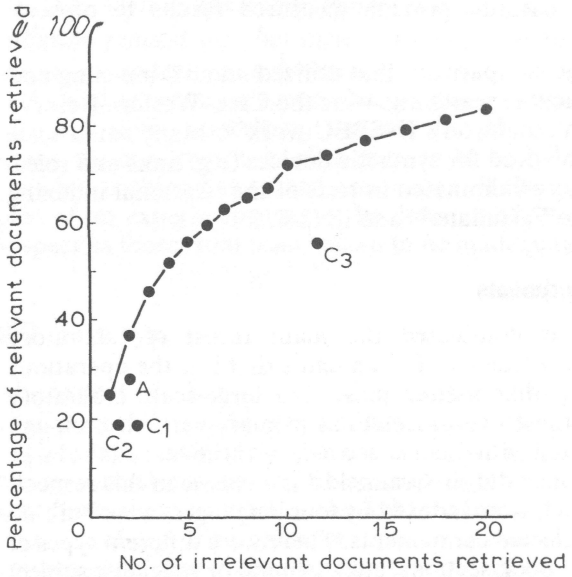


*Figure 8.1.* Retrieval comparison taken from Swanson's test[5].
• manual subject index, four results; •—• automated text retrieval

*Figure 8.1* shows D. R. Swanson's graph[5] with four results of manual searches on a subject index versus a performance curve based on full text interrogation with a thesaurus. As has been noted there was little control in this comparison: the machine result is superior. *Figure 8.2* reveals an opposite order of merit for the manual and automated comparison, and here the indexing exhaustivity was held constant on the Smart system[18]. So the automated result of Swanson may be more a reflection of higher exhaustivity than anything else. *Figure 8.3* is a hitherto unpublished result comparing ISILT manual with K. Sparck Jones automated[19]. As many of the variables as possible between the two methods were removed, the two remaining differences being: in manual the terms were slightly more specific, leading to a precision gain and a recall ceiling loss, and the definition of a subsearch (to generate the performance curve) differed a little, with an unknown effect. Here the manual system is a little better, but not at all recall levels.

If these results suggest that manual systems have lower recall capability it should be remembered that the machine searches produce high recall at such low precision levels that in practice searchers might not tolerate it. If the

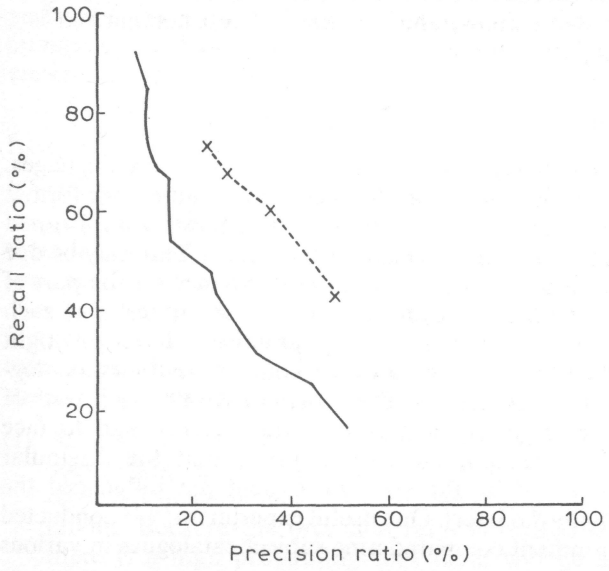


*Figure 8.2.* Retrieval comparison from SMART project, taken from Keen[18]. ×.---× manual, KWIC index to abstracts; ——— automated, abstracts with thesaurus

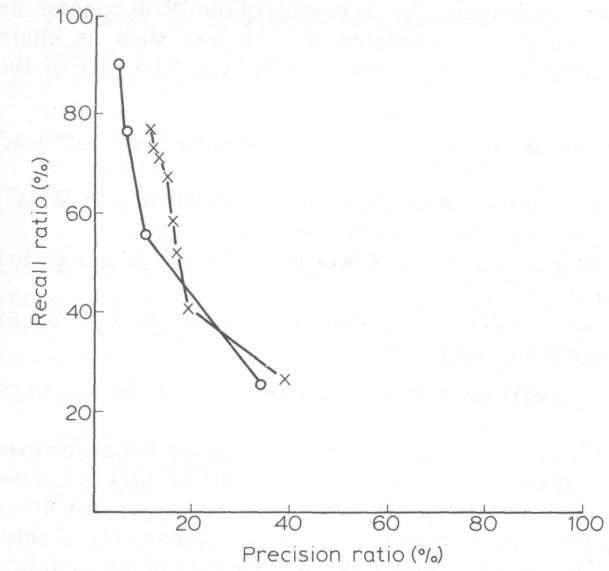


*Figure 8.3.* Retrieval comparison of ISILT and Sparck Jones[19]. ×——× manual post-coordinate, uncontrolled language, ISILT; o——o automated keyword, run Keen 800Ig T1, Sparck Jones.

superiority of the manual approach in precision at medium and low recall is a more reliable finding (at similar levels of indexing exhaustivity) it must be remembered that term weighting procedures developed more recently by K. Sparck Jones and G. Salton may well have narrowed or removed that gap. The trouble is: we just don't know, and this kind of test comparison is a major challenge to testing ingenuity.

### Printed index comparisons

Tests in this category could be regarded as experiments into index languages, indexing or searching, but they are separated out here because page format indexes seem to have been absent from experiments of these types. Printed indexes have also been late in being tackled by evaluators. This may be due to the relative satisfaction with their apparent performance on the part of their users. Or, it may be the great difficulties that face the tester of such heuristically-flexible page-scanning searching practices. Even in tight laboratory conditions the INSPEC comparison[20] admitted that methodological problems overlaid their test results. The Aberystwyth Off-shelf test[6] of published indexes covering library and information science had to face similar problems, and it was not possible to prove that the dissimilar document collections covered by the six indexes had not influenced the results unevenly (see *Figure 8.6* later). One useful experiment[21,22] conducted in an operational environment compared nine subject catalogues in various formats, including printed.

The hazards faced by a deeper laboratory test such as EPSILON[14–16] are that the care in control over the construction of the indexes may allow one index to exert undue influence on the others. However this was the only satisfactory way to tackle, once again, the arguments of the 1970s concerning the efficacy of subject indexes constructed by schemes such as chain procedure, PRECIS, articulated, and rotated (KWAC). The foci of the comparisons made were:

(1) Full versus no context, as preserved entry systems versus one with lead terms only.
(2) Direct versus indirect entry, as multiple entry rotated (e.g. KWAC) versus chain procedure (e.g. British Technology Index).
(3) Full versus partial provision of function words, as KWAC or articulated versus rotated term or PRECIS.
(4) Active versus predominantly fragmented term order, as KWAC or rotated versus articulated or PRECIS.

The findings of these comparisons will be indicated briefly in the list given later.

Tackling the printed index page-scanning mode deepened the problem of recording and analysing manual searching, and led in EPSILON to the use of audio-recording, index copy marking and a test technique involving scanning only selected index portions in order to measure accuracy in entry relevance prediction. This was the first time the criterion of presentation clarity was the subject of an experiment, and it would seem that this criterion appropriate to most information retrieval systems has been missed out of previous work.

## 8.3 The conclusions of information retrieval testing

Has all this testing activity led to a set of general conclusions about the design and operation of information retrieval systems, especially manual ones? It is perhaps no surprise that even the answer to such a question is a matter for opinion and debate. Both the content and the status of general findings are viewed in various ways.

### Laws, rules or principles?

If information retrieval is a behavioural science it is unlikely that inviolate laws await discovery. Researchers have therefore more wisely talked about there being hypotheses, rules or fundamental principles. Cyril Cleverdon[23] specified 13 hypotheses arising from Cranfield 1; Gerard Salton set out a set of rules governing automatic text analysis[24]; and Michael Keen and Jeremy Digger gave ten findings in the form of principles[9]. Cyril Cleverdon has referred to three principles he regards as fundamental[25], which may be stated in the present writer's terms as follows:

(1) As a search proceeds and retrieves an increasingly larger number of documents, so the numbers of relevant and irrelevant documents retrieved increase monotonically, as also do the measures of recall and fallout. Because the precision ratio is related to both these measures, there is a high probability that there will be an inverse relationship between recall and precision.
(2) If indexing exhaustivity is increased, so will the recall ceiling. For a given desired level of recall there is an optimum level of indexing exhaustivity: below this level recall will suffer, and above it precision will deteriorate. However, the optimum level may have a quite wide range of acceptable values[26].
(3) If indexing specificity is increased, the precision ratio rises. Specificity may be adjusted either by the semantic specificity of the index terms or the levels of term combination usable in searching. For a given desired level of precision there is an optimum level of specificity, though the range of values is not well understood.

The first of these three principles incorporates some of the important qualifications that safeguard a naive view of the recall/precision trade-off, as spelled out by Cyril Cleverdon[27], but misunderstandings and disagreement break out from time to time. The writer's view of the more detailed practical findings of manual laboratory tests adds the following ten matters:

(4) Different types of classificatory index language do not substantially differ in performance merit (Cranfield 1).
(5) Controlled index languages, such as classification, alphabetical headings and multiple entry systems (e.g. Uniterm, thesauri, etc.) differ little in performance (Cranfield 1, Cranfield 2, ISILT, Off-shelf).
(6) Index languages uncontrolled at the indexing stage do not have an inferior performance to controlled ones (Cranfield 2, ISILT).
(7) Extensive cross-references are not needed for high recall, and there is an optimum level above which precision suffers (Cranfield 1, Cranfield 2, ISILT).
(8) Syntactical devices used explicitly in searching (e.g. links, roles,

relations) as improvers of precision have a small and minority influence (ISILT, Farradane[13]).

(9) The index language vocabulary has a minor influence on performance compared with query negotiation, searching and indexing (Cranfield 1, Cranfield 2).

(10) A pre-coordinate file requires significantly more search effort and time to reach a given recall compared with a post-coordinate file (ISILT).

(11) Preservation of entry context allows significant rejection of non-relevant entries for very little recall loss (ISILT, EPSILON).

(12) Use of direct entry significantly reduces search time and effort: the indirect entry of chain procedure subject headings (as British Technology Index) has these penalties, for example (EPSILON).

(13) The varieties of function word provision and term order (e.g. in KWAC, articulated, PRECIS) perform indistinguishably (EPSILON).

It may be added that operational testing also adds its weight to these findings: for example, MEDLARS[28] bears out number (9), and WUSCS[22] bears out numbers (5), (6), (7), and (13).

**Measuring information retrieval system characteristics**

Conclusions and findings about information retrieval cannot be generally utilized unless measured relationships can be established between the variables studied and performance. For example, the best choice of indexing and index language as to term specificity—where users want a good precision ratio—needs a generalizable measure of specificity to replace the emotive 'named' index languages that usually figure in tests. A suitable measure has proved hard to find: indexing exhaustivity is a little easier, with Cranfield 2 testing five levels and showing that 33 terms per document was the best in that test environment[4]. For specificity in Cranfield 2 the first crude measure was that of vocabulary size[4], with large sizes taken to be more specific, but ignoring the influence of term use in indexing and searching that might well overlay the effect of size. A somewhat better measure was devised for the later ISILT test, where measures of specificity were related to the outcome of usage of the terms in indexing and searching, namely, measures based on size of retrieval output. But in ISILT only having three comparable index languages hardly revealed an interpolated optimum, so this approach was reapplied to the Cranfield 2 data on 29 index languages. *Figure 8.4* gives the resulting plot of specificity versus precision (taken from Keen and Digger[9]).

The connecting lines represent logical links between the different index languages: they are directions in which performance could be altered by varying the specificity of indexing or searching. Overall optimum specificity is that of language I3, single term word stems. Within the concept (phrase) languages there is a fall in precision either side of II12, simple concepts with complete species from hierarchy. This measure of specificity is not the last word on the matter, and still better measures need to be devised.

Measurement of cross-reference provision (linkage) was plotted against performance in ISILT. Search breadth also needs measuring beyond the crude use of co-ordination levels. The development of reliable and generally applicable systems characteristics measures would remove the need to test

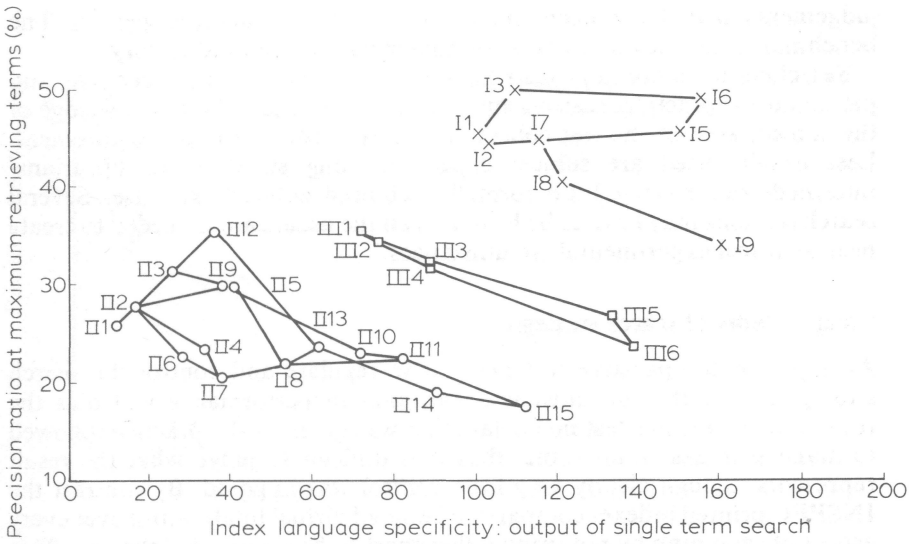


*Figure 8.4.* Cranfield 2 data re-presented from *Figure 16–16* in Keen and Digger[9].
× single terms; ○ simple concepts; □ controlled terms

variables of many values and combinations if correlations with performance could be established, so that systems having intermediate values could just be read off a suitable characteristics/performance plot.

## 8.4 Controlling searching in experiments

The kinds of experimental control needed to measure the variables in manual testing are illustrated best by the stage of searching. This stage encounters most of the problems posed at other stages and provides the severest test of an experiment. After considering the general methods of setting up the search stage of an experiment, six topics will be sampled to give the flavour of the problems and some of the current solutions.

### Conducting experiments

In an experimental situation there is the need to decide what constitutes a search and who is to perform it. In a laboratory one cannot have a real user conducting an unconstrained actual search at the time an information need arises. So the usual practice is either to obtain previously used search requests, or manufacture realistic ones, and have people other than the requesters do the searches. The searches themselves have often to be conducted in a closely prescribed manner, adopting particular procedures, recording the results in some way, and timing the process. What can never take place is a discussion of the meaning of the search request with its originator, and it is the inviolate written request that has to constitute the search query. Thus there can be no warrant to stray from the stated request, no radical alteration of it, if only because associated document relevance

judgements may have been made on its basis by another person. The benchmarks for relevance and searching cannot be allowed to vary.

Searchers for laboratory tests have usually to be specially recruited and paid, and for practical reasons students are often used. Some knowledge of the subject area of the test collection and requests is a usual requirement. Less usually used are subject experts of long standing, or librarians/intermediaries having little formally acquired subject expertise. Several search sessions may have to be held, and all the usual care is needed to create near identical experimental circumstances.

### Comparability of search strategies

All types of comparative test need to so regulate and control the search strategies used that no unwanted variation in performance will bias the results. In Swanson's test no explanation was given of the practice followed to make searches comparable, thus it is difficult to judge what the result represents, though clearly very little control was imposed. By contrast the INSPEC printed indexes comparison adopted virtual total control over every aspect of searching by requiring a flowchart to be rigorously followed. This led to problems, however, with the flowchart suiting one of the indexes better than the others, so failing to achieve the intended neutrality. Another problem was that since such artificial search procedures were used realistic times could not be obtained and the attempt to use standard times was not successful. Another test using fixed strategy methods was that of Case Western Reserve University, where searches designed for use on reasonably exhaustive indexing were applied also to titles, thus failing to match any documents at all with large numbers of the queries and giving very low recall indeed. In fact the searchers sometimes compensated for title searching by dropping terms from the formulation, but these were spotted as contrary to the rules. This result was therefore no realistic test of title searching at all.

Cranfield 1 saw clearly the search strategy problem and tackled it in several ways. The first round of testing encountered the problem of how long a searcher was justified in continuing a search when the one relevant source document was known to be in the file somewhere. Also, since scoring was to include the number of subsearches required to find the source document, the problem was to decide exactly what constituted a different subsearch, particularly the different kinds of entries in the four index languages. So search round two prescribed a limit beyond which the search strategy could not be broadened, and also defined a subsearch both in general and in terms of each system in as fair a manner as possible. Although this gave results that were taken to be quite satisfactory for the main variables under test, an analysis of system failures revealed cases where a search on one system had succeeded but the same query on another system had failed due to search formulation. Also, it was said 'it appeared often a matter of chance whether the correct programme (subsearch) was used on the first or fifth searches'. So the third round of searching was designed to 'eliminate as far as possible the variable of searching'[2] by adopting a standardized and fixed strategy for all four systems. This was done by making an initial free-mode search always on one of the systems as the yardstick, then applying the strategy in an appropriate identical manner to the other three. If one of these later systems

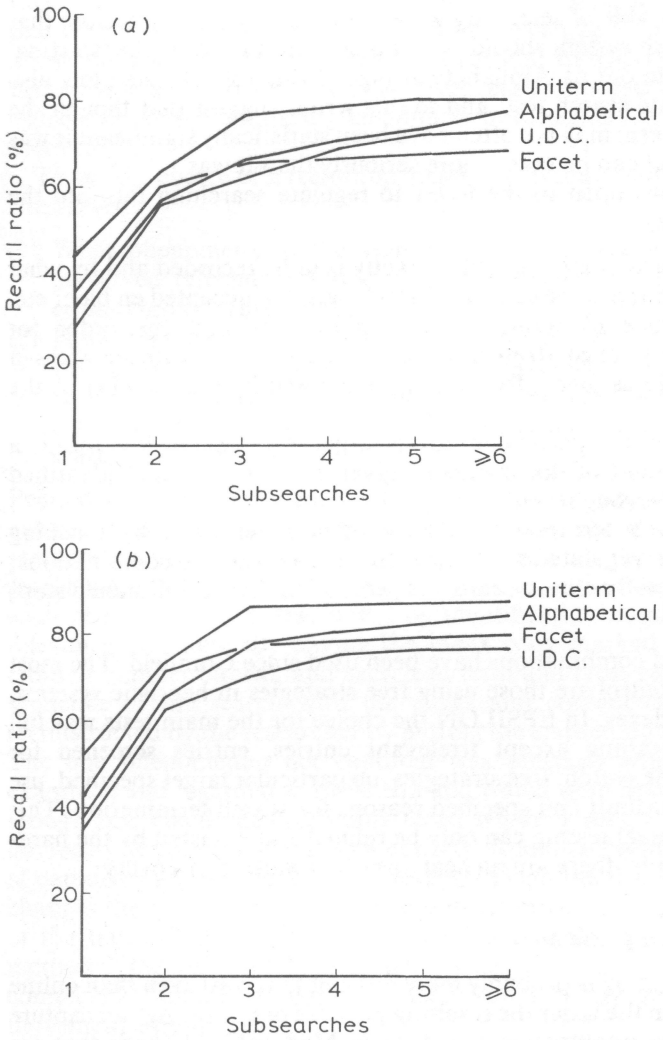


*Figure 8.5.* Cranfield 1 results from two search rounds, calculated from *Tables 3.6*, *4.8*, and *4.9* in Cleverdon[2]. (*a*) Search round two, (*b*) search round three

prompted an additional subsearch then this was allowed and used to create a modified version of all earlier searches.

The results of round three were taken to be very similar to round two, and in fact the similarity is most striking when the new presentation of *Figures 8.5*(*a*) and *8.5*(*b*) is compared, showing how recall improved with each new subsearch. The relative positions of Uniterm, Alphabetical and UDC are unchanged by search type: only Facet on round three performs a little better in not dropping so sharply at the third subsearch—not a surprising difference as the free-mode searches of round two often encountered too many chain index entries for pleasure and round three's fixed strategy rules would force such a search inexorably on to the end. But these plots strongly suggest that

fears about the variable of searching were unfounded. Even the feeling that for round three each system should have taken it in turn to be the 'starting' system is not borne out as Alphabetical played this role. These plots also represent a replicate search test, and to this writer suggest that though the superiority of Uniterm may not often have been statistically significant it was a reliable result that can be taken more seriously than it was.

The main options open to the tester to regulate searching fall into the following four areas:

(1) Procedural instructions, e.g. what exactly is to be recorded and by what method; are entries to be screened for relevance or accepted en bloc; etc.
(2) Rules for choice of terms, their combinations and the order for subsearches, e.g. fixed strategies used on all systems, with or without retrospective adjustment; free strategies devised by the searcher at the time; etc.
(3) Specification of any particular search target, e.g. high or low recall; a prescribed amount of documents of given relevance; actual identified documents to be sought; etc.
(4) Rules for search termination, e.g. whether determined by reaching specified target; regulated by the rules for term choice and combinations; use of a time cutoff or limit; searchers perception that search should stop; etc.

Many methods and combinations have been used since Cranfield. The most difficult cases to control are those using free strategies in heuristic systems, such as printed indexes. In EPSILON the choice for the main tests was full recording of everything except irrelevant entries, entries screened for relevance during the search, free strategies, no particular target specified, use of time as an upper limit and specified reasons for search termination. This crucial area of manual testing can only be refined and adjusted by the hard-slog of trial and error: there are no neat answers awaiting discovery.

### The search recording problem

Printed index searching is probably more difficult to record even than online interaction, since in the latter the resulting printout or search log can capture most of the process unobtrusively as done in Medusa[17]. Methods that are disturbing to the searcher are probably unavoidable, corresponding to the Heisenberg principle, and awareness of observation may cause a kind of Hawthorne effect. Some of the different methods for recording freely-devised strategies are:

(1) Searcher compiled record sheets, varying in detail.
(2) Searcher conducted marking of the index copy.
(3) Searcher verbalizing using audio recording.
(4) Observation by a second party, either person or camera.
(5) Searcher retracing progress in a post-search interview.

The progress of a search against time is often needed, requiring a stop-clock or timing device, or its derivation from audio or camera records. Torr, Fried and Prevel[29] concluded that a time-lapse camera was best for field testing of printed indexes, but apart from having to sit under a strongly lighted box

there was the need to trace out what was being read by a moving finger so that the film gave more than a head and an open page! All but the last of the above methods were used in the Off-shelf and EPSILON tests. Darkroom timers, which ticked obtrusively and were easily misread were soon replaced by a continuously running digital minutes and seconds display via a television monitor.

To aid the choice of a recording method, two central questions have to be considered:

(1) What phenomena in the search need to be recorded? E.g. citations perceived relevant; citations perceived irrelevant; index terms tried; cross-references used; index pages consulted; etc.
(2) How much needs to be known about each of the recorded phenomena? For example just how many of each, or actual identity of each one; searchers judgement about how relevant each one was; individual time for each one; ability to reconstruct the exact order of each event in search; etc.

Printed index marking with a time record and simple record sheet can achieve most of this. In EPSILON the text of the query, space for any searcher's notes, start and finish time, and reasons for termination of the search were the basic data on the record sheet, and as the search proceeded each relevant citation was noted by identification number, judgement of relevance and time. By having the index copy marked with the relevant citation number circled, and each lead term and cross-reference timed, the sequence of the search with the pages and content consulted could be reconstructed by the researcher by putting the elapsed times back into order. The one phenomenon not captured was precisely which citations were examined and regarded as irrelevant and which were never examined at all, though one could identify many cases where a set of index entries had obviously been examined in their entirety. Audio-recording has the potential of capturing all the phenomena needed, though the fear is that verbalizing changes the pattern of search and upsets its progress against time. Analysis of the tapes is also a problem, though in the one EPSILON use of this method[15] the searcher herself made transcripts after the searches were completed. Similar problems are likely to face any attempt to use eye-movement equipment.

## Search performance criteria and measures

Laboratory manual tests seem to have concentrated on measuring recall, precision, time and effort. There has been much debate about the mathematical properties of measures, and little recognition that even the matters of computation, aggregation and presentation can cause large differences[30]. A good example of the care needed in choosing a valid measure of a given criterion is the use of the precision ratio in testing browsable-heuristic systems.

In iterative systems where a stack of document entries is retrieved in toto the precision ratio is straightforward to calculate and is quite meaningful. But in the Off-shelf test it was difficult to get the searchers to spend the time accurately recording all the irrelevant citations they encountered, as they

were free to select and reject entries individually and did not have to accept every entry a given heading led to.

The resulting 'selected' precision ratios not unexpectedly favoured the one index that did not contain abstracts because swift title scanning was unrecordable. The remaining five indexes had remarkably similar precision levels. The INSPEC printed index testers chose recall and time as their main measures, so it can be suggested that time is a suitable replacement for precision in these circumstances.

A plot from the Off-shelf test[16] of relevant selected against time is given in *Figure 8.6*: we may regard this plot as the printed index testers equivalent of the 'Cranfield' recall/precision plot.

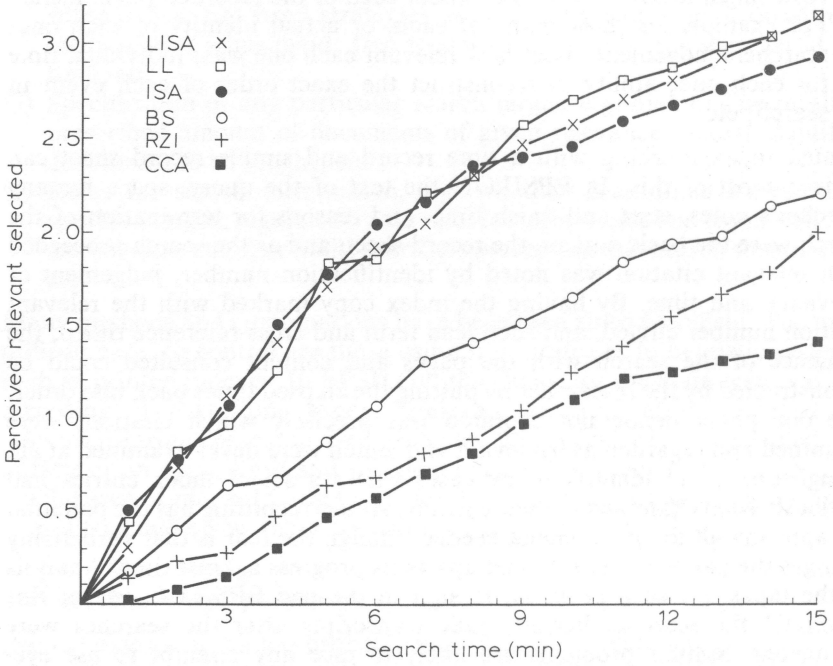


*Figure 8.6.* Results of information science searches of six indexes in the Aberystwyth off-shelf test, taken from *Figure 4.2* in Keen[16]

However it could be argued that the criterion of *search time* is conceptually distinct from *non-relevant entries*, and it is the recording process that fails: a longer time spent on an index search may not mean more irrelevant entries are encountered. So better methods were used in EPSILON to measure both precision and time, but again quite different index types exhibited similar precision results. Specifically the selected precision results of the comparable set of five Off-shelf indexes fell in the range 43–52 per cent, in EPSILON search tests 44–52 per cent and in EPSILON scanning tests 44–54 per cent. The suggested explanation is that there is a natural level of precision where searchers' tolerance for examining irrelevant is the governing factor, whatever the system type. This kind of result was also seen in WUSCS, with precision 43–48 per cent[22], but may not be limited to heuristic systems, as

one SDI experiment found the precision ratio to be maintained at 45 per cent even throughout substantial strategy changes (Medusa[17]).

Another severe performance measurement problem that is illustrated by the comparison in *Figure 8.6* is the difficulty of validly comparing what are, in this case, different test collections by virtue of there being different indexes. The question is whether the measures used minimize the unavoidable bias caused by the coverage of relevant items differing in each index. The hypothesis could be advanced that were there to be added an index of the same retrieval efficacy as the best ones, but containing a smaller number of relevant items, its performance curve would follow the best ones to begin with but fall away as soon as its recall rose. However, the counter hypothesis would be that paucity of relevant entries might make retrieval a more spread out affair, particularly if several index issues had to be consulted, thus giving a worse curve from the beginning on a plot of this type that uses elapsed time. So, a satisfactory way of comparing dissimilar systems has yet to be found.

### Human performance and preferences

It must have been quite a shock to the staff of Cranfield 1 to see their names heading the columns of a results table in which their performance (as indexers in this case) was open to public view. However there were no traces either of statistically significant or practically important differences: having survived the indexing of 18 000 documents the staff could probably index in their sleep! It is important to realize that this result didn't suggest that humans are consistent in every detailed decision, but that when judged by average performance outcome (surely the only test that matters) there were no real differences. The measurement of inter- and intra-person consistency in indexing has been a plague and a nuisance because it has been divorced from search outcome yet has been used to indicate quality and even performance.

An example of the validity of taking measured performance as the criterion of consistency was seen in an ISILT test of inter-searcher consistency. If to be consistent two searchers had to have the same search terms, combinations and subsearch order, then the average result would have been 0 per cent. If terms and combinations had to agree, but subsearch order need not, consistency would have been 13 per cent. With only term choice as the criterion (any combinations or order) the level would have risen to 32 per cent, a very similar level to many inter-indexer consistency results. But in searching, the identity of the search terms is less important than the outcome: the same documents can be retrieved by different terms. So, with retrieval of identical documents as the criterion consistency rose to 64 per cent. Still recalculating the same data, one could say that document identity is not as important as amount: and if this were the criterion consistency finally reached a high level of 81 per cent.

Individual searcher performance is important in laboratory manual comparisons of indexes when each searcher sees each index, but cannot be asked to repeat the same search request. In the Off-shelf experiment with six searchers and six indexes the variation in performance of the people was less than that of the indexes thus giving grounds for hope that skill had not overlaid the main variable being studied. Coping with this problem is a part of valid experimental design and statistics, an area rather neglected so far.

Gathering searcher preferences after exposure to a number of systems has been practised by researchers. One technique is to pose questions on criteria that are in fact being measured, to correlate subjective responses with more objective results. This has yet to be done accurately on the basis of individual responses and scores, rather than averages. This useful technique may eventually be used to see to what extent people's perception of performance matches with reality.

### Design of experiments

Free strategy searching on several systems faces the problems of many scientific experiments. The approach usually employed in information retrieval has been:

(1) Every request in the set must be searched against every system an equal number of times.
(2) No searcher must process any request more than once.
(3) Each searcher must conduct an equal number of searches on each system in a balanced manner during the test.

This has led to the use of Latin square designs: Cranfield 1 adopted this approach for search round one though it was not described as such and in practice there were only three searchers for the $4 \times 4$ square, with one person repeating the same requests after at least a one month time interval. This careful approach led to a comprehensive statistical appendix which is frequently overlooked[2]. Recent tests have used a similar approach and have looked for statistical significance using non-parametric tests such as the Sign test and Wilcoxon's Signed Ranks test.

The practicalities of conducting such experiments include the usual warming-up operations to minimize the learning effect. But more marked than this effect has been an end of session mixture of perfectionism and fatigue. In the Off-shelf test the last search received an increased time and resulted in less entries retrieved: a clear indication for future experiments to include one or two dummy searches with which to terminate. A more serious problem is that the use of within-subjects designs cannot avoid some carry-over effect, thus lessening the real differences in the systems measurements. Separate-subjects designs are often used in experimental psychology, so information retrieval researchers need to be more adventurous in this area. EPSILON made a special study of problems of design and statistics[31], and discussion of these matters is to be found in Chapter 5.

### Search diagnosis techniques

Post-search analyses of reasons for performance obtained are a vital part of operational testing and have also proved useful in laboratory testing. Analyses have usually been confined to searches in which failures occurred, divided into recall and precision failures. Success analyses might be an enlightening addition.

Cranfield 1 conducted many analyses of recall failures, and overall results showed 22 per cent due to searching, 67 per cent to indexing and 11 per cent to the index languages. Searching had a larger share of failures in analyses

conducted in other tests, often from 45 to 60 per cent. Only the Cranfield analyses revealed the extent to which failure is avoidable or not: 51 per cent of them were, and of these 22 per cent were due to lack of time allowed in indexing and 17 per cent due to question misunderstanding. The effort and subjectivity of conducting failure analyses are the problems. Too many laboratory tests have had to spend most of their time building their test indexes, and have had to cut back on search testing, and even miss out diagnostics altogether. Assignment of reasons to cases of failure can be complex, but the use of multiple reasons seems a reasonable technique. No wonder operational testing is a rarity, when even laboratory work poses such severe practical difficulties.

## 8.5 Test reliability and the future

As a new decade of evaluation testing is reached, a candid look at the past is needed. Tests have not yet covered all the variables we know about, let alone those remaining undisclosed. We don't know enough about effects of experimental scale, or the continued use of dated test materials. We do so often set up an investigation or experiment first, then afterwards explore the variables or pose the hypotheses. But, on the other hand, we now have plenty of test evidence to argue about, and we do have a clearer view of the design parameters in information retrieval.

It could be argued that only the scientific purist could expect a cleaner state of affairs, and such lack of progress is by no means confined to our own field. Information retrieval testing has frequently proceeded in a series of loosely linked investigations, with the inconclusive end to one piece of research providing the impetus for the next. For example, *Table 8.1* presents a set of results that are not understood and are a current anomaly: three test methods (A, B and C) were used to look at printed index entry processing speed. Methods A and B provide similar results: the four different entry types, though they have considerably different lengths of entry, are processed at very similar speeds. But why does method C conflict? Why don't the indexes with the longer entries take longer to process? Why don't the entries that prompt greater amounts of grammatical transformation take longer to process? Hopefully, future work will explore and eventually explain these anomalies.

**TABLE 8.1. Results of three tests from EPSILON taken from Tables C/2, C/4 and C/6 in Keen[16]**

| *Index* | *Entry length (terms)* | *Entries per minute* | | | *Entries grammatically transformed* |
|---|---|---|---|---|---|
| | | *Total search* | *Fully processed subset* | | |
| | | *A* | *A* | *B* | *C* |
| Rotated term | 4.6 | 3.50 | NA | 7.29 | 47% |
| Rotated string | 7.9 | 3.44 | 10.35 | 7.95 | 4% |
| Articulated prepositional | 6.9 | 3.41 | 11.20 | 8.04 | 32% |
| Shunted relational | 4.6 | 3.29 | 9.81 | 8.17 | 4% |

A: Search test; B: Scanning test; C: Audio test. All data are arithmetic means. NA: Not available.

Our handling of the human subject in manual systems testing may well have been over-cautious in the care with which groups of different people have been used as indexers, searchers, relevance judges, and so on. The idea of using the researchers in some of these roles has been strenuously avoided. We have argued strongly against the attempt to compare different retrieval systems without the rigour of laboratory control, but the semi-operational off-shelf approach may be valid. Perhaps the greatest rigidity in thinking has been that one well-conducted experiment settles both the issues the experiment was designed to investigate and queries about methodology. Experiments do need to return even to the fundamental parameters of exhaustivity and specificity so that understanding may be deepened and non-trivial design equations advanced. The 'single experiment' mentality fails to demonstrate repeatability, and the effort of replication should not now be an option. Manual testing has achieved much—now is not the time to stop.

## References


1. CLEVERDON, C. W. *Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems*, First Aslib Cranfield Project, College of Aeronautics, Cranfield (1960)
2. CLEVERDON, C. W. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, First Aslib Cranfield Project, College of Aeronautics, Cranfield (1962)
3. CLEVERDON, C. W., MILLS, J. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems*, 2 Vols, Second Aslib Cranfield Project, College of Aeronautics, Cranfield (1966)
4. CLEVERDON, C. W. The Cranfield tests on index language devices, *Aslib Proceedings* **19**, 173–194 (1967)
5. SWANSON, D. R. Interrogating a computer in natural language. In: *Information Processing* 62, Proceedings of IFIP Congress 1962, (Ed. C. M. Popplewell), North-Holland, Amsterdam (1963)
6. KEEN, E. M. A retrieval comparison of six published indexes, *UNESCO Bulletin for Libraries* **30**, 26–36 (1976)
7. CLEVERDON, C. W. Evaluation tests of information retrieval systems, *Journal of Documentation* **26**, 55–67 (1970)
8. AITCHISON, J. and CLEVERDON, C. W. *A Report on a Test of the Index of Metallurgical Literature of Western Reserve University*, First Aslib Cranfield Project, College of Aeronautics, Cranfield (1963)
9. KEEN, E. M. and DIGGER, J. A. *Report of an Information Science Index Language Test*, 2 Vols, College of Librarianship Wales, Aberystwyth (1972)
10. KEEN, E. M. The Aberystwyth index languages test, *Journal of Documentation* **29**, 1–35 (1973)
11. SARACEVIC, T. *et al. An Inquiry into Testing of Information Retrieval Systems*, 3 Vols, Comparative Systems Laboratory, Centre for Documentation and Communication Research, Case Western Reserve University, (1968)
12. AITCHISON, T. M. *et al. Comparative Evaluation of Index Languages, Part 1: Design; Part II: Results*, Reports R70/1 and R70/2, INSPEC, Institution of Electrical Engineers, London (1969, 1970)
13. FARRADANE, J. *et al. Research on Relational Indexing*, Final Condensed Report to OSTI, City University, London (1968)
14. KEEN, E. M. On the generation and searching of entries in printed subject indexes, *Journal of Documentation* **33**, 15–45 (1977)
15. KEEN, E. M. On the processing of printed subject index entries during searching, *Journal of Documentation* **33**, 266–276 (1977)
16. KEEN, E. M. *On the Performance of Nine Printed Subject Index Entry Types*, A Selective Report of EPSILON, College of Librarianship Wales, Aberystwyth (1978)
17. BARRACLOUGH, E. D. *et al. The Medusa Current Awareness Experiment*, Computing Laboratory University of Newcastle upon Tyne (1975)
18. KEEN, E. M. An analysis of the documentation requests. In: Report ISR-13, Section X, Department of Computer Science, Cornell University (1967)

19. SPARCK JONES, K. and BATES, R. G. *Research on Automatic Indexing 1974–1976*, 2 Vols, British Library Research and Development Report 5464, Computer Laboratory, University of Cambridge (1977)
20. AITCHISON, T. M., LAVELLE, K. H. and HALL, A. M. *Laboratory Evaluation of Printed Subject Indexes, Part 1: Design and Methodology; Part 2: Results and Discussion of Methodology*, Reports R70/5, R73/17, INSPEC, Institution of Electrical Engineers, London (1970, 1973)
21. HUNT, R. *et al. PRECIS, LCSH and KWOC: Report of a Research Project Designed to Examine the Applicability of PRECIS to the Subject Catalogue of an Academic Library*, 4 parts, University of Wollongong, New South Wales (1978)
22. KEEN, E. M. Review of HUNT, R. *et al.*, The Wollongong University Subject Catalogue Study (WUSCS), *Journal of Documentation* **34**, 356–357 (1978)
23. CLEVERDON, C. W. The Cranfield hypotheses, *Library Quarterly* **35**, 121–124 (1965)
24. LESK, M. E. and SALTON, G. *Design Criteria for Automatic Information Systems*, Report ISR-11, Section VI, Cornell University, Department of Computer Science (1966)
25. CLEVERDON, C. W. Information and its retrieval, *Aslib Proceedings* **22**, 538–549 (1970)
26. SPARCK JONES, K. Does indexing exhaustivity matter?, *Journal of the American Society for Information Science* **24**, 313–316 (1973)
27. CLEVERDON, C. W. On the inverse relationship of recall and precision, *Journal of Documentation* **28**, 195–201 (1972)
28. LANCASTER, F. W *Evaluation of the MEDLARS Demand Search Service*, National Library of Medicine, Bethesda, Maryland (1968)
29. TORR, D. V., FRIED, C. and PREVEL, J. J. *Program of Studies on the Use of Published Indexes*, 2 parts, General Electric Company, Information Systems Operation, PB169416 and PB169417 (1966)
30. SPARCK JONES, K. Performance averaging for recall and precision, *Journal of Informatics* **2**, 95–105 (1978)
31. KEEN, E. M. and WHEATLEY, A. Statistics and the EPSILON tests. In: *Evaluation of Printed Subject Indexes by Laboratory Investigation*, British Library Research and Development Report 5454, College of Librarianship Wales, Aberystwyth (1978)