4 EVALUATION

4.1 Introduction

This chapter consists of a discussion of some methods for evaluating information retrieval systems, leading to a description of the methods which were used for evaluating the systems described in the preceding chapter.

Section 4.2 gives an example of the type of formal experiment which was often used in comparisons of traditional "mechanistic" systems and also discusses the use of such formal tests on interactive systems. In Section 4.3 real users are introduced, and there is a discussion of experimental comparisons of online catalogues in use. Section 4.4 compares live and formal experiments with highly interactive systems. The remaining sections (4.5 - 4.9) describe the aims and methods of the present evaluation.

4.2 Formal experiments

Van Rijsbergen, in his book on information retrieval [VAN79, p3], states that "'real world' IR systems are evaluated in terms of 'user satisfaction' and the price the user is willing to pay for the service". Van Rijsbergen makes a firm distinction between experimental and operational systems, and discusses the formal evaluation of experimental systems by "comparing the retrieval experiments with standards specially constructed for the purpose". One justification of such formal evaluation is that it does give fairly "hard" results, enabling objective system comparisons. In any case, Van Rijsbergen was writing at a time when retrieval was usually rather mechanistic compared with the systems under consideration here. The response of a system to a query or search statement would be a set, ordered or unordered, of references, and the different sets and orderings resulting from different treatments of the search could fairly readily be compared. If we regard the search statement as given, the user only comes into it, if at all, as the person who assesses the quality of the output list of references. (This is something of an oversimplification: in the evaluation of a system using relevance feedback there may be a number of iterations, between which the output of the previous stage is assessed for relevance, but the process is still essentially mechanistic because there are predetermined rules about the number of references to be assessed, etc).

4.2.1 *An example*

An example of this formal approach is an experiment described by Van Rijsbergen and others [VAN81]. Their experiment is related to the present work because it was concerned with finding good methods of automatically selecting additional search terms for relevance feedback, given the terms in the query. Three test collections were used. For each collection there was a set of queries, for each of which all the relevant documents were known. The collections were rather small (1400 - 16000 documents) and covered fairly narrow subject areas. The mean number of relevant documents for each query ranged from 7.2 to 43.8. Results were given as mean precision at standard recall levels (10%, 20%, ...). Relevant documents from the 10 or 20 best matching documents retrieved using a simple coordination level strategy were used to assign weights both to the query terms and to terms which were closely associated with them (in the collection as a whole). A second search was then performed, and precision and recall calculated for the resulting set, after removing documents already retrieved in the initial search.

4.2.2 Formal experiments with interactive systems

It is possible to perform formal experiments even on highly interactive systems by constructing deterministic procedures for the execution of searches, followed by repetition of searches and examination of the retrieved records. This is best illustrated by an example.

A procedure for searching the full system described in Chapter 3 might run as follows:

(1) Type in search statement
(2) If no records are retrieved, stop.
(3) Look at each record on the first brief display screen in full.
For each one, if it is relevant, choose the "books classified near" option.
On the first screen of records classified near, look at each one not already chosen and choose or reject it.
(4) If no records have been chosen, stop; otherwise select the "more" option.
(5) Repeat from (2) to (4) until the procedure terminates.

(This procedure could be used on the qe system if the actions involving the "classified near" option are omitted, and on the dumb system by reducing it to steps (1), (2) and the first line of (3).)

Such an experiment requires the collection of suitable search statements, preferably from live use of a suitable installation (preferably one which caters for the desired type of users, accesses a similar database and prompts for input in a similar way). These are then executed by experimenters on the systems under test.

Tests of this type have the great advantage that they give well defined and repeatable results. Drawbacks include the fact that it is unlikely that any formal search procedure will accurately reflect the behaviour of a substantial proportion of real users. If the systems have been in live use examination of transaction logs may show that many searches do fall into one or a few formal patterns, suggesting repetition procedures of the type given above. This becomes less likely as systems become more complex. For systems using relevance feedback there is also the disadvantage that the experimenters have to make relevance judgments. It is quite often difficult to guess what sort of thing the original searcher was looking for. Richard Jones, in [WALK87b, Appendix 4], refers to a search for "sterling" where the user described his subject as "economics, sterling shares and gold" and appeared to be looking for items on the effect of sterling on shares. This difficulty is not of course resolved by using subject experts as repeaters of searches. It is clear that many searches cannot be realistically repeated except by the person who verbalized the search statement. However, it is possible to select a subset of real searches where the language is sufficiently public for the searches to be repeated quite confidently by experimenters (e.g. "care of the terminally ill", "abelian groups", "file management under CP/M"), and repetition experiments must be restricted to searches like these.

4.3 Experiments with real users

Many writers would agree with Van Rijsbergen that the most significant measure of the performance of an interactive IR system is the extent or degree of user satisfaction. While user satisfaction may be a *necessary* condition for a successful search it is by no means a *sufficient* one. One reason for this is that a system may delude users into thinking that there is nothing in the collection which satisfies their needs, or that they have found the most appropriate available material when this is not in fact the case. Many people would agree that it is a fairly serious matter for a user to go away from the catalogue wrongly believing that the library does not have a copy of the sought work. At the same time it is certainly not true that users always need exhaustive searches.

In the case of online catalogue evaluation it is not very useful to measure the proportion of satisfied users, because people tend to be satisfied with the system they are currently familiar with, however good or bad it may seem to an outsider. In the Council on Library Resources (CLR) Survey [MARK83] one of the questions users were asked was "In relation to what I was looking for, the search was very satisfactory / somewhat satisfactory / somewhat unsatisfactory / very unsatisfactory". Although the systems (and collections) whose users were questioned varied enormously 75% or more of users chose the "very satisfactory" or "somewhat satisfactory" responses. Similar results have been obtained by Walker and Jones [WALK87b] and others.

However, when apparently satisfied users are questioned in more detail it often emerges that they are aware of both functional and interactional shortcomings. Tagliacozzo [TAGL77] presented an analysis of the data obtained in a follow-up questionnaire sent to a sample of Medline users who had requested searches. A comparison of different items of the questionnaire revealed contradictions between an overall appraisal of the service and more specific responses on the outcome of the search. She advised caution in inferring the satisfaction of information needs from the users' evaluation of an information retrieval system.

When users of the Library of Congress online catalogues SCORPIO and MUMS were questioned during the CLR survey the reported satisfaction rate was 88%, but it nevertheless appeared that many people were experiencing considerable difficulties [ANDE82]. Indeed, there were not many features of the systems which individually elicited positive responses. The principal interviewer said "Whether patrons used the OPAC twice a week or twice a year, they found it necessary to relearn".

4.3.1 The NLM online catalogue comparison

In 1982 Siegel and others at the National Library of Medicine (NLM) conducted a comparative evaluation of two prototype online catalogues, CITE (2.4.2) and ILS, within the same environment [SIEG82, SIEG84]. Their methodology has strongly influenced recent comparisons of interactive retrieval systems, that of the Dewey Decimal Classification Online Project [MARK86] for example.

As well as a technical evaluation and feature analysis, the NLM experiments included (1) a questionnaire user survey covering searching requirements, demography and satisfaction with the systems; (2) a partially controlled Comparison Search experiment in which patrons conducted the same search of their own choosing successively on each of the systems and answered a number of questions relating to search outcome; and (3) a controlled Sample Search experiment, in which a panel of NLM library and other professional staff conducted matched but different searches on each system. The User Survey was a self-administered 60-question questionnaire, requiring 15 minutes for completion. It was given to everybody who did an online catalogue search during the test period. The compliance rate was over 80%. This questionnaire was a modified version of the "user" one from the CLR survey [MARK83].

The Sample Search experiment did not suffer from the Comparison experiment's obvious transfer effect, but it was in several ways less "realistic": the subjects could not be considered typical users, they may have had little knowledge of the topics for which they searched, and in any case their motivation was very different from that of users with a real information need. Six hundred people took part in the user survey, 60 in the Comparison Experiment and 20 in the Sample Search experiment. Data collection took 5 months.

Sample Search subjects (staff) each did 14 paired queries of 6 search types. The queries were matched for query type (author, subject, etc), level of search difficulty and expected retrieval size. Each search was timed, then questions were asked by the experimenter and the user was asked to assess the retrieved records on printed lists after each pair of matched searches. At the end there were some general questions and an open ended interview. In the Comparison Experiment patrons entering the catalogue area were asked to take part. The refusal rate was 25%. The experimenters (the same two people who carried out the Sample Search experiment) monitored the searches and recorded them on printouts for subsequent analysis. The experimenters decided when subjects should switch systems, and conducted brief post-search interviews using a subset of the Sample Search questions. The time for each subject, including a short introduction to each system, was less than 30 minutes.

The data gathered from the Comparison and Sample experiments in the NLM comparison was perhaps not analysed as thoroughly as it might have been, because CITE was markedly superior to ILS in subject searching (in the User Survey, 55% of 220 CITE subject searchers reported that the search was very satisfactory, 30% of 180 ILS users). Most users were rather satisfied with both systems.

In the Dewey Decimal Classification Online Project [MARK86], where the researchers used very similar Comparison and Sample search experiments to compare two catalogue systems which were much more similar to each other than the NLM systems, measures of recall and precision were calculated. Precision was calculated in the usual way as the proportion of the records retrieved (and displayed) in a search which are relevant. No attempt was made to estimate the true recall of a search, but the relative recall for a search on a system (misleadingly described as "estimated recall") was defined as the ratio of the number of relevant records obtained in the search of the system to the total number of relevant records obtained in a search for this topic on each of the two systems (either by one subject or by two different subjects). The time spent by the subject doing a search was also recorded.

4.4 Evaluation techniques for highly interactive systems

It is clear that it is not at all easy to find evaluation methods for interactive retrieval systems which are both realistic and meaningful. At one extreme there is evaluation which is carried out under conditions which are more or less equivalent to "natural" use of a retrieval system, a library catalogue in the present investigation. For such a live evaluation one would set up the systems in a library environment and use log analysis and perhaps an online or administered questionnaire. This method was used in the evaluation of two previous Okapi systems [MITE85b, WALK87b]. On one occasion it was supplemented by repetition of logged searches on different systems by the experimenters. There is complete lack of control over search topics as well as general conditions such as user attitudes and needs, so a large amount of data may be needed if significant conclusions are to be drawn.

The method of collecting data from live use followed by repetition of the real searches on other systems was referred to in the preceding paragraph. If the systems being compared are operationally very similar to each other this procedure has much to recommend it, at least as a supplement to tests under realistic conditions. This would apply, for example, to two systems which differ only in the indexing, or in the way they process the users' search statements. The method may also be used, with caution, in the comparison of a system which possesses additional facilities or more advanced interaction with a more basic one where there is a single fairly well defined procedure for executing a search. For such a comparison live data is collected from use of the more elaborate system. The same search statements are then executed by the experimenter, or by a computer program, on the simple system, and the results - the records retrieved - compared. This method could be used in evaluating the present systems, by collecting live data from use of the full and qe systems followed by repetition of the searches on the dumb system. The most obvious disadvantage is that such repetition cannot be expected to give an accurate picture of the behaviour of real users over complete sessions. Consider, for example, a session consisting of a single search which leads to a satisfactory result on the qe or the full system, but which when repeated on the dumb system finds only a few relevant records. If this were the beginning of a live session on the dumb system the (real) user would very likely have followed the initial unsatisfactory search with others using different terminology.

At the other extreme are purely formal experiments along the lines of those discussed in 4.2.2. These are useful in secondary experiments, following primary testing with real users. They can be used in gathering evidence about the likely effects of functional changes in the systems, for example in term weighting functions or the choice of terms for query expansion.

4.5 Aims of the evaluation

The primary object was to compare the systems (dumb, qe and full) with regard to effectiveness, efficiency and user acceptability. We already knew from informal testing that the qe system was sometimes "better" than the dumb system. Most of those who had tried it did not seem to find it difficult to operate, and, at least when used by its designers, it was often capable of increasing both the precision and the recall of searches. It was felt that the full system might be somewhat confusing to some users, but that its additional facility (looking at books classified near a chosen one) might sometimes outweigh any confusion factor. For an account of the performance measures chosen, see 4.7.

4.6 Planning the evaluation

We had to choose an experiment which could be expected to give results which were both significant and meaningful, and which would not require too much time and expense. The preferred method would have been analysis of log data obtained from live use, supplemented by the use of questionnaires on a sample of users. Unfortunately there were practical reasons which made this method impractical. PCL libraries use the integrated LIBERTAS system for catalogue access. Like most online catalogues this provides copy availability information, and also PCL allows users to make online reservations. It would have been very difficult to interface our experimental systems to LIBERTAS in such a way as to make these facilities available. Data would thus only have been gathered from the possibly unrepresentative sample of users who were prepared to use terminals which did not give availability information. More importantly, our database consisted of the PCL monograph collection as it was in mid-1985. Updating from a MARC tape is not particularly difficult, and SLS Ltd, the suppliers of the LIBERTAS system, were prepared to produce tapes for us. However, these tapes would not have included class numbers, which LIBERTAS treats as local data attached to copy numbers, not to titles. Nor would they have included information on the number and location of copies.

Thus we were compelled to choose some type of laboratory experiment, midway between the formal and live extremes. In a number of ways the method we chose lay somewhere between the Comparison and Sample search experiments used in the NLM experiment (4.3.1). We used subjects who were fairly representative of the users for whom the bookstock in the bibliographic database was intended. We gave them a choice of questions on which to search, leaving them free to choose topics of which they had a reasonable knowledge. If online catalogue users are classified by experience of the system into novice, those with some familiarity and frequent users, most people usually to fall into the middle category. Hence we tried to advance our subjects from the "novice" category by giving a brief introduction to each system before the session on it. Finally, we gave them tasks which are typical uses of a subject catalogue - trying to find reading lists for essays. To avoid prompting the exact search phraseology we used fairly extended questions in the form of essay titles. No subject did the same searches on different systems.

It was hoped that each subject would be able to use all the three systems, but we found that three search sessions with intervening questionnaires would take well over an hour. Previous work by Richard Jones with undergraduate subjects showed that some people showed signs of impatience or tiredness after 40 minutes of searching and answering questions [JONE88]. We decided that subjects should only use two systems each. There remained the question of the choice of systems and the order in which they were to be used. With hindsight, it may well have been better to use the full set of six permutations of two from three (taking account of order). However, it was decided that every subject should have a fairly brief session on the dumb system followed by a somewhat longer session on either the qe or the full system. The reasons for this appear to have included the following: the primary aim was to compare the qe and full systems, and we were fairly confident that at least the qe system would prove reasonably acceptable and effective; a brief session on the dumb system could be used to familiarize subjects with the interaction style and "feel" of any of the systems; we anticipated being able to obtain and process about 40-60 subjects, and doubted whether this would be enough to obtain useful results in a three-way comparison; two long sessions might lead to loss of concentration.

4.7 Performance measures

4.7.1 Recall, precision and efficiency

In evaluating batch-type retrieval systems the measures most often used have been recall and precision, applied to one-shot searches. Results have often been given as mean precision for a number of standard levels of recall (4.2.1). Some unified measures have been proposed, but precision at various recall levels does reflect the needs of different users for searches of different exhaustiveness. In so far as these concepts can be applied to interactive systems, the unit of interaction is not a single search but a session, where "session" is defined to consist of all the consecutive or almost consecutive searches by one user for items on a single topic.

With interactive systems recall is still a useful and meaningful concept - it would be difficult to argue that the ability to retrieve as many as possible of the relevant documents, *when required*, is not one of the major criteria for a satisfactory retrieval system. However, it is often the case that users do not require exhaustive searches. This is certainly true for the subjects in our experiment. They were asked to find suitable reading lists for the assigned topics. If one were trying to compare the reading lists one might be inclined to *deduct* points for lists which are very long. It is still the case, though, that recall-related measures are of fundamental importance. The ability to find, or rather to help the user find, inadequately described documents or documents on topics which are sparsely represented in the database is particularly important. For example, because of its classification browsing facility it might be the case that our "full" system turned out to be better than the qe system at helping users to find records which are relevant but lacking in descriptors. One valid test of the systems would be one in which subjects are requested to do searches which are as exhaustive as possible on a number of topics of varying difficulty.

Precision-related concepts, on the other hand, are not so easy to apply in the evaluation of highly interactive systems. All the same, some type of precision or efficiency measure must be important. Any system which allows the user to see all the documents in the database is theoretically capable of 100% recall. It is said that there are British Library readers who peruse the entire General Catalogue of Printed Books over a period of years. The retrieval process in almost any system involves user selection at some stage. In batch systems this is the process of scanning printed lists of retrieved records for the relevant ones. In interactive systems some or all of this selection may take place while the user is online. One obvious precision measure is the ratio of the number of relevant records found to the total number of records seen during the session. In the case of our systems, where records can be quickly scanned in single line format or seen in full, the time spent scanning records per relevant record may be a measure which better reflects the amount of effort expended. Measures of time do of course introduce another variable - that of the user's reading speed.

4.7.2 Relevance, recall and precision in the present experiment

There has been much theorizing about the concept of relevance: see for example [SARA75]. Various types of relevance have been distinguished, and distinctions made between the relevance and usefulness or pertinence of a document. Broadly, relevance is a measure of the degree to which a document is "about" the sought topic, while usefulness is more closely related to how well the document meets the individual user's needs at the time of a particular search. Many people would feel that relevance, in the sense of "aboutness", is something which can reasonably be determined in an objective way, for example by getting a panel of subject experts to assess the documents. Usefulness, on the other hand, is something subjective and ephemeral. Satisfactory operation of a system using relevance feedback will partly depend on users' assessments of the retrieved records being assessments of "aboutness" rather than usefulness. It may be detrimental if records are rejected because the user has read them already or even because they are too old (although in the latter case publication date is a feature the system can know about and may be able to use). It is even more likely to be detrimental if the user declares a record which is not about the sought topic to be relevant, on the grounds that it is interesting, for example.

One factor influencing the choice of task in the present evaluation - to compile a reading list for an unspecified person - was the likelihood that such a task would tend to externalize subjects' relevance assessments. All the records chosen by subjects as relevant were later assessed by librarians and subject experts (see 4.9). (Records looked at and rejected by the subjects were not assessed.) This enabled us to apply a precision measurement to the list of records chosen by each subject for each topic (the ratio of the number of records assessed relevant to the number of records chosen by the subject). We did not anticipate that the three experimental systems would show any difference in this respect, because the record displays on which subjects' assessments are based were identical, but this turned out not to be the case (5.6).

Because we did not ask our subjects to do exhaustive searches, recall figures could not be expected to be very useful indicators of system differences, except perhaps on topics where there seemed to be very few readily obtainable records. As in most experiments with databases of realistic size, it is impossible to measure absolute recall because the total number of relevant records for a topic can only be determined by examining the entire database. The recall-related figures which we used included the number of records chosen by a subject for a topic, and also the ratio of this figure to the total of all the (distinct) records chosen for the topic by all subjects on all systems. A second set of recall figures can be obtained by substituting the records found relevant by the assessors for the records chosen by the experimental subjects.

4.7.3 Transaction log analysis

Transaction logs reveal that some users perform long sequences of closely related searches. Systems with browsing or query expansion facilities might reduce the need to do this. Hence another measure worth looking at is the mean number of distinct searches in a session, or the proportion of sessions which contain a large number of searches. This is a recall-related rather than a precision-related measure.

A measure which is often used is the proportion of searches, or sessions which retrieve nothing at all. Looking at *searches* which retrieve nothing is only useful in the comparison of systems which treat users' initial searches differently (for example, in the comparison of a system which stems user input before searching with one which does not - see [WALK87b]). Up to about 40% of online catalogue sessions have this outcome, in some systems. In the present systems searches are treated identically until the user has retrieved something and chooses to see a full record, so there will be no significant system difference in the proportion of *searches* retrieving nothing. There may be a difference in the proportion of such *sessions*, but because of the experimental conditions the number of null sessions would be expected to be very low.

There are many other measures which can be obtained from transaction logs, for example the extent of use and non-use of specific features.

The preceding measures can be obtained automatically by computer analysis of transaction logs, but analysis by humans can show more, though with less certainty and reproducibility. It is often possible to find recurring patterns which are diagnostic of unsatisfactory design aspects, for example the use of invalid commands in certain situations, or frequent repetition of identical searches.

4.7.4 Users' opinions

An invaluable companion to log analysis is the information which can be obtained from users as responses to open-ended questions or requests for comments. Most such information can only be gathered orally. The exact content of the questions is not always particularly important. It may be quite useful to find out that more users prefer system A to system B, but it is far more interesting and valuable to find out *why* people prefer their favoured system. Users' apparent misconceptions and asides are as valuable as the reasoned and pertinent responses which one sometimes gets. The questions used in the present experiment are listed in Appendix 3.

4.8 Method

4.8.1 Subjects

As the systems to be evaluated used the Polytechnic of Central London's catalogue database it was originally intended that subjects should be drawn from the users of PCL libraries. It proved impossible to obtain a large enough sample, even with the inducement of a £4 fee for an hour's session + travelling expenses. Hence subjects were also obtained, mainly by advertising on student notice boards, from other polytechnics, colleges of London University, City University and two art colleges. Bookings were made mainly by telephone, and quite a high proportion of subjects arrived at the appointed time. Most of the subjects were undergraduates or students of equivalent level. A few were postgraduates and one or two were not students.

It seems likely that the subjects were a reasonably representative sample of academic library users, although impoverished students may have been over-represented. Subjects had very little knowledge of the nature of the experiment before arriving to take part in it. The advertisement mentioned "development of a computerised library catalogue", and the reminder, sent to most of them, contained "You will be asked to perform a few very simple tasks on a computer and to answer some questions. This should only take one hour.".

Fifty-seven subjects eventually took part in the experiment. Data from six of the subjects was not included in the analysis, in four cases because their English was so poor that they could not perform the tasks without a considerable amount of help. One subject talked so much that he did not have time to complete his tasks on both the systems within the allotted time. During one session the tape recorder failed to work, so the interview data was lost.

4.8.2 *Tasks*

It was important to make the evaluation as realistic as possible. An important characteristic of "live" information retrieval is that the searcher has an information need [ROBE81]. In our experiment, however, the searchers were paid subjects with no genuine need to search the catalogue. It was therefore crucial to provide them with topics to search for which at least meant something to them. Task sheets were prepared in five broad subject areas: social science (including politics, economics, history, sociology and psychology), engineering (mechanical, electronic and civil), computer science, life science (biology, botany, ecology) and history of art. These topic areas covered a large proportion of the available degree disciplines. For various reasons other major areas such as languages, law and the physical sciences were not covered. Subjects who were reading for a degree in any of these areas were asked to choose from the available topic areas. This choice could be made on the basis of personal interest or subjects studied at school. In spite of the fact that the subjects had a choice of topic area, and some choice of question within topic area, some of their relevance judgments were certainly affected by lack of subject knowledge.

In each topic area there were two task sheets each containing five questions, so that each subject could use a different sheet on each of two systems. For students of library and information science there were special sheets containing one question from each of the topic areas.

The questions were made fairly long and were phrased as far as possible so that they would not directly suggest the form of search statements (Appendix 4). Most of the questions were adapted from first year examination papers of the Polytechnic of Central London or of London University. An attempt was made to pair the task sheets by covering similar areas on each, but the order was varied. All the tasks were tried out on the systems before the task sheets were finalized. It was not possible to ensure that the sheets for different areas were of even approximately the same difficulty. This would have needed an extensive pilot study. It is likely that the art and social sciences sheets were somewhat easier than the science and engineering ones.

4.8.3 *Experimental procedure*

PILOT STUDY

A pilot was conducted using seven volunteers. This led to some changes in the intended procedure. Subjects seemed uneasy about working through the questions in sequence, as had been intended. One person remarked that she felt under pressure, as she was not familiar with all of the topics on her task sheet. It was therefore decided that subjects should be allowed to work on the questions in any order. The wording of the instructions to subjects was altered, and one of the task sheets was modified.

DESIGN

All subjects started with 15 minutes on the dumb system with one task sheet, followed by 25 minutes using either the qe or the full system with the other task sheet. For each topic area (pair of task sheets) systems and tasks were rotated as follows:

Subject 1	dumb system tasksheet A	qe system tasksheet B
Subject 2	dumb system tasksheet B	full system tasksheet A
Subject 3	dumb system tasksheet B	qe system tasksheet A
Subject 4	dumb system tasksheet A	full system tasksheet B

Each subject's session consisted of an introduction followed by a sequence of searches and taped interviews.

When they arrived, subjects were briefly interviewed to determine the nature of their academic background and their experience with online and other library catalogues (Appendix 3 contains all the questionnaires). There followed a brief explanation of the purpose of the experiment:

> The PCL Information Retrieval Research Team is designing a computerised library catalogue. At the moment they are working on a subject catalogue. You are probably aware that library users vary greatly in the amount of experience they have of using catalogues and computers so it is very important that computerised catalogues are easy to use for everyone. That's not too hard to accomplish if the system is quite basic but if it is fairly complex it becomes harder. By taking part in the experiment you will be helping us to investigate this problem. I will also be using the data as the basis for my MSc in Information Technology at Loughborough University.

Following this, subjects were given a very brief demonstration of the dumb system. They were then given a task sheet and were told that it represented a list of essay questions. They were asked to produce printed lists of the books which they felt would be most useful to a person writing one or more of the essays (the systems were set up so that they printed automatically at the end of each search). Subjects were free to choose whichever questions they wished, to concentrate on one or two or to work through several, but they were requested to produce at least one completed list in the time. No guidance was given about the number of references which should be chosen. The number of relevant books varied widely between questions and, in any case, it was felt that such a recommendation might prejudice the selection or rejection of items.

While subjects were searching the experimenter was sitting at a nearby desk, not overlooking the subject. Some subjects asked questions during their searches, but they were not answered apart from a general indication that they should read what was on the computer screen. The experimenter informed the subjects when their time had expired. The first search session was immediately followed by an interview which focused on the subject's experience with the dumb system. During the interview the system printed lists of all the records which the subject had seen (whether selected or not; lists of selected records for each search had already been printed). The experimenter then asked the subject to comment on the reasons for choosing or not choosing some of the records seen during the last completed search. The data collected during these discussions has not been analysed, but this period allowed the subjects a break from thinking about the mechanisms of the system they had just been using.

Following this, subjects were given a brief demonstration of the second system - full or qe - which they were scheduled to use. This demonstration was similar to that of the dumb system, but included use of the additional features of the second system. Subjects then spent 25 minutes performing searches from the other task sheet. This session again was followed by an interview (Appendix 3) and a discussion about the choice or non-choice of records. This interview contained additional questions which aimed to discover which of the two systems the subject had used was easier and which was the more helpful in finding useful books.

DATA

The data gathered consisted of transcripts of the recorded interviews, transaction logs automatically produced by the search systems and of printouts of all records seen and and of all records chosen for each search. The transaction logs (Appendix 5) contained enough information for the searches to be replayed, except that exact timing down to keystroke level was not recorded.

4.9 Objective relevance assessments

In an attempt to measure the joint performance of the systems and the subjects with regard to the quality of the lists of references produced, all the records chosen were assessed for relevance by independent assessors.

The assessors consisted of unpaid volunteers from the academic staff at Loughborough University, librarians from several institutions, and other people with specialist subject knowledge. For each question a list of all the records chosen by all subjects was compiled (this could have been done fully automatically from the transaction logs, but to reduce the amount of programming it was actually done by manual repetition of all the searches on a search program which logged the record numbers of all records chosen). These were made up into five lists, each covering all the questions in a topic area. Each of these lists was assessed by three people. The references for each question were presented in randomized order to compensate for the effect of the order of presentation on the assessors' judgments.

Assessors were asked to use their knowledge of the topics to make a quick assessment of the references from the information given in the references alone, and to tick the point on the scale which best described the way they felt about each book. They were asked to indicate how useful each reference would be to a person writing an essay on the topic, the "scale" being "very useful", "quite useful", "slightly useful", "not useful" and "other--please specify". Assessors' instructions and an example are given in Appendix 6.

This procedure was decided upon after a discussion with Professor S E Robertson of City University. In particular, Robertson advised the use of a four-point scale (with a fifth "don't know" category), on the grounds that it is easier for the assessors than a two or three point one, and that the scores should later be collapsed into a simple "good"/"not good" dichotomy. He also advised that it may be better to refer to *usefulness* rather than *relevance* and to *degrees* rather than *probabilities* of usefulness.

In scoring, "very useful" and "quite useful" were counted as relevant and "slightly useful" and "not useful" as non-relevant, and a consensus of the three judges was taken. Cases where a "don't know" assessment led to a tie should have been resolved by an additional assessor, but this was not done because of shortage of time, and these records were omitted from this part of the analysis.