5 ANALYSIS & RESULTS

5.1 Introduction

This chapter contains a general comparison of the three search systems based on statistical analysis of the transaction logs and on user comments and responses to questions. This is followed by a more detailed discussion of results relating to use of the query expansion and classification browsing options. The results of the external assessment of subjects' lists of chosen records are given. Finally there is a summary of users' comments on the systems.

5.1.1 *Terminology*

For brevity, some terms will be used in a specific and sometimes non-standard sense in this chapter. A *search* is the interaction of a user with the system starting with and resulting from the input of a single search statement. A *topic* is one of the questions given in Appendix 4. A *topic-session* is one or more consecutive searches by a single user aiming to retrieve records on a single topic. A *session* is a sequence of one or more topic-sessions by the same user on the same system, finishing when the user finishes or changes to another system. The three systems, which were described in Chapter 3, are often referred to as D (dumb system), Q (the system which allows query expansion by use of the "More" option but does not offer classmark browsing) and F (the full system, allowing both query expansion and classification browsing). Sets of records retrieved by the systems are referred to as *lists*, to reflect the fact that they are rather dissimilar from the unordered sets retrieved by most of the conventional boolean retrieval systems.

5.1.2 *Source and processing of data*

Statistics on numbers of searches and sessions, records retrieved, seen and chosen and their source, and on timings, were obtained from automatic analysis of the transaction logs generated by the search programs while they were in use during the experiment described in Chapter 4. A computer program was written to obtain this data from the logs. There is an example transaction log in Appendix 5. There was some manual editing of logs to insert subject numbers and the reference numbers of the topics which were being searched (Appendix 4). In a very few cases it was not possible for the experimenters to be certain of the topic, and these were omitted from analyses. The second main source of data was subjects' responses to questions (Appendix 3) and their comments. These were transcribed from tape recordings made during the sessions.

Secondary data was obtained in various ways. Sessions were replayed either manually or automatically, using a program which read users' keystrokes from logs and passed them as input to the search program. This enabled experimenters to see exactly what the original subject had seen. One of the experimenters (Stephen Walker) made assessments of the quality of the lists of records displayed, as well as observations on user behaviour with regard to use of commands, time spent reading screen displays etc. A similar method was used to compile complete lists of all

records chosen for each topic, together with the subject(s) who had chosen them and the source of the record (original search, query expansion or class browsing). The primary purpose of these lists was to produce printouts of records for external assessment for relevance (4.9), but they were also used to generate lists of the Dewey numbers at which all the chosen records were classified.

Note on statistical tests

Measures of system performance such as numbers of records chosen or seen are all more or less skewed and it may not be safe to use tests which depend on an assumption of normality. In testing for performance differences between systems we used Mann-Whitney U tests, which depend only on the rank-order of the observations. For testing differences such as the proportion of relevant records obtained under different conditions chi-squared tests were used.

5.2 General comparison of the systems

5.2.1 *Statistics on efficiency and effectiveness*

Some figures obtained from computer analysis of the transaction logs are given in Table 5.1.

Several measures of efficiency are possible. The ratio of the number of records chosen as relevant to the number of brief records seen is a precision-type measure (row 4 of Table 5.1). Here the Q and D were both significantly better ($P > .99$) than the F system. Similar results hold for the time spent per record chosen (row 5). It was thought possible that browsing and expansion facilities might lead to a lower mean number of searches per session. A search on the full (F) system has the potential to lead to the selection of a wider range of records than a search on the query expansion (Q) system (and both F and Q might show a reduction in the number of searches over the dumb (D) system). This was not the case. Row 1 of the table shows that the Q system had the lowest mean number of searches per session, followed by D and F, but the differences are not significant.

With our experiment it was not really possible to consider the question of relative retrieval effectiveness of the systems. If any of the systems had been markedly *in*effective this would have shown up, but even the dumb system is presumably at least as good for subject searching as most existing online catalogues. To measure effectiveness it would have been necessary to ask the subjects to try to carry out exhaustive searches, and this would have removed the experimental task even further from a realistic situation. It is true that the mean number of records chosen per topic-session is significantly greater for the Q and the F systems than for the D system (row 3 of Table 5.1), but all subjects used the D system first and a learning effect was to be expected. Any difference between the Q and the F systems is small and not significant at the 5% level.

Overall, users of the F system worked very much harder than users of the other systems without obvious benefit. However, it will appear later (5.6) that they tended to produce lists of chosen records containing a higher proportion of "good" records than those resulting from use of the other systems. The fact that F users looked at far more brief records is almost entirely due to the fact that every time they chose a record they

were asked whether they wanted to see "books shelved near this one" (Fig 3.12). Most users chose this option, at least near the start of their session (some soon appear to have become disillusioned), and the classified sequence was not, on the whole, an efficient source of relevant records (Table 5.4).

Table 5.1 Some comparative system use statistics

	Q	F	D	all systems
subjects	24	27	51	51
topic-sessions	65	64	108	237
t-sessions with no recs chosen (omitted from the analysis)	3	1	8	12
(1) searches/t-session	1.88	2.25	2.02	2.04
(2) brief recs seen/t-session	62.9	111.3	44.1	67.4
(3) records chosen/t-session	8.3	7.9	5.4	6.9
(4) brief recs seen/record chosen	7.5	14.1	8.2	9.8
(5) time/record chosen (secs)	40.7	56.1	52.5	49.7
(6) time/brief rec seen (secs)	5.4	4.0	6.4	5.1

5.2.2 *User opinions on system ease and helpfulness*

After they had used the dumb system followed by either the Q or the F, subjects were asked which system they had found easier to use and which one they had found more helpful in finding books relevant to the essay titles. The results are summarized in Table 5.2. This shows that two-thirds of the Q subjects felt the second (Q) system to be at least as easy as the dumb system, but more than half the F subjects felt that the dumb system was easier. The hypothesis that Q subjects are more likely than F subjects to find their second system at least as easy as the dumb system is accepted at the 5% level (chi-square = 3.43). On helpfulness the Q and F systems are more evenly balanced, 92% of the Q subjects and 78% of the F subjects preferring the second system. Although it appears that the Q system may be more often found helpful than the F the difference is not significant.

Table 5.2 Perceived ease and helpfulness relative to the dumb system

	Q system users (24)	F system users (27)
Ease:		
1st system easier	8	16
no difference	7	6
2nd system easier	9	5
Helpfulness:		
1st system more helpful	1	5
no difference	1	1
2nd system more helpful	22	21

Reasons given for the Q system being easier than the D system included:

- the "more" option (four subjects)
- you have to think less (two subjects)
- it directs you better (two subjects)
- easier to "home in on topic"
- cuts down the number of references
- learning effect (dumb system used first)

Reasons given for the D system being easier than the Q system:

- fewer options (two subjects)
- easier to keep track of the topic
- faster
- simpler

The F system was judged easier than the D system because of

- the "shelf" option (two subjects)
- the "more" option

The D system was judged easier than the F system because

- it's simpler (four subjects)
- it has fewer commands (four subjects)
- it has less choice (two subjects)
- it asks no questions
- it goes in one direction
- there is less pressure
- it's more flexible
- it's more structured
- it's easier to navigate round
- it's faster
- it's less distracting
- it's more user friendly

All but two of the 24 Q subjects thought that the Q system was more helpful than the D system. Reasons included the following:

the "more" option (twelve subjects)
gives a greater scope of books (four subjects)
homes in on the topic better (two subjects)
you don't have to think of so many words (two subjects)
it's easy to follow the topic
it gives you a second chance to find books

Twenty-one of the 27 F subjects found the F system more helpful than the D system:

the "more" option (eight subjects)
the class browsing option (five subjects)
it had more options (three subjects)
it found more books (three subjects)
you see more books
it did the work for you
you don't have to go through the entire list

Several fairly clear-cut conclusions follow from the comments on ease and helpfulness. The "more" option was seen as being effective as a way of homing in on a more focused list of books, and as reducing the amount of effort both in scanning records and, to a lesser extent, in thinking of alternative ways of expressing the topic. Class browsing was seen as helpful by an appreciable proportion of the F subjects, but its use involved a good deal of scanning of screens of records. It is worth noting that although F subjects chose more records from class browsing than from the "more" option (see below - Table 5.4), the latter was mentioned by more subjects than the former in response to the question about helpfulness. Two of the F subjects spontaneously remarked that if the "more" option were added to the dumb system this would provide the ideal combination. This combination is, of course, already present in the Q system.

5.2.3 *User opinions on system usefulness*

Subjects were asked

> About what proportion of the time did you feel that the computer was useful in helping you to find books which were relevant to the essay titles you were given?

Not surprisingly people found this question difficult to answer. Some were unable to suggest a figure. The numerical results are summarized in Table 5.3. The Q system looks to be useful more of the time than either of the other systems, although the results are barely significant. Many of the subjects made interesting comments about what was happening when the computer was not being useful. These supplement the responses to the questions about problems with the systems, and they are included in the comments discussed in 5.7.

Table 5.3 Proportion of time the computer was useful

proportion of time system was useful	Q system users	F system users	D system users	totals
50% or less	4 (20%)	9 (36%)	11 (23%)	24 (26%)
51% - 70%	2 (10%)	3 (12%)	14 (29%)	19 (20%)
more than 70%	14 (70%)	13 (52%)	23 (48%)	50 (54%)
	20	25	48	93

5.3 Use and performance of the query expansion and classification browsing facilities

We have already seen (5.2.2) that a substantial number of both Q and F users felt that the query expansion option rendered these systems more helpful than the dumb system. A significant proportion of F users also mentioned classification browsing in this context. Certainly, both the options were extensively used, despite the fact that the experimental subjects were not specifically urged to try them.

Table 5.4 shows the proportion of records which were chosen from lists retrieved by each of the three access facilities - the original list retrieved with the user's query terms, and lists retrieved by query expansion searches and classification browsing - on each of the three systems. On the Q system query expansion accounts for two-fifths of the records chosen. On the full system query expansion gave one-fifth of the records and class browsing more than two-fifths, so that the original list was a less important source than the expansion options.

Table 5.5 shows the performance of these options in use on the Q and F systems. Performance is classified as "good", "moderate", "bad" or "failure". A *failure* occurs when choice of the query expansion option led to no records being retrieved. Use of either of the options is *good* when it led to the user choosing three or more records from the first screen, or at least half the records retrieved if the (query expansion) option retrieved less than six records. It is *bad* if at most one record was chosen and that not from the first screen. Any other case is *moderately good*. This classification is fairly arbitrary, but is intended to reflect the fact that if query expansion works properly the best records will usually be very near the top of the list. If the user has to look at several screens to find relevant records from either of the options they are not working very well. Indeed, it was quite unusual for a user to look at all the records retrieved by query expansion.

Table 5.4 Breakdown of records seen and chosen by system and source (means per topic-session)

Source	Q system (65 t-sessions)	F system (64 t-sess)	D system (108 t-sess)	all systems
Original list				
brief recs seen	33.5 (53%)	30.6 (28%)	44.1 (100%)	37.5
records chosen	4.8 (58%)	3.0 (38%)	5.1 (100%)	4.6
brief recs/choice	6.9	10.1	8.2	8.2
'More' option				
brief recs seen	29.5 (47%)	23.9 (21%)	N/A	26.7
records chosen	3.5 (42%)	1.6 (20%)		2.6
briefs/choice	8.4	15.0		10.4
'Class' option				
brief recs seen	N/A	56.8 (51%)	N/A	56.8
records chosen		3.3 (42%)		3.3
briefs/choice		17.2		17.2
All sources				
brief recs seen	62.9 (100%)	111.3 (100%)	44.1 (100%)	67.4
records chosen	8.3 (100%)	7.9 (100%)	5.1 (100%)	6.9
briefs/choice	7.5	14.1	8.2	9.8

Note. The true figures for the 'More' option on the F system are slightly lower than those shown, and the 'original list' figures correspondingly higher. A bug in the search program caused two searches to be affected by misbehaviour of the 'Back' option, with the result that a few records shown as having come from the original list really resulted from query expansion.

Table 5.5 Performance of the query expansion and class browsing options in use

performance of option in use	query expansion Q system	query expansion F system	class browsing F system
good	23 (17%)	8 (6%)	16 (7%)
moderate	67 (48%)	59 (47%)	77 (36%)
bad	39 (28%)	53 (42%)	122 (57%)
fail	10 (7%)	6 (5%)	N/A
total	139	126	215
per session	2.14	1.97	3.36

5.4 Query expansion in detail

5.4.1 Statistics

On the Q system, the query expansion option was used a total of 139 times, by 20 out of 24 subjects and in 50 out of 65 topic sessions. It led to the retrieval of 42% of the records chosen on the Q system. On the F system it was used 126 times, by 26 of 27 subjects and in 53 out of 64 topic sessions. It led to the retrieval of 20% of the records chosen on the F system (Table 5.4). The extent of use of this option is slightly surprising, as although the facility was very briefly demonstrated prior to each session subjects were not specifically encouraged to use it, and the prompt is only one among several options (Fig 3.8). It may be that it would have a lower take-up in live use of the systems.

The proportion of failures (not retrieving any records) was low: 7% for the Q system and 5% for the F system. The combined figures for good and moderately good were 65% for the Q system and 53% for the F system (Table 5.5). It is clear that query expansion is a fairly prolific and relatively "easy" source of records perceived as relevant. It was significantly (at 5%) less useful and efficient on the F system than on the Q, both with regard to the number of records chosen and the number of brief records looked at for each one chosen. It appears that F subjects spent less time looking at the results of query expansion, perhaps because they had often already selected records from classification browsing. It is worth noting that only 15% of the query expansion records selected by F users were from the second or subsequent screen of records retrieved, as against 36% for Q users. This suggests that F users were more likely to feel that they had already chosen enough records.

5.4.2 Quality of lists of records from query expansion

The number of records chosen from query expansion by the subjects in the experiment cannot be expected to be a good indicator of the quality of the lists. The number of records already chosen will certainly affect users' behaviour. This is borne out by the fact that query expansion was far less fruitful on the F system than on the Q system (Table 5.5); it seemed unlikely that the lists produced by the option would be noticeably less good on the F system. Thus an attempt was made to assess the quality of the lists of records retrieved using query expansion by the subjects in the experiment. To do this, all the Q and F searches were repeated by one of the experimenters. He looked at the first screen of brief records retrieved by each query expansion search and graded the screens in accordance with the following scale:

A: At least four of the records look relevant (at least half, if fewer than eight records were retrieved)
B: Several of the records are worth looking at
C: One or two of the records might be worth looking at
D: It is unlikely that any of the records would be relevant.

"Relevance" was judged from the brief records only. The assessment was nearly always done with respect to the question as given in the appropriate topic sheet, not the user's actual search statement, as search statements do not always indicate what users are "really"

searching for, rendering it very difficult to assess relevance. For example, it would be difficult to know what the searcher for "Slump 1932" was looking for without knowing that the question was "How widespread was the Slump by 1932? ..". The question from the topic sheet was used even where the search statement was much broader than the sought topic, on the assumption that the user would have selected records which were relevant to the question. An example of this is the search "Welfare economics" for the question "Would perfectly competitive markets ensure maximization of social welfare?". In a few cases, however, the search statement was comprehensible but seemed rather remote from the sought topic, and here the search statement was used rather than the topic.

Table 5.6 is a summary of the results of this assessment. Thirty query expansion searches are omitted from the table, 16 because they failed to retrieve any records and the remainder because of unidentified program errors which rendered it difficult to repeat a few of the sessions accurately. For comparison, the experimenter's assessments are cross-tabulated with the gradings for performance in use as given in Table 5.5.

Table 5.6 Experimenter's assessment of query expansion searches

Experimenter's assessment of query expansion		Performance of query expansion in use: total	good	moderate	bad
A	Q system	54 (44%)	18	29	7
	F system	47 (42%)	3	35	9
	both systems	101 (43%)	21	64	16
B	Q	38 (31%)	5	21	12
	F	37 (33%)	3	16	18
	both	75 (32%)	8	37	30
C	Q	16 (13%)	1	9	6
	F	15 (14%)	1	4	10
	both	31 (13%)	2	13	16
D	Q	16 (13%)	0	4	12
	F	12 (11%)	0	0	12
	both	28 (12%)	0	4	24
Totals	Q	124	24 (19%)	63 (51%)	37 (30%)
	F	111	7 (6%)	55 (50%)	49 (44%)
	both	235	31 (13%)	118 (50%)	86 (37%)

Table 5.6 shows, as expected, that there is no detectable difference between Q and the F systems in the quality of the lists of records retrieved by query expansion searches. More significantly, it shows that three-quarters of the searches fall into the A and B categories. This is an encouraging result, particularly as the option is freely available, and was often used, even when only a single record has been chosen as the source of terms for expansion. However, this is not enough evidence to conclude that the free availability of query expansion is worth its

cost in computational resources. At this stage there are numerous unanswered questions. We do not know how many of the records chosen from query expansion were readily available from a previous list (the original list, a previous query expansion or, for the F system, a class browsing screen). Some of the experimental subjects, particularly some of those who used the Q system, did a query expansion search after almost every choice of a record. In these cases, once more than three or four records have been chosen, the lists retrieved by successive query expansions tend to be almost identical (except that records which have been chosen do not appear in subsequent lists). These repeated searches do not usually result in the loss of potentially relevant records, but nor do they help the user; it would probably be more efficient for the user to select more of the available relevant records before doing another query expansion search. This raises questions about the presentation of the "More" option which are discussed in 6.2.1.

5.4.3 *Users' comments on query expansion*

Users of both the Q and the F systems were asked the following question after their session:

> Did you use the 'more' option? Did it help you to find more useful books?

Twenty-two of the 24 Q system subjects said they had used the option. The true figure was 20, and all these said that it had helped them (the other two said that it had not helped). Of the 27 F system subjects all said they had used it (true figure 26), 19 said that it was helpful, 7 were uncertain and 1 said that it had not been helpful. This reflects the smaller proportion of books chosen from query expansion by full system users (Table 5.4).

Subjects were asked if they would like to comment on the facility. Fifteen of the Q and 18 of the F subjects did comment. The comments were mainly appreciative but there was a certain amount of criticism of the way in which search results were presented. Q subjects were on the whole more positive than F subjects. Some of the F subjects may not have clearly distinguished the "More" option from the classification browsing, and they certainly made less extensive use of "More" than the Q subjects; eight of the F subjects used expressions of uncertainty ("I think ..", "It was useful in some ways but ..", "I think it was 50/50.") against only three of the Q subjects.

Nine Q subjects and 11 F subjects reiterated that they had chosen books from the "More" option or that it had been useful. Six of the Q subjects said that they had used the facility "a lot" or "all the time". One subject said that he had chosen one book from there which he thought he would not otherwise have found.

The facility can bring about a shift in the emphasis of a search, not always for the better.

> It seemed to enable you to get more specific books
> It helped bring out other ranges of books
> It was helpful to feed you into new areas
> It seemed at times as if it was getting a bit too specific
> It was useful in some ways but it was still getting away from the search

Two Q subjects said that it was helpful when the initial search had not

been very successful.

> It found more books especially on the subjects I didn't know anything about.

> The first section didn't come up (I was looking for the 1932 slump) and when I pressed 'More' for specific books it was much more helpful.

Two subjects thought that it was time-consuming, and one that it was quite tiring enough browsing through the original list.

> It makes it more time-consuming in some ways by breaking it down further

Three users felt it to be a fault that query expansion sometimes retrieves records from the original list which have not been either chosen or rejected.

> I found other books listed in the original category [i.e. original list], so whether it was finding more books or whether I was just wasting time I don't know

This is one of the aspects of the presentation of interactive query expansion systems which the designers had given considerable attention to. It is perhaps noteworthy that more of the subjects did not comment on this repeated retrieval of the same records.

5.5 Classification browsing in detail

5.5.1 *Statistics*

On the F system, the only system offering this facility, the classification browsing ("books shelved near ...") option was used a total of 215 times, by all the 27 subjects who used the system, and in 63 out of 64 topic sessions. It led to the retrieval of 42% of the 507 records chosen by users of the F system. The extent of use is not at all surprising, since it was offered by means of a yes/no choice every time a record had been chosen relevant (Fig 3.12).

Invocations of this function were classified in the same way as for the query expansion option (above), except that there are of course no cases of failure to retrieve any records. The combined figures for good and moderately good are 43%, substantially lower than the corresponding figure for query expansion (Table 5.5). In 54% of cases the user chose no records at all. It is clear that classification browsing is often not useful, yet over 40% of records chosen on the F system came from this option. It was relatively inefficient as a source of records: Table 5.4 shows that a mean of more than 17 records were looked at (in brief) for every one chosen.

It was noticed that the classification seemed to be particularly ineffective in the area of computing, where there were several thousand records broadly classified at eight numbers within 001.64 (electronic data processing). In two-thirds of the 50 invocations of this option in searches for computing topics no records were chosen. The Dewey Classification has since been revised in this area, and PCL records have been reclassified.

5.5.2 Quality of lists of records from classification browsing

In an attempt to find out more about the operation of classification browsing 38 of the full system sessions were repeated by one of the experimenters, who graded the screens of records which the user had seen displayed following choice of the browsing option in a similar way to that described above for query expansion (5.4). The 38 sessions were all those of the first 17 subjects who used the full system except for four sessions and two part-sessions which were omitted because, for technical reasons, it was not easy to repeat the searches exactly as they had been performed. Two further part-sessions were omitted because the user's search statement did not seem likely to retrieve records on the sought topic. The screens were assessed on the basis of the brief titles displayed; no account was taken of the actual definition of the Dewey class marks (nor was any account taken of whether the user actually chose any records). The definitions of the categories were as follows:

A (good): a reasonable proportion of the records seen appeared to be about the sought topic. Example: 574.875 (cytology - membranes and cell wall) in a search for "Ion transport".

B (possible): some of the records seen appeared to be not too distantly related to the sought topic. Examples: 658.403 (management decision making and information management) in a search for "Management information system design", 155.422 (child psychology - infants) in a search for "Influence of the mother on the child".

C (scattered): a few of the records were somewhat related to the sought topic: the records seen covered a wide range of topics. This category usually appeared when there are very broad classification codes. Example: 621.38 (electronic and communication engineering) in a search for "Computer data in telephone network".

D (remote): none of the records seen (apart from the one used as the pivot) appeared at all closely related to the sought topic. Example: records at 362.17.. (specific medical services) in a search for "Computers in medicine".

These gradings may be compared, cautiously, with the identically lettered ones used in the assessment of screens from query expansion (Table 5.6). Assessment of the classified displays was not limited to the first screen displayed. This was tried initially, but the experimenter felt that it gave results which would be less likely to reflect the behaviour of users. Query expansion usually gives the best records very near the top of the display list, with similarity decreasing steadily and relatively smoothly. Class browsing displays are far more erratic and unpredictable, and it seemed more reasonable to make assessments based on what users had actually chosen to see. Table 5.7 summarizes the findings. More detailed results, with topic references, search statements and Dewey numbers, are given in Appendix 8.

Table 5.7 Experimenter's assessment of a sample of the classified displays

Experimenter's assessment of classified displays	Performance of class browsing in use			
	total	good	moderate	bad
A (good)	22 (22%)	7	9	6
B (possible)	49 (49%)	3	22	24
C (scattered)	18 (18%)	0	4	14
D (remote)	12 (12%)	0	0	12
Totals	101	10 (10%)	35 (35%)	56 (55%)

The experimenter's assessments of the classified displays in Table 5.7 suggest that classification browsing might be more useful in live use of a system than the results during the experiment suggest. While only 22% of uses gave a good proportion of relevant records, another 49% looked as though they may provide a few hits. The reasons for the discrepancy between the experimental use figures and the experimenter's assessments appear to be (1) that the subjects in the experiment were not trying to do exhaustive searches, and (2) that in many cases substantially the same display was seen more than once: subjects chose the option more than once at the same Dewey number, and so were less likely to want to choose any records after the first occasion.

5.5.3 *Users' comments on classification browsing*

Users of the F system were asked the following question after their session:

'Did you opt to look at books shelved near the one you had chosen? Did this help you find more useful books?'

All 27 F subjects had used this option. Twenty said it had helped them find useful books, five said it had not been useful and two were uncertain. The logs show that three subjects chose no records from classification browsing and three chose only one record.

Eighteen of the subjects commented on their experiences with the classification option. Despite the favourable reaction reported above, all of the comments were in some degree critical. Subjects drew attention to the fact that the facility was not always useful and that it could be time-consuming or confusing. Some of the following comments were made in answer to the question "Did you have any problems using the computer?", but they are more appropriately given here.

Nine subjects made remarks to the effect that the facility had been sometimes useful and sometimes not.

> It was not as useful as I thought it might be.
> [It was useful] in a small amount of cases - not as much as the 'more' option

though.

Eight subjects remarked that books which are close together in the classification sequence are not necessarily about the same subject, or that related books were separated by books on a different subject.

> Just because books were on the same shelf didn't mean they were really relevant.

> [It] usually got you off the track.

> I had to discipline myself not to waste time looking at everything.

> [I] got confused going through all the different topics.

> There's a difference between looking along bookshelves and being able to see things nearby. I put in a keyword and found a book and looked three or four screens up and down from that, and right at the extremes of those [I] found lots of major books. It would have been helpful if I could have skipped about ten books. Rather than looking along a shelf, looking down about five shelves.

One subject said

> Get rid of that!

5.6 "Objective" precision of chosen records

Perhaps the most surprising result of this experiment is that when the records chosen by the subjects were assessed for relevance by independent assessors, as described in 4.9, the F system gave markedly higher precision than the Q system. The figures are given in Table 5.8, with a breakdown by source of records (original list, "more" option, "class" option). In fact the F system is significantly better in this respect than both the D and the Q systems. The D system also gave higher precision than the Q system, but the difference in this case is less marked. It must be emphasized that the precision which was measured is that of the lists *chosen by the subjects*, not that of the lists produced by the systems.

It is interesting to speculate about the reasons for the overall higher precision achieved by F users. Records were displayed in exactly the same full and brief formats on all the systems. Once a record had been displayed in full format F users chose about the same proportion as did Q and D users (over all systems the proportion of displayed full records chosen was about 72%, or, putting it the other way round, about 28% of "promising" records were rejected after they had been seen in full). From Table 5.1, row 4, it can be seen that F users retrieved and presumably looked at considerably more brief records than did users of the other two systems - about 14 for each record chosen as against seven or eight on the Q and the D. It is tempting to guess that F subjects rapidly became more discriminating about the choice of records, knowing that an almost unlimited number of screens of brief records were readily available.

Table 5.8 Assessed precision of subjects' choice of records

number/proportion of relevant records by system and source	Q system rel/total (precision)	F system rel/total (precision)	D system rel/total (precision)	All systems rel/total (precision)
source: original list	194/280 (69.3%)	145/182 (79.7%)	371/535 (69.3%)	710/997 (71.2%)
source: 'more' option	124/211 (58.8%)	62/90 (68.9%)	N/A	186/301 (61.8%)
source: 'class' option	N/A	150/203 (73.9%)	N/A	150/203 (73.9%)
totals	318/491 (64.8%)	357/475 (75.2%)	371/535 (69.3%)	1046/1501 (69.7%)
(omitted, see note)	(51)	(32)	(44)	(127)

Note. Forty choices (29 records) are omitted because of missing or "don't know" relevance assessments. These would probably have little effect on the precision figures. A further 87 choices are omitted because they are duplicates: many subjects chose the same record more than once in different searches on the same topic.

5.7 User comments about the systems

In response to the questions about "problems with the computer" and suggested improvements (Appendix 3) some of the experimental subjects made comments which should be of interest to retrieval system researchers and designers, although by no means all the comments bear on matters relating to query expansion. Many comments were also given in explanation of what was happening when the computer was not being useful (5.2.3). There were a considerable number of mildly critical comments, but the interviewer sought these. She did not seek compliments. The finding that most of the subjects were readily able to make adequate use of any of the systems is a gratifying result, although the query expansion system appeared to offer the best combination of ease of use with effectiveness.

5.7.1 Choice of search terms and retrieval of non-relevant records

There were more than 20 comments about the necessity of weeding out non-relevant records ("Looking through the books", "Computer bringing up unwanted books", "I put in AI and got lots of foreign books"). Most subjects seemed to accept that this was necessary, but there were half a dozen complaints about the systems retrieving records which had little relation to the sought topic. Some of these were about false drops due to false coordination or homography and others were about records being retrieved under just some of the user's search terms. Some users feel that the system has, or at least ought to have, a linguistic knowledge that extends to the recognition of noun phrases which describe a topic, even if it is unable to find any records in response; others think that the onus is on them to find the right way of describing the sought

subject.

> It's really defining what you want - I put 'social welfare' into the search and it came up with 'social work' which wasn't the same thing. Sometimes you put something in and it comes up with something totally absurd - it's actually the dictionary definition of what you put in rather than the concept.

Two users expressed the wish to see records under each of their search terms separately. One expressed doubts about the feasibility of this:

> I would have liked to have been able to have separated out the words that I selected so that instead of it going for a combination of the words I used and that was it, I would have been able to tell it that I wanted to look at all the economics books. But then I'd probably have about 300 books to look at.

There were actually about 6000 books indexed under economics.

There were 16 comments about the difficulty of thinking how to express the topic ("Thinking of words", "Can't define what you want", "Picking the right phrase"). A few users thought that the system ought to be able to help them in the choice of terms.

> I wasn't specific enough in requesting precisely the thing I wanted so I seemed to get more general titles to scan through.

> Words should be made more flexible - tell you other words that mean the same thing. For example, another word to use instead of 'design'.

5.7.2 Record content and display; the recognition of relevance

There were more than 20 comments about the difficulty of assessing relevance on the basis of the displayed information. Most of these are connected with the lack of subject information in the source bibliographic records, and there is not much that can be done without a new approach to cataloguing. About 80% of the records in the database, and nearly all the records retrieved by the subjects, had LCSH or PRECIS headings or both. Several subjects suggested that additional information should be presented on request "at a lower level" [than the "full" record].

> Once it's selected a book and you've looked at it you could have more details of what the book's about. It gave who wrote it and the publisher but it would be more useful to actually have a general idea of what the book's about.

Three subjects asked for full non-abbreviated titles in the brief display:

> It would be better if the titles weren't abbreviated - they should run onto the next line. Where there are lists of reports from seminars perhaps they could be under one heading and then you go further in but maybe that's too complicated for the computer.

Only one subject complained about the constant switching between brief and full display, although it is likely that an appreciable proportion of users would be conscious of this if they were using the systems frequently, or if they were using a terminal connected over a slow network.

> Is it possible to save time when selecting a book? There were some I could

> say I wanted just from the title - I didn't need to see full details The title either told me that I knew the book or gave me sufficient information or the author or year did. It would have saved time to be able to select it from the first screen.

No subject mentioned that a large proportion of books could be *rejected* without seeing a full display. This seems to be because rejecting records was not seen as a significant part of the process of conducting a search; choosing a record adds it to the list which will be printed, but it is not clear to users that if unwanted records are not rejected they may appear again in a list resulting from query expansion. Five or six of the subjects complained that the "More" option often included records which they had seen before. It seems clear that records not chosen are often regarded as rejected, and the system should have some way of taking account of this.

5.7.3 *Problems connected with the list of chosen records*

In the evaluation experiment subjects were asked to use the systems to compile a printed list of references (4.8.3). This was the carrot to motivate subjects to choose references as relevant, and it is unlikely that a similar procedure would work in normal, live use of a retrieval system. It had been anticipated that there would be interactional problems connected with these lists of chosen records. As expected, there were two sources of difficulty. It was not possible to edit the list by including a previously rejected record or by removing a record. There were several sessions where the subject repeated a search to obtain the desired printed list. Nor was there any facility for retaining the same list over a number of related searches. There were ten comments about editing the list and seven about carrying it over from one search to the next.

> If I say I don't want it to remember a book it doesn't mean that I want it to forget it. I couldn't go back. I might make a mistake pressing 'no'.

> Let me keep adding to the list I've got or delete things from the list. I'm stuck with my original list unless I start again. I can't do things incrementally.

> When I changed to a new topic I didn't really look that hard to see if there was a way I could go back and forth between my sections or not? Sometimes you type a topic in and there's nothing. The first question I was doing, it was very difficult to find books that I really wanted (propaganda and art). There's so much about world war two and before that - that's interesting but I wanted to find some more relevant books so I had to type in some new topics and I couldn't refer back to the old list I already compiled. But I would have typed it out and had it next to me...

5.7.4 *General interaction and presentation*

Six full system users made comments to the effect that the system had too many options or that they did not know where they were or what was happening.

> I think you need fewer options actually.

> It's more confusing than the last one. I couldn't remember which one I was on - whether I was on the books next to it or on the subject search or which.

> I did get a little confused between 'more' and the shelf option. There were too many books to look at.

One or two users suggested showing more options at a time rather than less so that it was not necessary to go through so many stages. There were few complaints about the umbrella "Restart" option (Fig 3.14), but the logs and other evidence show that some users had problems finding out how to finish, and the difference between "New", "Edit" and "Quit". This did not seem seriously to affect their searches. Many subjects did not use the "View" option, and one or two were plainly unaware of it because they suggested such an option.

Only one subject suggested that there should be online help or additional instructions. Several suggested the use of colour either to highlight prompts or to distinguish between the different lists of records. One subject suggested using a mouse instead of mnemonic commands.

> Display: clarity of screens - the way it's arranged; colour screens; different typeface - direct the user to different parts of the screen; icons; use a mouse - not many people can type.

> Some kind of summarised key to the instructions. Is it possible to have more colour on the screen? I know you underline the key letters but it would be nice if they could be in a different colour or a different print.

One subject (QE system), an MA student in Manpower Studies, might be a useful design consultant. In response to the question about system improvements:

> What I did like was when I spelt 'german' wrong the computer prompted me to have another go at that word whereas in the other system [LIBERTAS at PCL] if you get it wrong it's all wrong and you have to start from scratch. I like [the input box] - it focuses your eyes.
>
> If there are any shortcuts which can speed up the process when you make an error so you can switch out and come back in again still keeping all the material you've used so far. I didn't really spot anywhere where you could speed it up in that sense.
>
> The instructions - it's difficult when you've only got a single colour. If you've got colour you can bang them out at people. I noticed instructions were underlined and highlighted but it seemed a little bit packed to me - perhaps if they were spaced out a bit more you could discern them more quickly. Another minor improvement, although it's difficult if you're using windows, is getting instructions in the same place. When you're using these systems you're using books and notes too so your eyes are going away from the screen. If everything's in the same place on each manoeuvre then you're saving a hell of a lot more time. I'm not sure if it's possible using windows.

5.8 How people searched the systems

No interactive system can be assessed without seeing in detail how people really use it. Full transcripts of sessions are extremely long. Appendix 8 contains an illustrated summary of one search and a commentary on another.