# 7 Comparison Search Experiment Participants

## 7.1 Introduction

Although previous researchers had enlisted the Comparison Search Experiment to compare and evaluate computerized library systems, the experiment conducted in this study had three unique aspects: (1) the system recruited library patrons, (2) interviewed them, and (3) recorded their searches and interview responses. Since monitors were not present to encourage patrons to complete the entire Comparison Search Experiment, the researchers expected patrons to leave before they finished the full administration of the Comparison Search Experiment (see Table 6.3).

Before the researchers could submit Comparison Search administrations to the data analysis, they had to manually review individual administrations to determine whether patron search topics were in scope and whether patron responses to pre- and post-search questionnaires were valid. This chapter describes the results of this manual review process.

## 7.2 Michigan Pretest

Prior to the installation of the experimental online catalog in the two participating libraries, the project director conducted a pretest of the Comparison Search Experiment. She recruited 13 students enrolled in the School of Information and Library Studies at the University of Michigan to take part in the pretest. In interviews lasting one hour, the project director observed students and encouraged them to comment on all aspects of the experiment, e.g., user interface, wording of pre- and post-search questions, system functionality. On the basis of the pretest at the University of Michigan, the project team made several changes to the experimental system and to Comparison Search Experiment procedures. Examples of changes are:

- Changed wording of system options, instructions, explanatory messages, and questions and responses on pre- and post-search questionnaires.
- Added response categories to pre- and post-search questionnaires.
- Changed colors of highlighted options.
- Increased timeouts.
- Disabled certain keys on the keyboard.
- Identified and fixed bugs.

## 7.3Installation Period Length

Following the Michigan pretest, the researchers transported and assembled the Gateway microcomputer bearing ASTUTE to the first of two data collection sites — Mardigian Library at the University of Michigan-Dearborn. The system was available to library patrons for a five-week period beginning on November 11, 1993. The Mardigian Library staff liaison instructed evening and weekend staff in system start up, shut down, and backup procedures. The liaison and other library staff took the responsibility of daily system maintenance, i.e., turning the system on and off, making periodic backups of transaction log files. Periodically, the ASTUTE project team would visit Mardigian Library to make sure that the system was functioning properly and make backups of transaction log files.

The project director conducted the Comparison Search Experiment with Mardigian Library staff in the fifth week of system installation. When the project director conducted the first of four interviews with staff, the system unexpectedly crashed. The systems programmer studied the problem and found system and transaction log files had been corrupted when users unexpectedly turned system equipment on and off.

ASTUTE project staff studied logged searches to determine the extent of the damage. They decided that the number of corrupted searches was so large that they doubted the reliability of collected data. They informed the Mardigian Library staff liaison of the problem who invited the project team back to the library for a second installation and data collection period.

Investigating the problem of corrupted system and log files, the ASTUTE project team introduced the following changes to ASTUTE to ensure that files would not be corrupted in future system installations at participating libraries:

- The main problem was that the <Reset> button on the front of the computer enticed users to push it. This sometimes corrupted the database

tables. To solve the problem, we disconnected the wire that connected the <Reset> button to the computer.

- We programmed the autoexec.bat file to automatically restart the ASTUTE experimental online catalog in the event that the machine was restarted or rebooted. This discouraged students from using the machine for other purposes.
- ASTUTE project team members personally monitored the log tables frequently to be sure the machine was in use.
- Checks were also made to ensure that the computer equipment and mouse were functional.
- When several track balls disappeared from the mice at participating libraries, we replaced non-functioning mice with inexpensive mice.
- We added a password protection feature to exit the ASTUTE experimental online catalog. This discouraged students from using the machine for other purposes.

The researchers transported and assembled the Gateway microcomputer bearing the updated ASTUTE to the second of two data collection sites — Lilly Library at Earlham College. The system was available to library patrons for a fifteen-week period beginning on February 23, 1993. The Lilly Library staff liaison instructed evening and weekend staff in system start up, shut down, and backup procedures. The liaison and other library staff assumed the responsibility for daily system maintenance, i.e., turning the system on and off, making periodic backups of transaction log files. In lieu of periodic visits to this distant site to make sure that the system was functioning properly and make backups of transaction log record files, the Lilly Library liaison sent transaction log records to the ASTUTE project team via file transfer protocol (ftp) or ground mail. The ASTUTE project team reviewed the logs as soon as possible to make sure everything was functioning properly.

ASTUTE was installed in Lilly Library for a much longer time period than was given in the original schedule in the project proposal. Lilly Library staff advised the project director to extend the system installation period for several reasons: (1) Earlham College had far fewer students than UM-Dearborn (1,100 vs. 7,400 students), (2) the ASTUTE project team had installed ASTUTE in Lilly Library during the second half of the College's winter quarter and library staff had not had the opportunity to inform students of the system's availability, and (3) keeping ASTUTE installed in Lilly Library through the College's spring semester would give library staff the opportunity to tell students about the system in the many classes staff teach and in their daily interactions with students.

Having installed ASTUTE in Lilly Library, the ASTUTE project team turned their attention to making changes to the ASTUTE experimental online catalog to ensure a successful second administration at Mardigian Library. The team installed ASTUTE in Mardigian Library on March 12, 1993 and left it there for five weeks. This time period coincided with the second half of the university's thirteen-week winter semester. The ASTUTE project team regularly visited Mardigian Library to make sure that the system was functioning properly and to make backups of transaction log record files. The Mardigian Library staff liaison sent transaction log records to the ASTUTE project team via two-day campus mail. The ASTUTE project team reviewed the logs as soon as possible to make sure everything was functioning properly.

Table 7.1 gives the dates of system installation in the two libraries and number of administrations of the Comparison Search Experiment with library patrons and staff.

**Table 7.1. System Installation Information**

| Location | Date | Search Administrations* |
|---|---|---|
| Mardigian Library | Nov. 11–Dec. 14, 1992 | 556 |
| Mardigian Library | Mar. 12–Apr. 19, 1993 | 826 |
| Lilly Library | Feb. 23–May 28, 1993 | 238 |

*Includes library patron and staff searches.

Sections 7.4–7.6 describe the characteristics of usable and unusable searches from the three data collection periods.

## 7.4First Comparison Search Experiment Administration at UM-Dearborn

The first data collection period at UM-Dearborn lasted five weeks from November 11, 1992 to December 14, 1992. The experimental online catalog administered a total of 556 search experiments. The log includes twelve administrations of the Comparison Search Experiment with library staff. The researchers manually reviewed the queries entered in these experiments to determine usable administrations for submission to data analyses. Figure 7.1 divides search administrations into five components for "usable" and "unusable" search administrations.

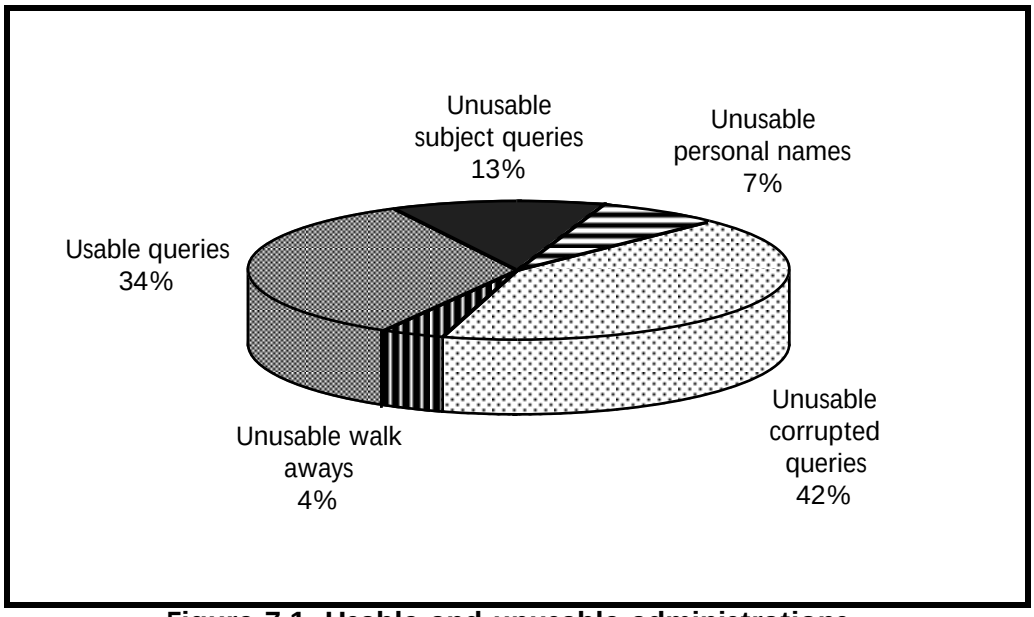


**Figure 7.1. Usable and unusable administrations of the first Comparison Search Experiment at UM-Dearborn**

Reviewing logged data, the ASTUTE project team determined that experimental data logged during the one week preceding and one week following Thanksgiving had been corrupted by the user action of repeatedly turning the microcomputer on and off. The corruption affected 234 searches. Thus, a large number of searches in the middle of the log were unusable. Since the team was concerned about the overall reliability of the data, they disregarded data from the first collection period and planned a second data collection period at UM-Dearborn. The only data they kept were the twelve administrations of the Comparison Search Experiment with library staff.

The ASTUTE project team treated data from this first data collection period as a pretest. The team introduced changes to the experimental online catalog to ensure that data files would not become corrupted in future installations of the catalog in participating libraries and disconnected certain keys on the keyboard.

Search experiments in the "walk-aways" category in figure 7.1 accounted for 4% (22 of 556) of UM-D search administrations. These were search administrations in which users completed one or more pre-search questions but they did not continue with their searches. These users probably walked away from the system and it eventually reset itself to the introductory screen savers.

Unusable administrations accounted for 20% of first-comparison administrations. Unusable administrations of the Comparison Search Experiment were connected with the content of user queries. Such queries probably would have failed to produce useful

retrievals in the experimental online catalog because they were out of scope, that is, they did not match the subject coverage of the catalog's database at UM-Dearborn (i.e., computer science and technology).

"Usable" administrations of the Comparison Search Experiment accounted for about one-third of the queries logged in the first data collection period at UM-Dearborn. Over half (58%) were full administrations of the search experiment. A large percentage (26%) of "usable" search administrations contained all events except a completed post-search questionnaire.

## 7.5 Second Comparison Search Experiment Administration at UM-Dearborn

### 7.5.1 Introduction

The second data collection period at UM-Dearborn lasted five weeks from March 12, 1993 to April 19, 1993. The experimental online catalog administered a total of 826 search experiments. The researchers manually reviewed the queries entered in these experiments to determine usable administrations for submission to data analyses. Subsections below describe the characteristics of usable and unusable search experiments. Figure 7.2 divides search administrations into four components. The largest component was for "usable" search administrations for subject queries.

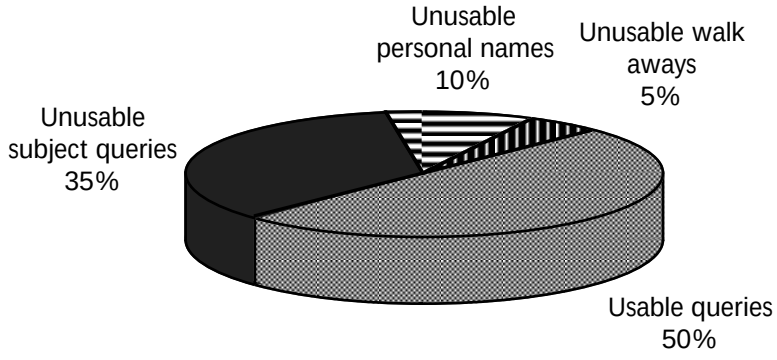

**Figure 7.2. Usable and unusable administrations of the second Comparison Search Experiment at UM-Dearborn**

Of the 826 search administrations, about half (412 of 826 administrations) were usable. Over a third (35%) were unusable queries that users entered using the experimental online catalog's subject searching capabilities for subjects generally. About 10% were unusable queries entered using searching capabilities for personal names. The remainder (5%) were search administrations in which users completed one or more pre-search questions but they did not continue with their searches. These users probably walked away from the system and it eventually reset itself to the introductory screen savers.

Figure 7.3 depicts daily activity in terms of the total number of search administrations and usable search administrations.

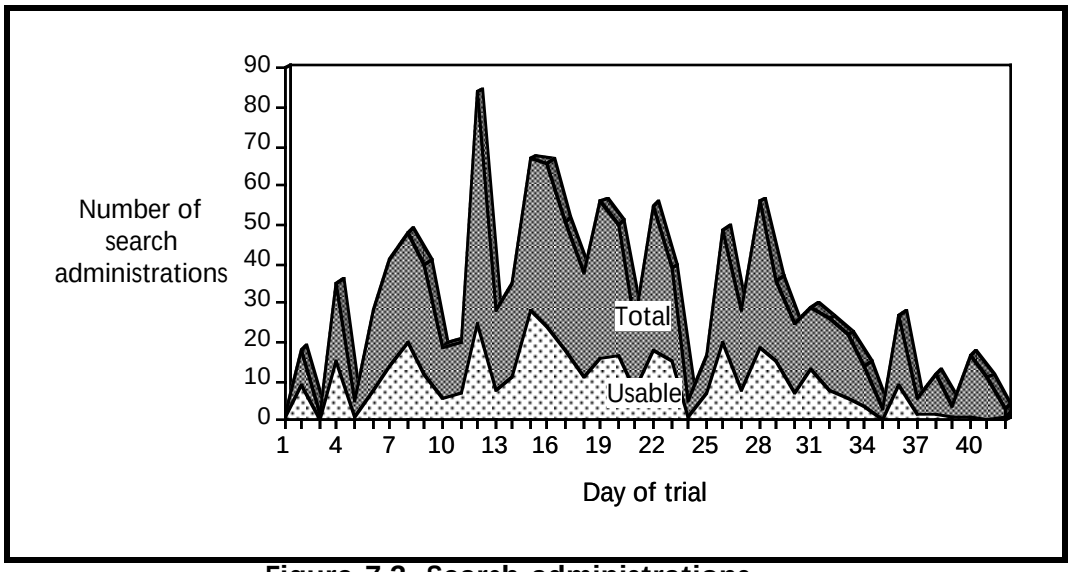


**Figure 7.3. Search administrations at UM-D on a daily basis**

An increase or decrease in the number of usable administrations was accompanied by a corresponding increase or decrease in the total number of unusable administrations. Toward the end of the data collection period, the total number of unusable administrations increased without a concomitant increase in usable administrations. Reviewing the logs, the researchers discovered a key-remapping error that had been introduced into the system on the thirty-fourth day of data collection. When users typed the letter "r," the system understood it as the letter "i." Thus, the researchers placed several administrations with queries such as "etheinet" and "image piocessing" in the unusable category called "system error." These administrations would have been usable

had the keys not been remapped. This might help explain why there were many unusable administrations toward the end of the data collection period.

Figure 7.4 shows daily activity by day of the week. Searching was heaviest in the early part of the week (Tuesdays, Wednesdays, Thursdays). It decreased toward the early part of the weekend and was heavy again on Sundays. In a related study, Neal Kaske (1988, 366) found subject searching especially heavy at the Engineering branch library at the University of Alabama on Wednesdays and weekends.

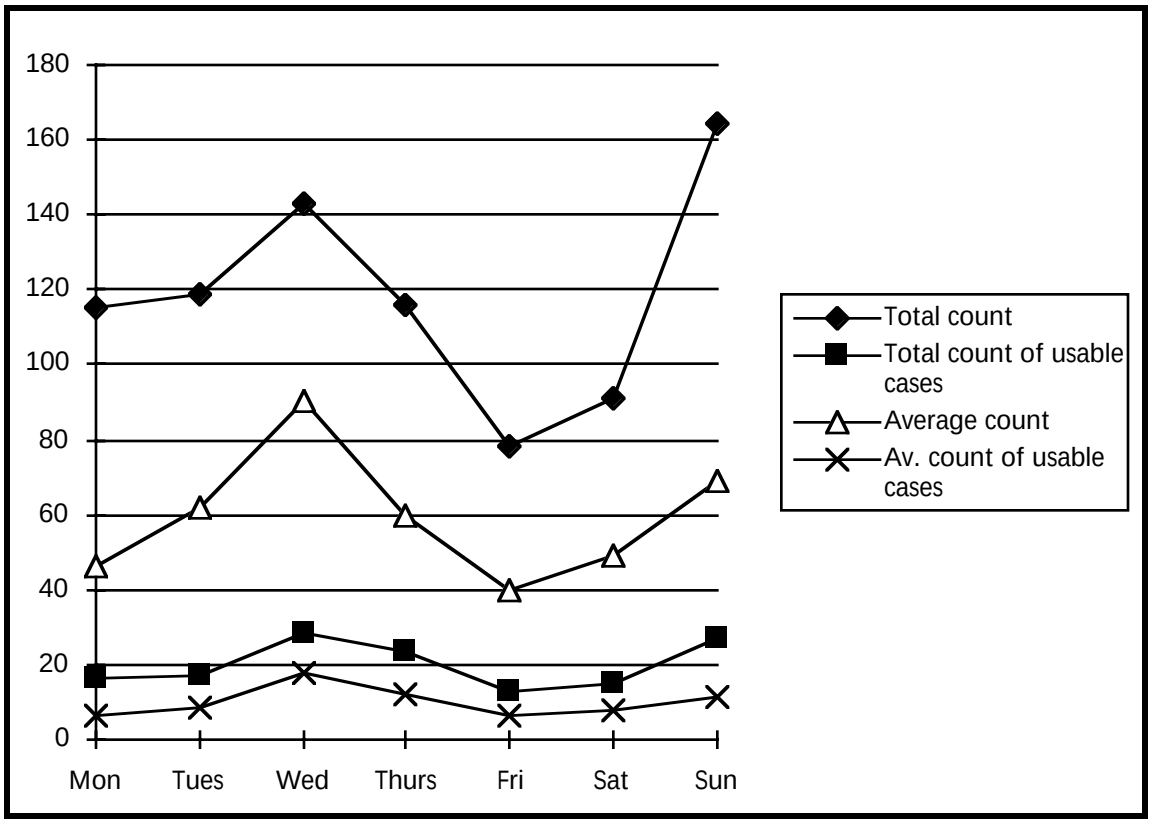


**Figure 7.4. Usable search administrations at UM-D by day of the week**

## 7.5.2 Usable Administrations

Table 7.2 summarizes the characteristics of the 412 usable administrations in the second Comparison Search Experiment at UM-Dearborn.

**Table 7.2. Usable Administrations in the Second Comparison Search Experiment at UM-D**

| Events | Completed events | | | | |
|---|---|---|---|---|---|
| Pre-search questionnaire responses | x | x | x | x | x |
| Pinstripe System search | x | | x | x | x |
| Blue System search | | x | x | x | x |
| Post-search questionnaire responses | | | | | x |
| Suspect post-search questionnaire responses | | | | x | |
| Number of usable search administrations (N=412) | 59 (14%) | 56 (14%) | 55 (13%) | 71 (17%) | 171 (42%) |

A large percentage (42%) of usable administrations of the Comparison Search Experiment were full administrations. Of the four partial administration categories, the largest percentage (17%) contained the four complete events; however, responses to the post-search questionnaire were suspect. That is, post-search questionnaire responses were default responses that users could have selected by repeatedly pressing the <Enter> key or by repeatedly entering the same number. In the data analysis, the researchers ignored suspect post-questionnaire responses and combined these administrations with search administrations in which pre-search questionnaires and searches in Blue and Pinstripe Systems had been completed. The latter administrations accounted for 13% of usable search administrations. When they were combined with the former, the two categories of usable search administrations amounted to almost a third (30%) of usable search administrations. Remaining usable administrations had searches in only one of the two test systems (Blue or Pinstripe) and each of these two categories amounted to 14% of usable administrations.

A total of four librarians took part in the Comparison Search Experiment at UM-D. These librarians contributed twelve full search administrations; however, searches for two full administrations were discarded because of a system error. This left ten full search administrations.

Chapters 8–12 feature the analysis of usable Comparison Search Experiment administrations.

### 7.5.3 Unusable Administrations

The second administration of the Comparison Search Experiment at UM-Dearborn yielded 414 unusable search administrations. Forty search administrations were unusable because users completed one or more pre-search questions but they did not enter a query into the catalog. These users probably walked away from the

experimental online catalog. After four minutes had elapsed, system's timeout function reset itself to the initial introductory screen savers

Table 7.3 summarizes the characteristics of unusable administrations in which users entered queries into the experimental online catalog. The queries in these unusable administrations were entered through subject searching capabilities or through personal-name subject searching capabilities. The system's choice between the two types of capabilities was based on user responses to the question following the pre-search questionnaire that asked users whether their query involved the name of a person.

**Table 7.3. Unusable Administrations of the Second Comparison Search Experiment at UM-D**

| | Capabilities for Subjects | | Capabilities for Personal Names | |
|---|---|---|---|---|
| **Query categories** | No. | % | No. | % |
| Out of scope | | | | |
| Subjects | 128 | 44.1 | 1 | 1.2 |
| Personal names | 10 | 3.5 | 21 | 25.0 |
| Names and subjects | 0 | 0.0 | 9 | 10.7 |
| Playing | | | | |
| Subjects | 17 | 5.9 | 0 | 0.0 |
| Personal names | 3 | 1.0 | 2 | 2.4 |
| Names and subjects | 2 | 0.7 | 10 | 11.9 |
| Title | 2 | 0.7 | 0 | 0.0 |
| Meaningless input: | | | | |
| Blank | 69 | 23.8 | 1 | 1.2 |
| Character | 1 | 0.3 | 0 | 0.0 |
| Gibberish | 19 | 6.6 | 9 | 10.7 |
| Letter | 8 | 2.8 | 0 | 0.0 |
| Letters | 6 | 2.1 | 0 | 0.0 |
| Known-item search | | | | |
| Title search | 3 | 1.0 | 0 | 0.0 |
| Author-title | 0 | 0.0 | 23 | 27.3 |
| Author | 1 | 0.3 | 2 | 2.4 |
| Other unusable searches | | | | |
| Sex term | 10 | 3.4 | 3 | 3.6 |
| Expletive | 2 | 0.7 | 0 | 0.0 |
| Command | 1 | 0.3 | 0 | 0.0 |
| Subject entered as a name | n/a | n/a | 3 | 3.6 |

| | | | | |
|---|---|---|---|---|
| System error | 8 | 2.8 | 0 | 0.0 |
| Total | 290 | 100.0 | 84 | 100.0 |

Most queries in unusable administrations were out of scope, that is, they did not match the subject coverage of the catalog's database at UM-Dearborn (i.e., computer science and engineering). Since the catalog was restricted to subject searches, queries requesting author and/or titles were also considered unusable.

#### 7.5.3.1 Out-of-scope Subject Queries

Of the 290 queries entered through subject searching capabilities, the largest percentage (44.1%) were out of scope. Examples of these queries were "astrophysics," "croatia," "euthanasia," and "activity based costing."

Users averaged almost two access points per search. Two searchers entered as many as seven access points. One of the two searchers entered the query "interplationr." When the Pinstripe System informed the user that it was unable to find material and suggested that the user check the query's spelling, the user responded by clicking on the <Do search> six times before walking away from the experimental system. Thus, the system counted this access point seven times. The other user initially entered the query "saas." He responded to several Blue-system messages informing him that the system was unable to find material by entering new queries, e.g., "soft tissue" and "cancei" [sic].

Of the three subject searching capabilities available to the Pinstripe System for retrieving bibliographic records, the keyword-in-record search was most successful in making retrievals. The Blue System had seven subject searching capabilities available for retrieving bibliographic records. The alphabetical approach was the most successful. The exact approach was also responsible for retrieving bibliographic records. Examples of user queries which were successful in retrieving records through the exact approach were "physics," "differential equations," "trigonometry," and "foreign relations." Although users retrieved records for these and other queries, the researchers felt that the retrieved material did not focus entirely on their topics of interests and that users could retrieve a large quantity of material on such topics in catalog that was encyclopedic in coverage. Even though these queries retrieved bibliographic records, the researchers deemed them out of scope in a catalog limited to computer science and technology material.

Of out-of-scope subject searches, 42.2% of searches were conducted in one of the two experimental online catalogs. The average number of access points per search was comparable to out-of-scope subject searches conducted in both systems.

There was only one out-of-scope *subject query* entered through *personal-name subject searching capabilities*. The query was "medical statistics." The user searched only the Blue System which was unable to retrieve bibliographic records.

#### 7.5.3.2 Out-of-scope Queries for Personal Names

Users entered out-of-scope personal name queries through both general subject searching and personal-name subject searching capabilities. Examples of out-of-scope queries for personal names that were entered using the system's subject searching capabilities were "assad," "bush, g," "plotkin," and "sinclair upton." Users entered about one query per search.

In the Blue System, search trees for personal-name searching invoked keyword approaches before responding to user queries with the alphabetical approach regardless of the extent to which user queries matched an indexed name. Users conducted only two searches for out-of-scope personal names in a single system — the Blue System. The two queries were "hausdorf" (last name) and "kumeso michael matabi" (last name, first name, topic). Both searches in this system gave users the opportunity to select personal-name subject headings through the alphabetical approach. The user in the former search selected the name "Land, Edwin Herbert," and displayed the one bibliographic record bearing this name. The user in the latter search selected the name "Ford, Henry," and displayed the eight bibliographic record bearing this name. Both users gave displayed records "very useful" ratings.

There were nineteen searches in which users searched both Blue and Pinstripe Systems. The Blue System gave users the opportunity to select personal-name subject headings through the alphabetical approach. Although listed names did not refer to the person in the user query, several users selected listed names, displayed bibliographic records, and gave displayed records "very useful" ratings. Examples of user queries and personal-name headings users selected were:

| User query | Selected name heading |
|---|---|
| corniliustacitus | Salomon, Erich |
| pasteur, louis | Zeppelin, Ferdinand |
| tolstoy, leo | Adams, Ansel |
| seppalainen | Shute, Nevil |
| Zhao, X | Zeppelin, Ferdinand |

Three queries processed by the Pinstripe System retrieved bibliographic records through the keyword-in-record approach for personal names. Surprised by these retrievals, ASTUTE project team members repeated these searches to determine why they produced retrievals. They analyzed the results and realized that they had erroneously

included main- and added-entry fields for personal names (USMARC fields 100 and 700) when they had selected bibliographic record fields for the database supporting this search. Since author names were given in these fields, retrievals in keyword-in-record searches for personal names were bolstered by the addition of name elements from these fields. In Pinstripe-system searches, retrievals were also made on words in these author-name fields. Although users gave retrieved records "very useful" relevance assessments, retrieved records were not about the persons named in their original user queries.

#### 7.5.3.3 Out-of-scope Queries for Names and Subjects

Users entered nine queries for names and subjects through personal-name subject searching capabilities. Examples of these queries are given below in terms of the elements users entered:

| Last name | First name | Topic |
|---|---|---|
| tolstoy | peter | history |
| kennedy | robert | assassination |
| charlemagne | charlemagne | education |
| shakespeare | william | actors in shakespeares time |
| piaget | jean | psychology |

The Pinstripe System submitted these queries to keyword-in-record or alphabetical searches. The Blue System tried to find both first and last name and topic elements through keyword-in-record searches. If unsuccessful, the Blue System ignored the first name element in keyword-in-record searches. As a last resort, the Blue System displayed an list of personal-name subject headings in the alphabetical neighborhood of the name elements.

In the nine searches, users searched both Blue and Pinstripe Systems. Six of the nine users began their search in the Pinstripe System. Subject searching approaches in the Pinstripe System failed to retrieve bibliographic records. Since the Blue System always responded with the alphabetical approach, it gave all nine users the opportunity to browse personal-name subject headings and retrieve bibliographic records. Three of the nine users selected listed headings. For example, the user who entered the query on Robert Kennedy selected subject headings such as "Mapplethorpe, Robert" and "Lindbergh, Charles A." The three users displayed a total of six records connected with selected name headings. Even though none of the records specifically mentioned the names they entered, users gave five "very useful" and one "not useful" ratings to the records.

#### 7.5.3.4 Playing in Comparison Search Experiments

Users entered queries in Comparison Search Experiments that were probably not serious queries. The researchers felt users were "playing," and, thus, categorized such queries into "playing" categories. Although a total of twenty-four queries were submitted queries to searching capabilities for subjects generally, the queries were topics, names, titles, and combinations of names and topics. Examples of each are:

- "things" (topic)
- "petty" (name)
- "star trek: the lost generation" (topic)
- "Ronnie rahal the greatest baseball player ever" (combination)

Users could be quite persistent in their "playing." Two users entered six access points. In one of the two searches, the persistent user entered the access points "muslim" (twice), "pitbulls," "nursing" (twice), and "sex." The latter query matched the initial word in the subject heading "Sex discrimination in education, thus, the Blue System responded with the alphabetical approach but the user did not select a listed subject heading. The Pinstripe System responded to the latter query with a keyword-in-subdivided-heading search and displayed the same heading. The user selected this heading, displayed one record, and rated it "not useful." Even though users were "playing," the Blue and Pinstripe Systems responded with the alphabetical approach to several queries. Keyword approaches (e.g., keyword-in-record, title-keyword) were also successful in producing retrievals.

Users were probably playing when the entered twelve queries for which the system performed subject searches for personal names. Examples are queries with the following elements:

| Last name | First name | Topic |
|---|---|---|
| Matt Van Kirk | Matt | Summer term |
| taraban | kimberly | love |
| jakeson | michael | |
| jakeson | michael | singing muise |
| kumeo | michael | ninja turtle part 3 |
| wazir | mohammad | student |

Users conducted searches in both systems except for one query which the users searched in the Blue System. Three of the eleven searches conducted in both systems began in the Pinstripe System. In twelve searches, the Blue System responded with the alphabetical approach. Even though listed personal-name headings were not on the topics specified in

user queries, four users selected personal-name headings from the alphabetical list, displayed a total of nineteen bibliographic records, and rated fourteen of them "very useful," one of them "possibly useful," and the remainder "not useful." For example, the user who entered the query "taraban kimberly love" selected the heading "Taylor, Frederick Winslow," displayed seven bibliographic records, and rated all seven records "very useful."

#### 7.5.3.5 Meaningless Input

Several types of queries constituted meaningless input: blank lines, character(s), gibberish, and letter(s). Sixty-nine subject searches and one personal-name search were characterized by blank lines. These searches were conducted in a single system. Thirty-nine searches began and ended in the Blue System; thirty-one searches began and ended in the Pinstripe System. Most likely users completed the pre-search questionnaire and walked away before entering their queries.

The character "]" was entered in a Pinstripe subject search. It retrieved no records.

A single letter or several letters were entered into fourteen subject searches. Examples of queries placed in the "letter" category are "a," "i," "j," and "s." Examples of queries placed in the "letters" category are "dD," "aa," and "aaaa." The Blue System invoked the alphabetical approach in response to several "letter" queries because the system found longer subject headings beginning with the letter entered by users. For example, the Blue System displayed an alphabetical list of subject headings beginning with the letter "a" in response to the query "a." A user chose the subject heading "APL (Programming language)," displayed over two dozen titles, gave all displayed titles "very useful" ratings, and continued selecting subject headings and cross references from the alphabetical list. The Pinstripe System responded to "letter" queries with keyword searches and the alphabetical approach. Even though their original queries made little sense, users gave displayed bibliographic records "very useful," "possibly useful," and "not useful" ratings.

Queries placed in the "gibberish" category were made up of letters, spaces, numbers, and/or characters. Nineteen subject searches and nine searches for personal names were placed in this category. Examples of gibberish are:

- jkutf[765f76b elina E\\\EEE
- fardin hajji ddJ
- vnccx6rxyrxf
- ihyfol75fg7u 6e4

Fourteen users began their searches in the Blue System and fourteen users began their searches in the Pinstripe System. Although the Pinstripe System occasionally responded with the alphabetical approach, users displayed no bibliographic records in subject searches. In three personal-name searches, both Pinstripe and Blue Systems responded with the alphabetical approach. Even though their original queries made no sense, users gave displayed bibliographic records "very useful," "possibly useful," and "not useful" ratings.

#### 7.5.3.6 Known-item Searches

Users entered three queries through general subject searching capabilities that were probably known-item searches for titles. All three queries were similar: "schaum's" (twice) and "schaum's outline series dynamics." Since title words were included in searchable title-keyword and keyword-in-record databases, searches in Blue and Pinstripe Systems produced relevant retrievals through title-keyword and keyword-in-record searches.

Users entered a total of twenty-three queries bearing name and topic elements through personal-name subject searching capabilities that were probably requests for known items. The researchers searched the University of Michigan's Mirlyn online catalog to verify known-item requests. Although there were few queries for known items, they could be subdivided into the following specific categories: (1) out-of-scope, unverified known-item query, e.g., "jordon albert arcitecture" (sic), (2) out-of-scope, verified known-item query, e.g., "Paine Thomas Common Sense," (3) in-scope, unverified known-item query, e.g., "treddnick nick microprocessors," and (4) in-scope, verified known-item query, e.g., "kuo control systems."

Five queries produced relevant retrievals through keyword-in-record searches in Blue and Pinstripe Systems. Examples of these queries were "kuo control systems," "bastow car suspension and handling," and "juvinall robert machine design." Users retrieved relevant records because of the presence of main- and added-entry headings for personal names in the keyword-in-record database (USMARC fields 100 and 700) that the team erroneously included the searchable database of personal names.

A total of three queries were entered through subject and personal-name searching capabilities that could have been author names: "pascal," "dick," and "dijkstra." Since the user entering the query "pascal" used personal-name searching capabilities, he could have been looking for biographical information on the French scientist Blaise Pascal. In view of ASTUTE's subject coverage (computer science and technology), the user was probably searching for the computer language named Pascal. The user might not have known how to respond to the question that guided the experimental system in

its choice of subject or personal-name searching capabilities because name of the programming language was the same as the name of a famous scholar.

#### 7.5.3.7 Remaining Unusable Categories

Users entered a total of thirteen queries that the researchers placed in the "sex" terms category. Examples of "sex" queries entered through personal-name subject searching capabilities are "el-awad sex" and "Perry James sex." No queries submitted to personal-name subject searching capabilities retrieved titles. Examples of such queries entered through general subject searching capabilities were "sex (three times)," "oral sex," and "pornography." In subject searches, the query "sex" retrieved titles in the Blue System through alphabetical and keyword approaches and in the Pinstripe System through keyword approaches because the subject heading "Sex discrimination in education" was assigned to one bibliographic record. Thus, users had the opportunity to retrieve and display titles in several searches. Users gave two titles "not useful" ratings and one title a "very useful" rating in the two searches in which they displayed titles.

Users entered expletives in two searches. The lone command entered into the Pinstripe System as a query was "quit."

Three queries were computer science or engineering topics entered into the experimental online catalog through personal-name subject searching capabilities. The queries were "joint application development," "lagrange," and "dell computers." The user entering the query "lagrange" could have been looking for biographical information on the French mathematician Joseph-Louis Lagrange. However, a user entered the query "lagrangian dynamics" five minutes before the "lagrange" query. Thus, the researchers felt that the two users were the same person who wanted information on a topical subject — not on a personal name. The user entering the query with the last and first name elements "dell computers," respectively, could have responded positively to the question on persons that guided the system in its selection of personal name subject searching capabilities because company named Dell Computers sounded like it was named after a person with the last name of "Dell."

It was understandable why users responded positively to the question on persons that guided the system in its selection of personal-name subject searching capabilities. Some fine tuning to the wording of this question may be necessary to minimize the likelihood of users taking actions that invoke subject searching approaches for personal names when their queries contain proper adjectives or nouns.

Eight queries were discarded because of system errors. Three queries were discarded because a user action caused the "r" key to be mapped to the "i" key. Five were discarded because the system incorrectly coded user relevance assessments with a "0"

(zero) instead of the numbers 1–3. Both errors occurred during the last week of data collection at the University of Michigan-Dearborn.

## 7.6 Comparison Search Experiment Administration at Earlham College

### 7.6.1 Introduction

The data collection period at Earlham College lasted fifteen weeks from February 23 to May 28, 1993. The experimental online catalog administered a total of 238 search experiments. The researchers manually reviewed the queries entered in these experiments to determine usable administrations for submission to data analyses. Subsections below describe the characteristics of usable and unusable search experiments. Figure 7.5 divides administrations into four components. The largest component was for usable search experiments for subject queries.

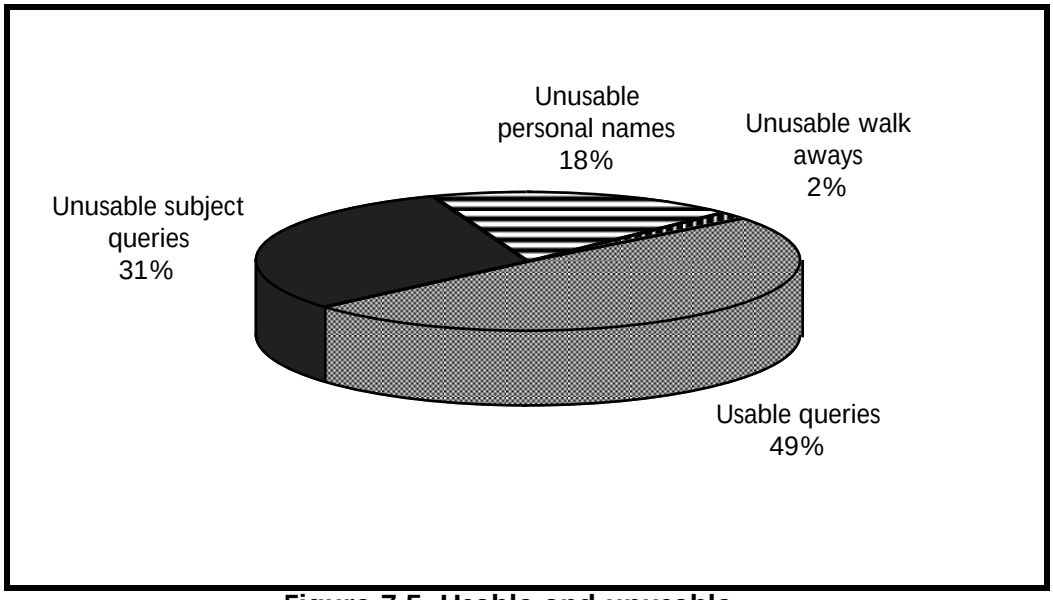


**Figure 7.5. Usable and unusable Comparison Search Experiment administrations at Earlham**

About half of the search administrations were usable (116 of 238 administrations). Usable search administrations include 21 administrations of the Comparison Search Experiment with library staff. A little less than a third of search administrations were unusable queries that users entered using the experimental online catalog's subject searching capabilities for subjects generally. About 18% were unusable queries entered using searching capabilities for personal names. The remainder (2%) were search administrations in which users completed one or more pre-search questions but they did

not continue with their searches. These users probably walked away from the system and it eventually reset itself to the introductory screen savers. The percentages of search experiments in these four categories were comparable to percentages at UM-Dearborn. There were a greater percentage of "walk-aways" and lower percentage of unusable personal-name queries at UM-Dearborn.

Figure 7.6 depicts daily activity in terms of usable and unusable search administrations.

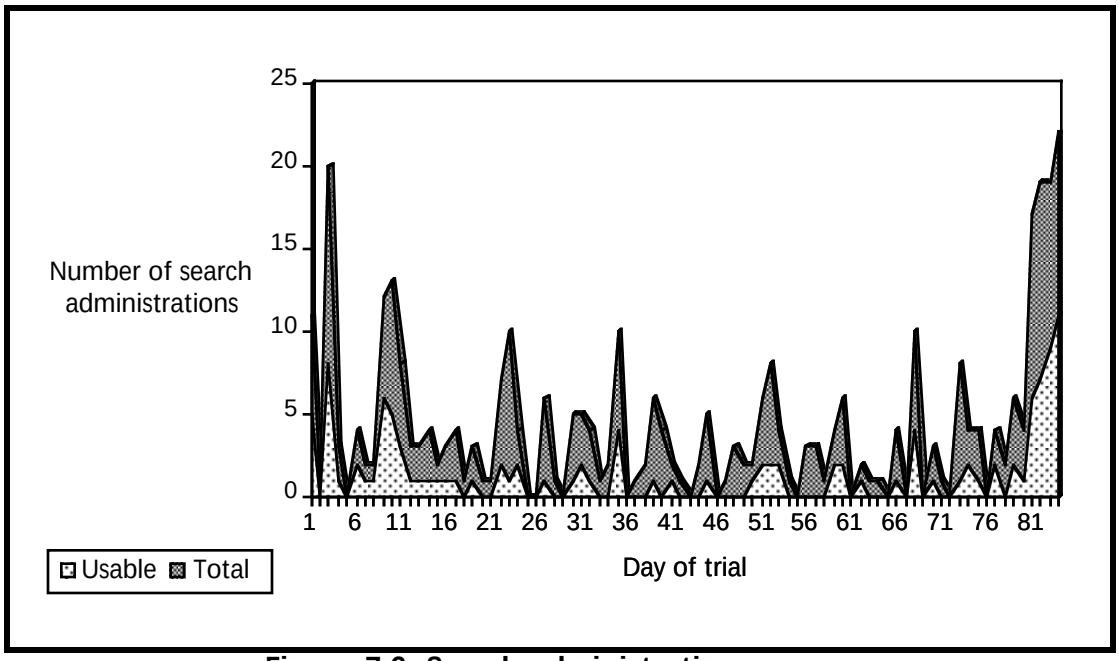


**Figure 7.6. Search administrations at Earlham on a daily basis**

An increase or decrease in the number of usable administrations was usually accompanied by a corresponding increase or decrease in the total number of search administrations. However, there were several days in which all search experiment administrations were deemed unusable. Toward the end of the data collection period, the number of usable administrations increased considerably. The researchers administered the Comparison Search Experiment with library staff at this time. Thus, the interviewer monitored staff searches to make sure they completed pre- and post-search questionnaires and advised them on the appropriateness of their queries in view of the experimental online catalog's subject coverage. This explains why there were many usable administrations toward the end of the data collection period.

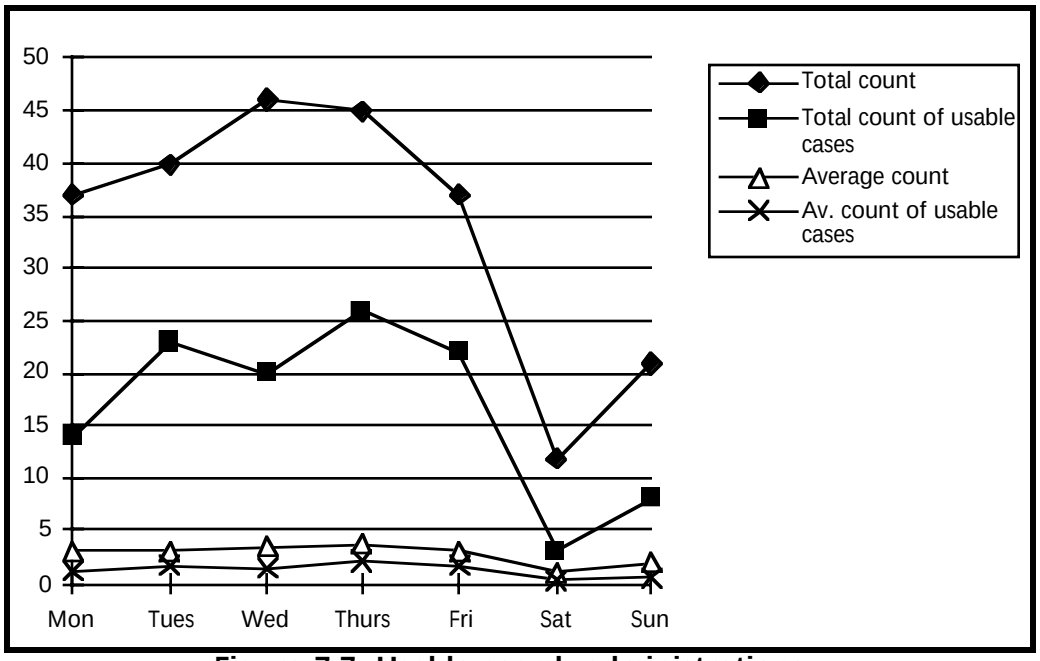


**Figure 7.7. Usable search administrations at Earlham by day of the week**

Figure 7.7 shows daily activity by day of the week. Searching was heavier on weekdays than on weekends.

## 7.6.2 Usable Administrations

Table 7.4 summarizes the characteristics of usable administrations in the Comparison Search Experiment at Earlham College.

**Table 7.4. Usable Comparison Search Experiment Administrations at Earlham**

| Events | Completed events | | | | |
|---|---|---|---|---|---|
| Pre-search questionnaire responses | x | x | x | x | x |
| Pinstripe System search | x | | x | x | x |
| Blue System search | | x | x | x | x |
| Post-search questionnaire responses | | | | | x |
| Suspect post-search questionnaire responses | | | | x | |
| Number of usable search administrations (N=116) | 12 (10%) | 18 (16%) | 20 (17%) | 9 (8%) | 57 (49%) |

The administration of the Comparison Search Experiment at Earlham College yielded 116 usable search administrations. A large percentage (49%) of usable administrations

of the Comparison Search Experiment were full administrations. Of the four partial administration categories, the largest percentage (17%) contained three complete events; the only missing event was the post-questionnaire. About 8% of administrations contained suspect post-questionnaire responses. In the data analysis, the researchers ignored suspect post-questionnaire responses and combined these administrations with search administrations in which pre-search questionnaires and searches in Blue and Pinstripe Systems had been completed. The latter administrations accounted for 17% of usable search administrations. When they were combined with the former, the two categories of usable search administrations amounted to one-quarter of usable search administrations. Remaining usable administrations had searches in only one of the two test systems (Blue or Pinstripe) and accounted for 16% and 10% of usable administrations, respectively.

Percentages of usable search administrations were similar at UM-Dearborn and Earlham. The largest percentages were for full administrations (42% at UM-D; 49% at Earlham). Administrations involving Pinstripe-System searches were more common at UM-Dearborn (14%) than at Earlham (10%). Administrations containing suspect post-search questionnaire responses were higher at UM-Dearborn (17%) than at Earlham (8%).

A total of four librarians participated in the Comparison Search Experiment at Earlham College. These librarians contributed eighteen full search administrations and two partial administrations in which they completed no post-search questionnaires. The partial administrations were database failures. Both librarian and interviewer decided it would be best to forego completing post-search questionnaires for database failures because the librarian could not draw comparisons between systems. Instead, the librarian entered new searches for different topics.

Chapters 8–12 feature the analysis of usable Comparison Search Experiment administrations.

### 7.6.3 Unusable Administrations

The administration of the Comparison Search Experiment at Earlham College yielded 122 unusable search administrations. Four search administrations were unusable because users completed one or more pre-search questions but they did not enter a query into the catalog. These users probably walked away from the experimental online catalog. After four minutes had elapsed, system's timeout function reset itself at the initial introductory screen savers

Table 7.5 summarizes the characteristics of the remaining 118 unusable administrations. In these administrations, users entered queries into the experimental

online catalog. The queries in these unusable administrations were entered through subject searching capabilities and through personal-name subject searching capabilities. The system's choice between the two sets of capabilities was based on user responses to the question following the pre-search questionnaire that asked users whether their query involved the name of a person.

**Table 7.5. Unusable Comparison Search Experiment Administrations at Earlham**

| | Capabilities for Subjects | | Capabilities for Personal Names | |
|---|---|---|---|---|
| **Query categories** | **No.** | **%** | **No.** | **%** |
| Out of scope | | | | |
| Subjects | 57 | 76.0 | 2 | 4.7 |
| Personal names | 5 | 6.7 | 6 | 14.0 |
| Names and subjects | 0 | 0.0 | 29 | 67.4 |
| Names, subjects, and call number | 0 | 0.0 | 1 | 2.3 |
| Playing | | | | |
| Subjects | 2 | 2.7 | 0 | 0.0 |
| Meaningless input: | | | | |
| Blank | 7 | 9.3 | 0 | 0.0 |
| Letter | 1 | 1.3 | 0 | 0.0 |
| Known-item search | | | | |
| Title search | 1 | 1.3 | 0 | 0.0 |
| Author-title | 0 | 0.0 | 5 | 11.6 |
| Other unusable searches | | | | |
| Sex term | 2 | 2.7 | 0 | 0.0 |
| Total | 75 | 100.0 | 43 | 100.0 |

Over 80% of queries in unusable administrations were out of scope, that is, they did not match the subject coverage of the catalog's database at Earlham (i.e., American history). Since the catalog was restricted to subject searches, queries requesting author and/or titles were also considered unusable. Categories for playing, meaningless input, and other unusable searches accounted for much smaller percentages of administrations of the Comparison Search Experiment at Earlham than at UM-Dearborn. Furthermore, subcategories for gibberish, expletives, letters, command, and a few others were not needed to characterize Earlham subject queries.

#### 7.6.3.1 Out-of-scope Subject Queries

Of the 75 queries entered through subject searching capabilities, the largest percentage (76.0%) was out of scope. Examples of these queries were “illegal abortion,” “health conditions south,” “koreans in japan,” and “manhattan project.”

The majority of out-of-scope searches were conducted in both Blue and Pinstripe Test Systems. About the same number of respondents started their searches in the Blue (19) and Pinstripe Systems (17). Users who searched in a single system averaged almost two queries per search. Users who searched both systems averaged a little over one access point per search. One searcher entered six access points: “teinmen,” “steinem” (twice), “virginia woolf” (twice), and “animals.”

Of the three subject searching capabilities available to the Pinstripe System for retrieving bibliographic records, the keyword-in-record search was most successful in making matches. The Blue System used five of the seven subject searching capabilities available for retrieving bibliographic records. The alphabetical approach was the most successful. The exact approach was also responsible for retrieving bibliographic records. Examples of user queries which were successful in retrieving records through the exact approach were “spiritism,” “illness,” “harlem renaissance” (three times), and “natural resources.” Although users retrieved records for these and other queries, the researchers felt that the retrieved material did not focus entirely on their topics of interests and that users could retrieve a large quantity of material on such topics in catalog that was encyclopedic in coverage. Even though these queries retrieved bibliographic records, the researchers deemed them out of scope in a catalog limited to American history.

Twenty-one users entered out-of-scope subject queries in only one of the two experimental online catalogs. One searcher entered four access points. This searcher entered the query “woodstock.” When the Pinstripe System informed the user that it was unable to find material and suggested that the user check the query’s spelling, the user responded by clicking on the <Do search> three times before walking away from the experimental system. Thus, the system counted this access point four times. Only one search in single-system administrations yielded titles. Users displayed retrieved bibliographic records in only one of the two searches in each system and displayed a single title.

There were only two out-of-scope *subject queries* (1.3%) entered through *personal-name subject searching capabilities*. The queries were “woodstock” and “wood stock.” One user searched both Blue and Pinstripe Systems and one user searched only the Blue System. Both users were unable to produce retrievals for these queries.

#### 7.6.3.2 Out-of-scope Queries for Personal Names

Users entered out-of-scope personal-name queries through both general subject searching and personal-name subject searching capabilities. Three of the five personal-name queries users entered through the former were "Tanner, Henry," "marilyn monroe," and "rodin." Users searched both systems for the four of the five queries and the Blue System only for the "rodin" query. Neither the Blue nor the Pinstripe System yielded retrievals for these queries.

Users entered six out-of-scope personal-name queries through personal-name subject searching capabilities in the experimental online catalog at Earlham. Users searched all six queries in the Blue Test System. These six out-of-scope personal-name queries were:

| User query | Selected personal-name heading |
|---|---|
| singer | |
| chopin kate | |
| poe edgar | |
| einstein albert | |
| resnick | Woodring, Harry Hines. |
| stone harlan | Stoner, Anna Louisa Wellington. |

As a last resort, the Blue System always responded to personal-name queries with the alphabetical approach. In these six searches, the system responded with this approach which showed users a list of personal-name headings in the alphabetical neighborhood of entered name element(s). In two of six alphabetical searches, users selected a listed heading and displayed one bibliographic record connected with the heading. Although the original user queries and selected personal-name headings seemed to have nothing in common, users gave "very useful" and "possibly useful" ratings to retrieved titles.

#### 7.6.3.3 Out-of-scope Queries for Names and Subjects

Users entered twenty-nine queries for names and subjects through personal-name subject searching capabilities. Examples of these queries are given below in terms of the elements users entered:

| Last name | First name | Topic |
|---|---|---|
| debs | eugene | labor |
| houdini | harry | spirit mediums |
| charlemagne | charlemagne | education |
| socrates | | philosiphy (sic) |
| jones | jim | biography |

The Pinstripe System submitted these queries to keyword-in-record searches. The Blue System tried to find both first and last name and topic elements through keyword-in-record searches. If unsuccessful, the Blue System ignored the first name element in keyword-in-record searches. As a last resort, the Blue System displayed an list of personal-name subject headings in the alphabetical neighborhood of the name elements.

In the twenty-nine searches, three users searched a single system once — the Blue System. Of the users searching both systems, a few more users started in the Pinstripe System (14) than in the Blue System (12). Subject searching approaches in the Pinstripe System failed to retrieve bibliographic records. Since the Blue System always responded with the alphabetical approach, it gave users the opportunity to browse personal-name subject headings and retrieve titles. Nine users selected listed headings. Examples of original user queries and selected personal-name subject headings were:

| User query | Selected personal-name heading |
|---|---|
| stanton elizabeth suffrage | Steele, Eliza R. |
| welter barbara women's history | Welles, Gideon |
| nash diane civil rights movement | Nasser, Gamal Abdel |
| Baker Ella Student Non-violent coordinating committee | Baker, Newton Diehl |
| Jackson LaToya The life of LaToya Jackson | Jackson, Andrew; Jackson, Rachel |

Eight of nine users who selected listed personal-name headings from alphabetical lists also displayed titles connected with them. Of the twelve titles displayed, users rated almost all "not useful." Although they rated two records "very useful," it was reassuring that users were taking relevance assessments seriously in these instances.

One user query was a combination of call number, name, and subject (or title) elements. The query was "pr 935 s2.6" (last name element), "christy " (first name element), and "cooking or dogs" (topic element). Although the Blue System displayed personal-name headings beginning with the letter "p," the user did not select one.

#### 7.6.3.4 Remaining Subject Queries

The thirteen remaining subject queries were distributed into six query categories. Examples of queries entered in searches placed in the "playing" were "lizzie's big butt," "heidi," and "poop." Both Blue and Pinstripe Systems were unsuccessful in producing retrievals for these queries.

Seven searches were characterized by blank lines. These searches were conducted in a single system with five users searching the Blue System and two users searching the Pinstripe System. In these search administrations, users probably completed the pre-search questionnaire and walked away before entering their queries.

In the one search bearing a single letter ("s"), the keyword-in-subdivided-heading search of the Pinstripe System and the alphabetical approach of the Blue System were successful in displaying subject headings. The user, however, did not select listed heading(s).

The two queries placed in the "sex" category failed to retrieve titles in the Blue or Pinstripe Test Systems.

A single user entered a query for a known item using general subject searching capabilities. The user entered the query "the age of jackson." After submitting it to exact and alphabetical approaches, the Blue System retrieved nine bibliographic records through a title-keyword search because title words make up the searchable database for this search. Of the nine records, the user gave one record a "very useful" rating and the other eight records "possibly useful" and "not useful" ratings.

#### 7.6.3.5 Remaining Personal-name Queries

Users entered a total of five queries for known items using personal-name searching capabilities. Although two queries were entered on different days, they could have been entered by the same user because they asked for the same materials. The queries were "schlesinger aurthur urban history." A similar query "schlesinger arthur the age of jackson" was entered about the same time as the other two queries but not immediately before or after them. The query "parker sarah comanche indians" could have been an author-title search but the researchers could not verify this item in searches of operational online catalogs. Through the Blue System's alphabetical approach, users retrieved titles for two of the five queries but they gave them "not useful" ratings.

## 7.7 Reliability of Relevance Assessments

This was not the first study of user queries to discover that users entered terms that were not serious subject queries. Several studies (Drabenstott and Vizine-Goetz, 1994; Peters 1989; Henty 1986; Walter 1987; Hunter 1991; Dickson 1984; Lester 1991) characterized such queries as "sex terms," "malicious entries," "gibberish," "garbage," "wrong file," etc. Although it seemed unlikely that such queries would produce retrievals, there were several situations in which the Blue and Pinstripe Systems gave users opportunities to display subject headings and bibliographic records. For example,

the Pinstripe System gave users the opportunity to select subject headings and display bibliographic records connected to them because it randomly selected the alphabetical approach in response to user queries regardless of the extent to which they matched listed headings. In personal-name searches, the Blue System responded with the alphabetical approach to name-query elements as a last resort. Keyword searches in both systems were sometimes successful in retrieving records for single letters or sex terms.

The discouraging factor with these retrievals was the preponderance of "very useful" ratings users gave to displayed bibliographic records. Let us illustrate this phenomenon in examples of two searches. In the first example, a user at UM-Dearborn entered the single letter query "j" into the Pinstripe System. The system displayed two titles retrieved in a keyword-in-record search. The user gave one title a "very useful" rating and the other title a "possibly useful" rating even though the two records were on totally different subjects, viz. CAD and jet propulsion systems. In the second example, an Earlham user entered the surname "resnick" into the experimental online catalog. The Blue System responded with the alphabetical approach. The user selected the personal-name heading "Woodring, Harry Hines," displayed the one title connected with it, and gave it a "useful" rating even though it did not mention a person named resnick. (The retrieved title was a biography written by Keith D. McFarland of Harry H. Woodring, Secretary of War under Franklin D. Roosevelt.)

These examples and the relevance assessments made by many other users in searches for unusable queries made us skeptical about the reliability of relevance assessments and post-search questionnaire responses. We examined the method we used to gather such data to determine how it could be changed in future administrations of online questionnaires.

Figure 6.11 shows the Pinstripe System's display of a title retrieved in a search for "civil rights" bearing a pop-up window that gives users three categories for rating usefulness: (1) "very useful," (2) "possibly useful," and (3) "not useful." The user could click on one of these three categories using the mouse or use a combination of arrow keys and <Enter> key to select a category. The first-listed category (i.e., "very useful") was given in white type and set in a medium-orange color against a light-blue background. The other categories were not highlighted but they were given in light yellow type and set in a navy blue background. The simplest action for the user to take was to hit the <Enter> key. This made the system choose the first-listed category, i.e., "very useful" and record it to level 3 transaction log records for relevance assessments. Reviewing relevance assessments for unusable queries, the researchers felt that many users performed this simple action in response to relevance assessment pop-up boxes. This action could have

been more prevalent in searches for unusable queries that were meaningless, were for play, sex terms, or expletives, because users were not serious about their searches.

If the ASTUTE project team could redesign the relevance assessment pop-up window, they would make the selection process a much more deliberate step on the part of the user than hitting the <Enter> key. There are several possibilities and combinations of possibilities for redesigning this window. If the first-listed category were to remain highlighted, it should be changed to a category such as "not useful" or "don't know." Thus, a user's selection of the first-listed category by the simple action of hitting the <Enter> key would not "stack the deck" in favor of "useful" relevance assessments but it would allow for a conservative estimate of useful titles. Another approach would involve requiring users to type in a number or letter corresponding to the category of their choice instead of using arrow keys to select listed categories. This would make their selection more deliberate than repeatedly hitting the <Enter> key. The researchers also felt that users would be more conscientious about selecting categories that reflected the relevance of displayed items if they knew that the system would use such assessments to find additional titles. Thus, online catalogs that feature relevance assessment capabilities should inform users that their assessments will be used for finding additional titles to ensure that users give serious consideration to their relevance assessments.

## 7.8 Reliability of Post-search Questionnaire Responses

The experimental online catalog always concluded search administrations with post-search questionnaires. Although many users whose search administrations were deemed unusable failed to complete post-search questionnaires, about the same number completed them or most of them. Many users who entered out-of-scope queries were serious about their questionnaire responses in that they chose categories that documented the system's inability to produce retrievals for their queries. Other users such as those who were not serious about their queries (e.g., queries in "letter," "sex term," "gibberish" categories) were not serious about their responses to the questionnaire and chose responses that did not make sense in view of their search results. For example, there were searches in which neither system retrieved titles but users chose answers to post-search questions that implied that one or both systems retrieved titles.

Another problem with post-search questionnaire responses was similar to the problem with relevance assessments. In unusable *and* usable search administrations, users repeatedly selected the first-listed response category in post-search questionnaires. A manual review of searches and relevance assessments made the researchers skeptical about the reliability of these post-search questionnaire responses. We examined the

method used to gather such data to determine how it could be changed in future administrations of online questionnaires.

The experimental online catalog displayed each question and response categories separate from other questions and response categories. The first-listed category was given in a bright yellow type and set in a medium-orange color against a dark blue background. The other categories were given in a light yellow type against a dark blue background. The simplest action for the user to take was to hit the <Enter> key. This would make the system choose the first-listed category and save it to level 1 transaction log records.

In the analysis of usable search administrations (see chapters 8–12), the researchers disregarded post-search questionnaires in which users selected most or all first-listed response categories.

If the ASTUTE project team could redesign the post-search questionnaire, they would make the selection process a much more deliberate step on the part of the user than hitting the <Enter> key. There are several possibilities and combinations of possibilities for redesigning the selection process. If the first-listed category were to remain highlighted, it could be changed to a category such as “don’t know” or “don’t care.” First-listed categories could also be different from question to question. When users selected one or more such categories, the truthfulness of their responses to other post-search questions could be in doubt. Another approach would involve requiring users to type in a number or letter corresponding to the category of their choice. This would make their selection more deliberate than repeatedly hitting the <Enter> key.

## 7.9Chapter Summary

This chapter describes the manual review of search administrations that the researchers performed to determine whether patron search topics were in scope and whether their responses to pre- and post-search questionnaires were valid.

The three installations of the experimental online catalog lasted for five- to fifteen-week periods at the two participating libraries. Data collected in the first of two data-collection periods at UM-D were discarded because of a system error that corrupted data for over two-fifths of the search administrations (see section 7.4). The system error was introduced by the unexpected user action of repeatedly turning the microcomputer on and off. The programming team made changes to the experimental online catalog to ensure that the data would not become corrupted in later system installations.

In the second, five-week data-collection period at UM-D, the experimental online catalog administered a total of 826 search experiments (see section 7.5). In the fifteen-week data-collection period at Earlham, the experimental online catalog administered a total of 238 search experiments (see section 7.6).

The manual review of searches revealed that about 50% of search administrations at UM-D and Earlham were usable (see sections 7.5.1 and 7.6.1); however, the manual review of post-search questionnaires and questionnaire responses revealed that large percentages (58% at UM-D, 51% at Earlham) of search experiments were less-than-full administrations (see sections 7.5.2 and 7.5.3). That is, users completed a search in one system, both systems, and/or failed to answer the post-search questionnaire. Since monitors were not present to encourage patrons to complete the entire Comparison Search Experiment, the researchers expected patrons to leave before they finished the full administration of the Comparison Search Experiment. The manual review process was essential because it identified full and less-than-full search administrations which were, on occasion, handled differently in the statistical analysis of search results (see chapter 9). The failure analysis described in chapters 10–12 focuses on both full and less-than-full usable search administrations.

Unusable administrations were divided into categories for queries that were out of scope, for play, meaningless input, known-item searches, and other unusable searches (e.g., sex terms, expletives) (see sections 7.5.3 and 7.6.3). Discussions of unusable search administrations revealed that users selected subject headings, displayed titles, and rated some titles "very useful" even though they were not about the topics users originally entered in their searches. Discouraged with the positive relevance assessments users gave to displayed titles in unusable search administrations, the researchers made suggestions for redesigning the procedures for collecting relevance assessments in future experimental and operational systems (see section 7.7). They also made suggestions for redesigning system-administered post-search questionnaires to minimize the collection of unreliable data in future experimental systems (see section 7.8).

## References

Dickson, Jean. 1984. "An analysis of user errors in searching an online catalog." Cataloging & Classification Quarterly 4, 3 (Spring): 19–38.

Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

Henty, Margaret. 1986. "The users at the online catalogue: A record of unsuccessful keyword searches." *LASIE* 17, 2 (September/October): 47–52.

Hunter, Rhonda N. 1991. "Successes and failures of patrons searching the online catalog at a large academic library: A transaction log analysis." *RQ* 30, 3 (Spring): 395–402.

Kaske, Neal K. 1988. "A comparative study of subject searching in an OPAC among branch libraries of a university library system." *Information Technology and Libraries* 7, 4 (December): 359–72.

Lester, Marilyn Ann. 1989. *Coincidence of user vocabulary and Library of Congress Subject Headings: Experiments to improve subject access in academic library online catalogs*. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Peters, Thomas A. 1989. "When smart people fail: An analysis of the transaction log of an online public access catalog." *Journal of Academic Librarianship* 15, 5 (November): 267–73.

Walter, Dennis R. 1987. "The user at the online catalogue: A record of unsuccessful keyword searches — another case study." *LASIE* 18, 3 (November/December): 74–81.

# 8 Characteristics of System Users and Searches

## 8.1 Introduction

Chapter 8 introduces readers to the characteristics of system users and searches. Characteristics of system users are gleaned from pre-search questionnaire responses. Search characteristics come from an analysis of "usable" search administrations. The analysis reveals interesting findings about the time spent searching, length of user queries, and matches of user queries with the experimental system's controlled vocabulary. Search characteristics include an analysis of titles displayed in the Blue and/or Pinstripe System. The analysis is necessary to identify titles retrieved more than once by one or both systems and such titles given conflicting relevance assessments. This chapter concludes with decisions that the ASTUTE project team made to resolve conflicts in preparation for the statistical analysis in the chapter that follows.

## 8.2Experimental System Users

All respondents whose questionnaires were deemed "usable" answered the three questions in the pre-search questionnaire. The first question in the pre-search questionnaire asked respondents how many times they had used the experimental ASTUTE online catalog. Figure 8.1 summarizes respondent answers at UM-Dearborn and Earlham College to the first pre-search question, viz. "How many times have you tested this system?" Pre-search questionnaire responses came from 412 users at UM-D and 116 users at Earlham.

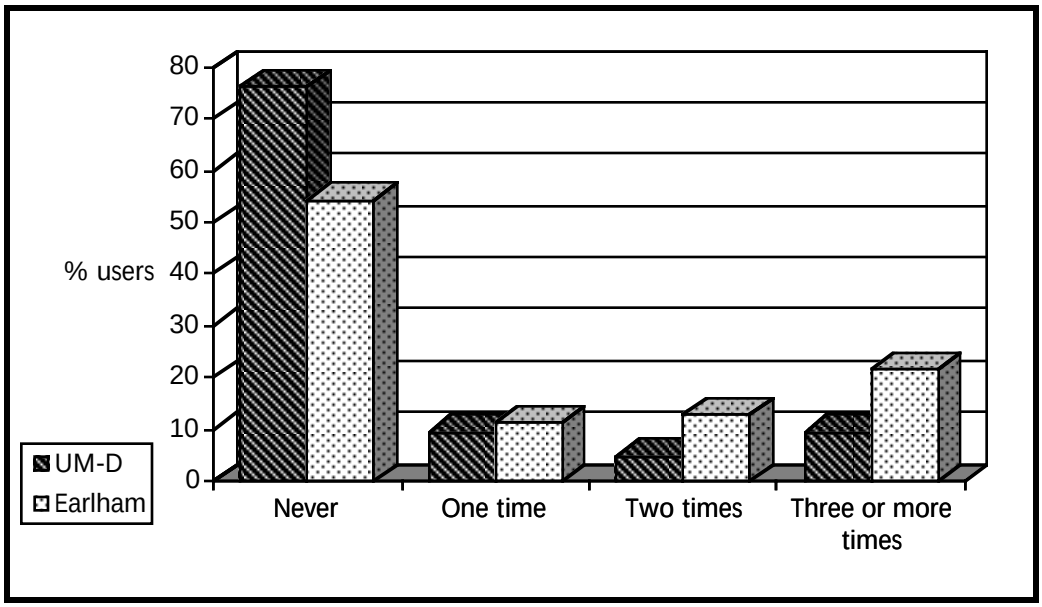


**Figure 8.1 Previous use of ASTUTE**

About three-quarters of UM-D users were searching ASTUTE for the first time. A little less than half of Earlham users were repeat users.

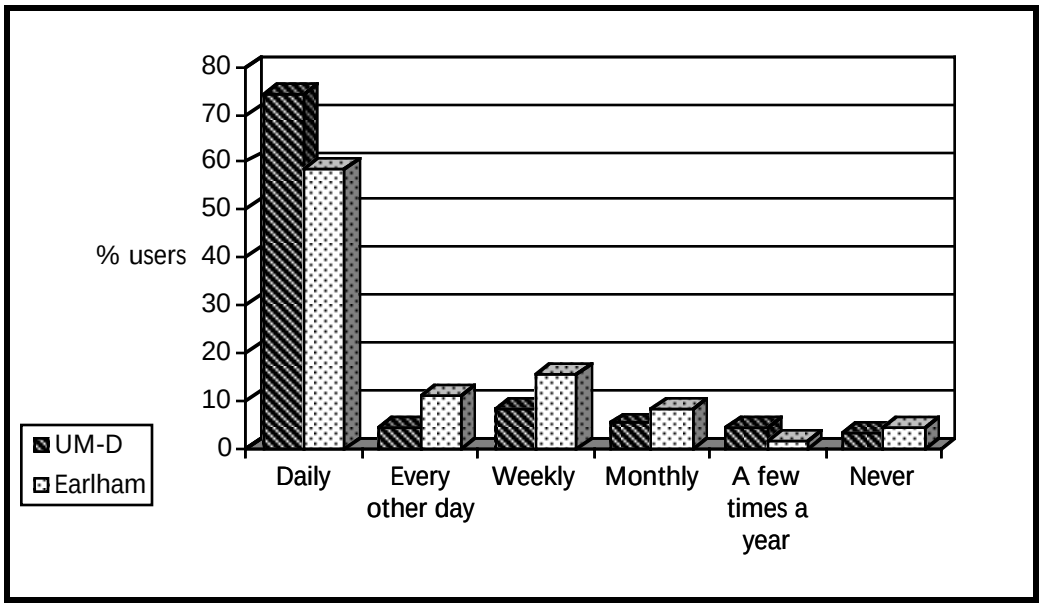


**Figure 8.2. Other computer use**

Figure 8.2 summarizes how respondents completed the statement "I use a computer system other than ASTUTE" which was the second pre-search question.

A large percentage (74.0%) of systems users at UM-D used computers on a daily basis. Such a high percentage was anticipated because ASTUTE's database contained materials on computer science and engineering. Students who studied these fields were likely to be frequent computer users. System users at Earlham were also frequent computer users. Almost seventy percent of Earlham respondents used computers daily or every other day.

Table 8.1 gives the major field of study of experimental system users at UM-D and Earlham.

**Table 8.1. Major Field of Study**

| Major field of study | UM-D (n=412) | Earlham (n=116) |
|---|---|---|
| Computer science | 29.1 | 12.9 |
| Engineering | 46.1 | 0.0 |
| Business | 3.6 | 1.7 |
| Education | 0.5 | 2.6 |
| History | 0.5 | 25.0 |
| Mathematics | 1.2 | 0.9 |
| Other humanities fields | 1.9 | 12.9 |
| Other social science fields | 1.0 | 11.2 |
| Other science fields | 4.4 | 4.3 |
| None of the above | 11.7 | 28.5 |
| Total | 100.0 | 100.0 |

Large percentages of UM-D respondents were computer science or engineering majors (29.1% and 46.1%, respectively). Thus, the majority of UM-D respondents were performing searches for topics in their major field of study. This was not true of respondents at Earlham College where ASTUTE's database focused on American history. Only one-quarter of Earlham respondents were history majors; consequently, the majority of Earlham respondents were not performing searches for topics in their major field of study.

A little under one-quarter of Earlham respondents were majoring in fields in the humanities or social sciences (12.9% and 11.2%, respectively). Although we have no details on the specific fields of respondents who selected these categories, such respondents might be familiar with history from coursework in cognate subjects. The large percentage of computer science majors at Earlham is suspect. This major was the first in a list of ten categories. Respondents could have been in a hurry to complete the pre-search questionnaire and hit the <Return> key (which selected the first-listed category, i.e., computer science) rather than moving the arrow keys down to the appropriate category and hitting the <Return> key. (See section 7.8 for a discussion of suspect questionnaire responses.)

## 8.3Time Spent Searching

Tables 7.2 and 7.4 described the events that respondents completed in usable search administrations. Respondents could have completed as few as two events, i.e., pre-search questionnaire and one search in the Blue- or Pinstripe-test System, or as many as four

events, i.e., pre-search questionnaire, one search in the Blue-test System, one search in the Pinstripe-test System, and post-search questionnaire. Table 8.2 summarizes the average time it took respondents to complete each type of search administration.

**Table 8.2. Time Spent Per Search Administration Type**

| **Events** | **Completed events** | | | | |
|---|---|---|---|---|---|
| Pre-search questionnaire responses | x | x | x | x | x |
| Pinstripe System search | x | | x | x | x |
| Blue System search | | x | x | x | x |
| Post-search questionnaire responses | | | | | x |
| Suspect post-search questionnaire responses | | | | x | |
| **Time spent searching** | | | | | |
| UM-D | 4.3 | 4.6 | 4.9 | 4.6 | 6.4 |
| Earlham | 4.3 | 4.2 | 4.4 | 3.0 | 6.4 |

Respondents who searched a single system or who searched two systems but failed to complete the post-search questionnaire spent between four and five minutes searching. Respondents who searched both Blue and Pinstripe Systems and completed a post-search questionnaire spent the longest time searching — 6.4 minutes.

Search administrations with suspect responses to post-search questionnaires consisted of administrations with default responses that users could select by repeatedly pressing the <Enter> key or by repeatedly entering the same number. At Earlham, such search administrations were quite short — three minutes. At UM-D, the time spent searching in such administrations was comparable to the times spent searching a single system.

Figures 8.3 and 8.4 show the average length of time users took to complete each event in search administrations: (1) searches in the Blue and/or Pinstripe Systems (designated “Search”), (2) three pre-search questions (designated “Pre-search”), and (3) eleven post-search questions (designated “Post-search”) at UM-D and Earlham, respectively. Average search times summarized in figures 8.3 and 8.4 were calculated from 412 searches at UM-D and 116 searches at Earlham.

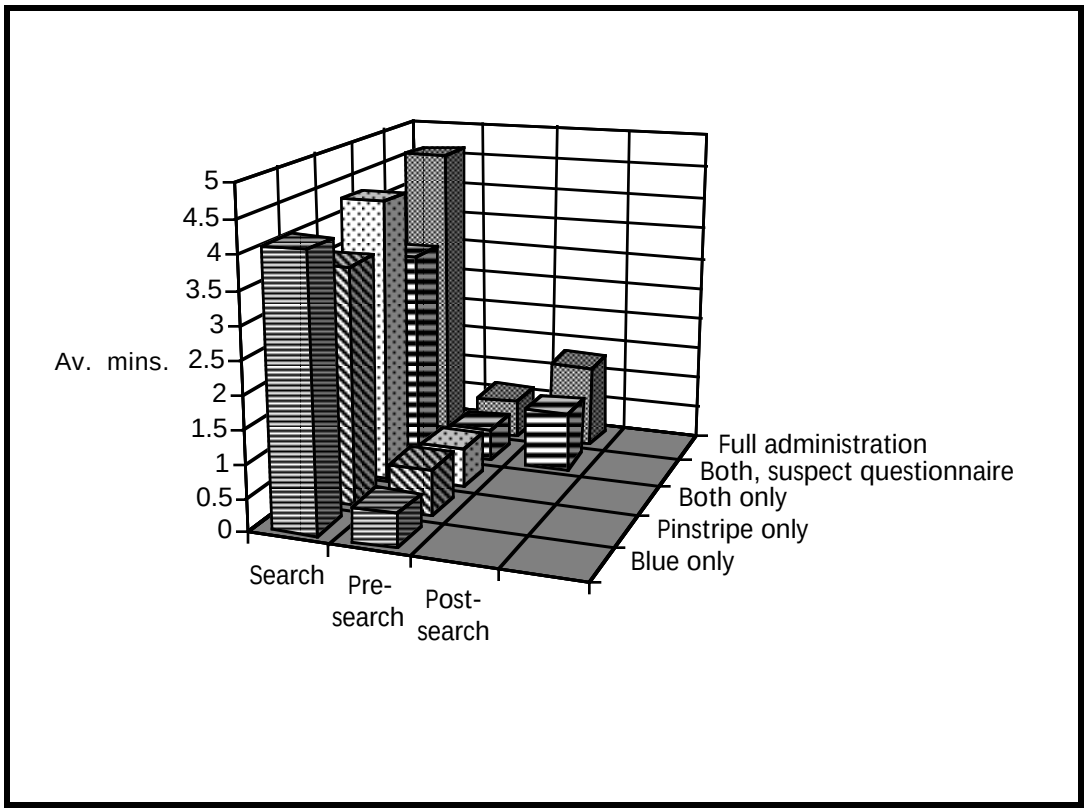


**Figure 8.3. Time spent searching per event at UM-D**

Searchers spent less than a minute completing the pre-search questionnaire. Time spent searching was quite low for search administrations bearing suspect post-search questionnaires — a little more than three minutes at UM-D and two minutes at Earlham. Time spent searching was the longest for full search administrations — over four minutes at both UM-D and Earlham. Completing post-search questionnaires took less than a minute in search administrations with suspect questionnaire responses and took a minute and a half in full search administrations.

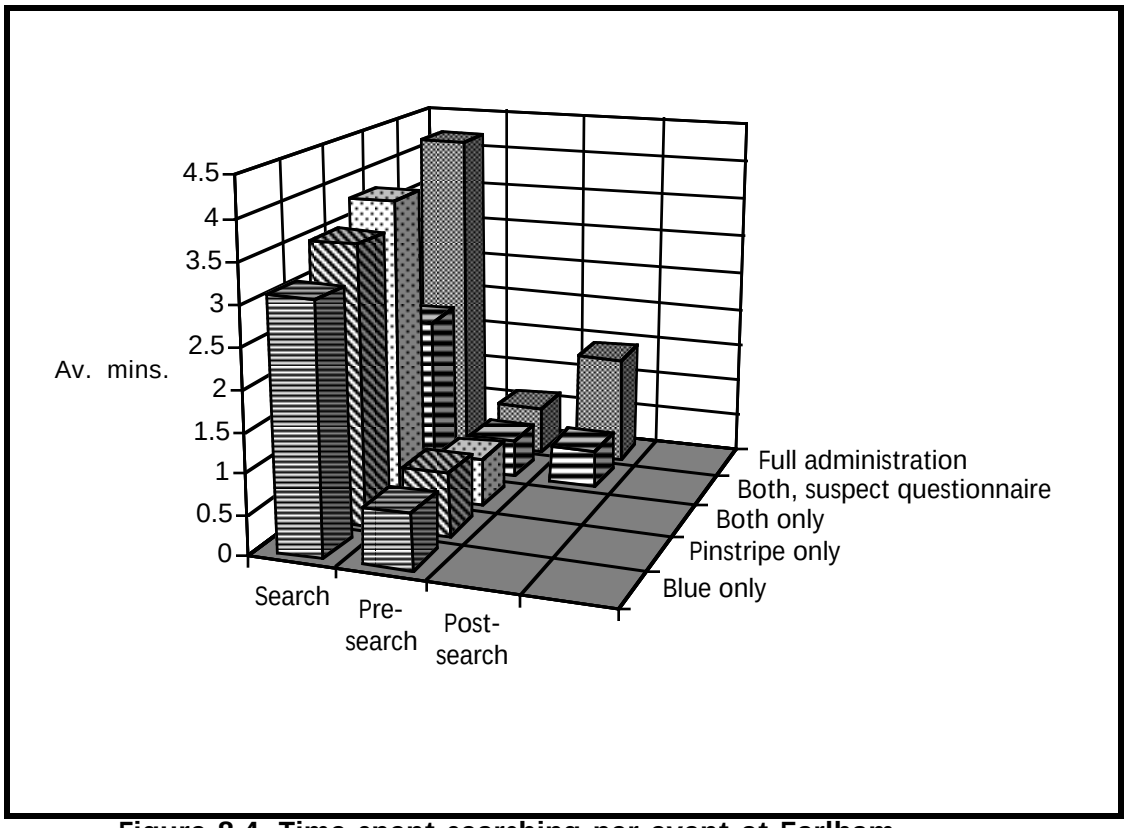


**Figure 8.4. Time spent searching per event at Earlham**

Since end users conducted searches in two systems, it was entirely possible that they would tire of searching the experimental online catalogs and truncate their searches in the second catalog. To determine whether this was true, we reviewed the time spent searching the two systems. Table 8.3 gives the time spent searching the Blue and Pinstripe Systems according to whether they were the first or second system searched. (Times given in Table 8.3 do *not* include search administrations in which users searched only one of the two test systems.)

**Table 8.3. Time Spent Searching Each System**

| | First system searched | | Second system searched | | Total by library | |
|---|---|---|---|---|---|---|
| **Library** | **Blue** | **Pinstripe** | **Blue** | **Pinstripe** | **Blue** | **Pinstripe** |
| UM-D | 2.9 | 2.1 | 1.4 | 1.9 | 2.1 | 2.0 |
| Earlham | 2.8 | 1.2 | 1.9 | 2.5 | 2.3 | 1.9 |
| Total by system | 2.8 | 1.9 | 1.5 | 2.1 | 2.2 | 2.0 |

When UM-D users began their search in the Blue System, they spent 2.9 seconds searching this system and 1.9 seconds searching the Pinstripe System. When UM-D users

began their search in the Pinstripe System, they spent 2.1 seconds searching this system and 1.4 seconds searching the Blue System. Thus, at UM-D, users spent more time searching the first system than the second system. When Earlham users began their search in the Blue System, they spent 2.8 seconds searching this system and 2.5 seconds searching the Pinstripe System. When Earlham users began their search in the Pinstripe System, they spent 1.2 seconds searching this system and 1.9 seconds searching the Blue System. The latter users actually spent more time searching the second system than the first system.

Overall, users spent about two minutes searching each of the two experimental systems. Searches in the Blue System ranged from less than a half minute to about twenty minutes. One UM-D user entered the query "radiator design" into the Blue System. When the Blue System failed to make a match, the user entered the query "heat transfer." The Blue System responded with the exact approach to this cross-reference. The user spent the next twenty minutes browsing subdivided forms of the matched heading, browsing narrower terms, displaying titles assigned to the matched heading and to narrower terms, displaying over fifty titles, and giving varied relevance assessments to retrieved titles. Although the keyword-in-record approach in the Pinstripe System retrieved about ninety titles for the user query "heat transfer," this user only spent a little over two minutes reviewing the results in the Pinstripe System.

A few searches in the Pinstripe System lasted as long as thirty minutes. One UM-D user began searching in the Blue System for "composites" and spent two and a half minutes browsing the results of the alphabetical approach. When this user terminated his search of the Blue System, ASTUTE responded with the results of an alphabetical search in the Pinstripe System. The user browsed the alphabetical list, selected terms of interest, browsed broader and narrower terms connected with selected term, displayed over eighty titles, and varied his relevance assessments of displayed titles. His search of the Pinstripe System lasted almost twenty-nine minutes.

## 8.4User Queries

ASTUTE was designed to handle one query per search experiment administration. That is, it prompted users to enter a query, gave search results in the first system, responded to a user action to start over or terminate the search by switching to the second system and automatically giving search results in the second system, responded to a user action to start over or terminate the search by initiating the post-search questionnaire, and terminated the search at the completion of the questionnaire. Thus, the average number of queries per search was expected to be low in comparison to

studies of operational online catalogs (Markey 1984, 66; Drabenstott and Vizine-Goetz 1994, 169, 186).

When ASTUTE failed to make matches of one or more words in user queries, the system prompted users to check and correct spelling. Some users responded by entering entirely different queries. Table 8.4 summarizes characteristics of user queries.

**Table 8.4. Characteristics of User Queries**

| | UM-D | | Earlham | |
|---|---|---|---|---|
| **Characteristics** | **Blue** | **Pinstripe** | **Blue** | **Pinstripe** |
| No. of queries (av.) | 1.3 | 1.3 | 1.1 | 1.1 |
| No. of queries (range) | 1–8 | 1–6 | 1–8 | 1–6 |
| No. of query words (av.) | 1.8 | 1.6 | 2.9 | 3.0 |
| No. of query words (range) | 1–7 | 1–6 | 1–12 | 1–11 |
| No. of words in queries that make matches (av.) | 1.5 | 1.5 | 2.1 | 2.4 |
| No. of words in queries that fail to make matches (av.) | 2.0 | 1.9 | 3.2 | 4.1 |

At UM-D and Earlham, the average number of queries per search was about the same, about 1.1 or 1.3 queries per search. This low average was expected because of ASTUTE's unique design.

At UM-D, the average number of query words was almost two words. At Earlham, the average number of words in queries was almost three words. The average number of words in queries entered by UM-D respondents was typical for queries for topical subjects and geographical names that were likely to produce retrievals in online catalogs (Drabenstott and Vizine-Goetz 1994, 169, 186). The average number of words in queries entered by Earlham respondents was typical for queries for queries that were not likely to produce retrievals, e.g., queries for personal names, for combinations of names and topics (Drabenstott and Vizine-Goetz 1994, 200, 228, 234).

Queries ranged from one to twelve words. Examples of long queries are:

- digital design with verilog hdl
- fundamental of data structure in c
- king (surname); coretta (given name); scott (middle name); learn about what happened in "1927 "
- king (surname); martin (given name); luther (middle name); about the bycotts they had way back in Mississippi (topic)

- monitoring performance of telephone operators
- women in the westward movement

The last two rows in Table 8.4 give the average number of words in queries that made matches and failed to make matches in the Blue or Pinstripe Systems. When queries made matches in the Blue or Pinstripe Systems, the systems responded with the results of a particular subject searching approach. That is, the systems retrieved titles or subject headings bearing words in user-entered queries. At both UM-D and Earlham, queries that failed to make matches were longer than queries that made matches.

We compared the average length of queries that made matches to the average length of queries that failed to make matches to determine if the difference between the lengths was significant. The results of t-tests examining query length and matches are given in Table 8.5.

**Table 8.5. T-tests Comparing Query Lengths**

| Match type | Number | Mean length | Standard deviation |
|---|---|---|---|
| ***T-test results for Blue System at UM-D*** | | | |
| No match | 172 | 1.9709 | 1.0284 |
| Successful match | 80 | 1.5125 | 0.6363 |
| T = -3.6713 | Degrees of freedom = 250 | | Probability = 0.0003 |
| ***T-test results for Pinstripe System at UM-D*** | | | |
| No match | 121 | 1.8595 | 0.9601 |
| Successful match | 164 | 1.4512 | 0.6296 |
| T = -4.3297 | Degrees of freedom = 283 | | Probability = 0.0000 |
| ***T-test for results for Blue System at Earlham*** | | | |
| No match | 43 | 3.2326 | 2.3077 |
| Successful match | 16 | 2.1250 | 1.2583 |
| T = -1.8153 | Degrees of freedom = 57 | | Probability = 0.0747 |
| ***T-test for results for Pinstripe System at Earlham*** | | | |
| No match | 25 | 4.1200 | 2.2420 |
| Successful match | 44 | 2.4091 | 1.4518 |
| T = -3.8469 | Degrees of freedom = 67 | | Probability = 0.0003 |

All t-tests gave significant results except for the t-test on mean lengths of queries in the Blue System at Earlham. Although the difference in mean lengths was greater than differences in the mean lengths for queries entered into the experimental online catalogs at UM-D (i.e., 1.1076 words vs. 0.4584 or 0.4083 words), the t-test failed to

achieve significance. Its failure to achieve significance could be due to the low sample size for Earlham queries in the Blue System. Generally speaking, the length of queries had a significant effect on the ability of the experimental online catalog to make matches of its subject vocabulary. The longer queries were, the less likely the experimental systems would make a match.

## 8.5Retrieved and Displayed Titles

### 8.5.1 Opportunities to Retrieve Titles

The Blue and Pinstripe Systems responded to user-entered queries that matched indexed subject terms with controlled vocabulary approaches (e.g., exact, alphabetical, keyword-in-heading) or free-text approaches (e.g., title-keyword, keyword-in-record). The former approaches provided users with intermediary browsing lists of subject headings from which users selected headings of interest. The Blue System's selection of a particular subject searching approach was governed by search trees that required queries to make various matches of controlled vocabulary or free-text terms. The Pinstripe System selected approaches randomly.

Search administrations in the Blue or Pinstripe System that contained a "1" in the L2_Results field of transaction log records gave users the opportunity to display bibliographic records. Search administrations in the Blue or Pinstripe System that contained a "0" in the L2_Results field did not give users the opportunity to display bibliographic records (Table F3 in Appendix F). The latter searches were roughly equivalent to searches in which zero titles were retrieved.

ASTUTE always wrote a "1" to the L2_Results field when the Pinstripe System chose the alphabetical approach at random in response to user queries regardless of the closeness of user queries to subject headings in the alphabetical list. For example, the Pinstripe System responded with the alphabetical approach to the user query "treatment of women slaves during the civil war." Examples of listed subject headings were "Treaties," "Treaty of Ghent," "Trenton, Battle of, 1776," and "Trial by jury;" these subject headings had nothing to do with the user's topic of interest. Thus, the existence of a "1" in the L2_Results field did not necessarily mean that the Pinstripe-system's response (subject headings or titles) was useful to users.

In Table 8.6, Blue and Pinstripe searches are categorized according to search administration type. Number columns give the number of search administrations per search administration type. Percentage columns give the percentage of searches which gave users the opportunity to retrieve and display titles (that is, the percentage of

searches in which the system made a match, and, thus, wrote a "1" to the L2_Results field of transaction log records). Table 8.6 summarizes the percentages of Blue and Pinstripe searches that gave users the opportunity to pursue the results of subject searching approaches.

**Table 8.6. "Opportunities" by Search Administration Type**

| | % UM-D | | | | % Earlham | | | |
|---|---|---|---|---|---|---|---|---|
| | Blue | | Pinstripe | | Blue | | Pinstripe | |
| Search administration type | No. | % | No. | % | No. | % | No. | % |
| Search in two systems, no or suspect questionnaire | 126 | 77.8 | 126 | 56.3 | 29 | 75.9 | 29 | 62.1 |
| Search in two systems, questionnaire | 171 | 78.9 | 171 | 73.7 | 57 | 84.2 | 57 | 54.9 |
| Total | 353 | 79.3 | 356 | 73.9 | 104 | 83.7 | 98 | 62.2 |

**Key:**

No. = number of searches in this search administration type

% = percentage of searches with opportunities

Table 8.6 omits results for searches in a single system. Percentages of match for search administrations in which users conducted a search in a single system cannot be compared to one another because the searches in the Blue and Pinstripe Systems were for different topics.

Percentages of opportunities were lower in the Pinstripe System than in the Blue System. The percentages for the Pinstripe System (at both institutions) would have been even lower if we had subtracted the number of Pinstripe-system searches in which the alphabetical browsing list of subject headings produced by the alphabetical approach contained subject headings that did not match the initial word(s) in longer subject headings.

Of the analyses described so far, the analysis summarized in Table 8.6 provides the first comparison of system performance. The analysis demonstrated that the Blue System was likely to provide more opportunities for guiding users to useful information on their topics of interest than the Pinstripe System provided.

### 8.5.2 Retrieving Titles

Table 8.7 summarizes the number of titles retrieved in Blue and Pinstripe searches and categorizes titles by the particular search administration type.

**Table 8.7. Retrievals by Search Administration Type**

| | % UM-D | | | | % Earlham | | | |
|---|---|---|---|---|---|---|---|---|
| | Blue | | Pinstripe | | Blue | | Pinstripe | |
| Search administration type | No. searches | No. titles | No. searches | No. titles | No. searches | No. titles | No. searches | No. titles |
| Search in one system only | 56 | 983 | 59 | 3,124 | 18 | 242 | 12 | 112 |
| Search in two systems, no or suspect questionnaire | 126 | 1,969 | 126 | 5,222 | 29 | 220 | 29 | 1,076 |
| Search in two systems, questionnaire | 171 | 2,967 | 171 | 7,927 | 57 | 694 | 57 | 736 |
| Total | 353 | 5,919 | 356 | 16,273 | 104 | 1,156 | 98 | 1,924 |

Pinstripe-system users retrieved many more titles per search than Blue-system users. This was especially evident for UM-D users who retrieved over two times the number of titles in the Pinstripe System than in the Blue System. High-posted searches in the Pinstripe System were likely to come from the keyword-in-record search which was one of the three search approaches implemented in this system. Keyword-in-record searches made retrievals based on matches of words in all subject-rich fields (see section 5.8.3).

### 8.5.3 Displaying Retrieved Titles

Providing users the opportunity to display titles did not guarantee that users would choose seemingly relevant subject headings from a list that would take them one step closer to browsing titles connected with the subject heading. Furthermore, it did not guarantee that users would browse more than the first title displayed in free-text approaches.

For example, an Earlham user entered the personal name query “jefferson” (last name), “thomas” (first name), and “inventions” (topic) into the Blue System. Exhausting controlled vocabulary and free-text approaches for name and topic elements, the Blue System displayed an alphabetical list of subject headings for personal names that contained the name “Jefferson, Thomas, 1743–1826.” The user selected this name but she did not display one of the several dozen titles bearing this name.

Failing to take advantage of the opportunity to display seemingly relevant bibliographic records was not an uncommon phenomenon. It occurred in both Blue and Pinstripe searches and amongst users at both institutions. In the failure analysis of experimental

system searches (chapters 10–12), this problem was called "perseverance" and was studied in depth to determine why it happened and what system design changes would minimize it (see sections 10.2.1.1, 10.2.2.1, 10.3.1.1, and 10.3.2.1).

Table 8.8 summarizes the number and percentage of titles users retrieved and displayed in Blue- and Pinstripe-system searches.

**Table 8.8. Retrieved and Displayed Titles**

| | % UM-D | | | | % Earlham | | | |
|---|---|---|---|---|---|---|---|---|
| | Blue | | Pinstripe | | Blue | | Pinstripe | |
| **Search administration type** | No. titles retrieved | % titles displayed | No. titles retrieved | % titles displayed | No. titles retrieved | % titles displayed | No. titles retrieved | % titles displayed |
| Search in one system only | 983 | 29.9 | 3,124 | 10.2 | 242 | 35.1 | 112 | 45.5 |
| Search in two systems, no or suspect questionnaire | 1,969 | 26.3 | 5,222 | 13.6 | 220 | 40.9 | 1,076 | 9.4 |
| Search in two systems, questionnaire | 2,967 | 36.6 | 7,927 | 12.9 | 694 | 41.9 | 736 | 21.2 |
| Total | 5,919 | 32.1 | 16,273 | 12.6 | 1,156 | 40.3 | 1,924 | 16.0 |

Experimental-system users retrieved fewer titles in the Blue System than in the Pinstripe System but they displayed a larger percentage of titles per search in the Blue System than in the Pinstripe System. Blue-system users at UM-D displayed nearly one-third of the titles they retrieved. Pinstripe System users at UM-D displayed about one-eighth of the titles they retrieved. At Earlham, Blue-system users displayed about 40% of the titles they retrieved. Overall, Pinstripe-system users at Earlham displayed only 16% of the titles they retrieved.

### 8.5.4 Duplicate Titles in Retrievals

Every time the Blue and Pinstripe Systems displayed a title, they asked searchers to rate the usefulness of the title and gave them three usefulness responses to select: (1) "very useful," (2) "somewhat useful," and (3) "not useful." When users displayed titles more than once, they occasionally gave different relevance assessments. Statistics on displayed titles included titles that users displayed more than once (Tables 8.7 and 8.8). Searchers had many opportunities to display the same title time again and again. Examples of such opportunities are:

- Repeating the display of previously displayed title(s) by selecting <Previous title> or <Next title> options
- Retrieving and displaying the same title(s) in subsequent Blue-system searches invoked by user selection of the <Expand search> option.
- Retrieving and displaying the same title(s) in a Pinstripe-system search following a Blue-system search or in a Blue-system search following a Pinstripe-system search.

At UM-D, Blue- and Pinstripe-system users displayed a total of 470 titles more than once. Of this total, 82.8% of titles were displayed twice across systems, 11.5% were displayed three times across systems, and 5.7% were displayed more than three times across systems. At Earlham, Blue- and Pinstripe-system users displayed a total of 183 titles more than once. Of this total, 80.1% of titles were displayed twice across systems, 12.5% were displayed three times across systems, and 7.4% were displayed more than three times across systems. Three titles were displayed as many as eight times across systems.

### 8.5.5 Resolving Conflicting Relevance Assessments

To compare the Blue and Pinstripe Systems using precision and estimated recall, we had to resolve conflicting relevance assessments. We reviewed relevance assessments for titles displayed more than once to determine how to resolve conflicting relevance assessments. We discussed user relevance assessments with project consultant Paul Kantor and showed him various reports on titles displayed more than one time. We noticed that the more users displayed the same titles, the more likely they were to give positive relevance assessments to titles, i.e., “very useful” or “possibly useful” ratings. After much discussion, we decided to use the last relevant assessment that searchers gave titles displayed more than once and to disregard other relevance assessments for the same title.

Two examples help to describe how we resolved conflicting relevance assessments. Let’s say a user displayed a title twice in the Pinstripe System and gave it a “not useful” the first time she displayed it and a “possibly useful” rating the second time she displayed it. (The user did not display this title in her search of the Blue System.) We kept the second rating, i.e., “possibly useful,” and disregarded the first display and rating of this title in the Pinstripe System. Let’s say a user displayed a title two times in the Pinstripe System and gave it a “not useful” rating followed by a “very useful” rating. The user then switched systems, displayed the very same title in a Blue-system search, and gave it a “possibly useful” rating. We kept the last rating, i.e., “possibly useful,” converted the

first rating of this title in the Pinstripe System to "possibly useful," and disregarded the second and third ratings of this title in the Pinstripe System.

In Table 8.9, the number of titles displayed in the Blue and Pinstripe Systems are given before and after the ASTUTE project team resolved conflicting relevance assessments.

**Table 8.9. Titles Displayed Before and After Resolution of Conflicting Relevance Assessments**

| Search administration type | UM-D | | | | Earlham | | | |
|---|---|---|---|---|---|---|---|---|
| | No. in Blue System | | No. in Pinstripe System | | No. in Blue System | | No. in Pinstripe System | |
| | Titles displayed (before) | Titles displayed (after) | Titles displayed (before) | Titles displayed (after) | Titles displayed (before) | Titles displayed (after) | Titles displayed (before) | Titles displayed (after) |
| Search in one system only | 294 | 286 | 319 | 304 | 85 | 74 | 51 | 50 |
| Search in two systems, no or suspect questionnaire | 519 | 430 | 711 | 653 | 90 | 85 | 101 | 80 |
| Search in two systems, questionnaire | 1,087 | 966 | 1,021 | 954 | 291 | 225 | 156 | 152 |
| Total | 1,900 | 1,682 | 2,051 | 1,911 | 466 | 384 | 308 | 282 |

The numbers of titles displayed before and after conflicting relevance assessments differed by as many as 218 titles (in UM-D's Blue System) and as few as 26 titles (in Earlham's Pinstripe System). Overall, conflicting relevance assessments affected less than ten percent of the titles that users displayed in the experimental systems.

Chapter 9 gives precision and estimated recall of experimental online catalog searches. To calculate these measures, we used the numbers given in the "Titles displayed (after)" columns of Table 8.9. Thus, precision and estimated recall scores discussed in this report were based on users' last relevance assessment of titles displayed in the last system they searched.

### 8.5.6 Rating Displayed Titles

Resolving conflicting relevance assessments required making the following changes to the relevance assessments users gave to displayed titles:

- Changing relevance assessments of titles in the first system searched to match the last relevance assessment users gave to titles in the second system searched.
- Disregarding multiple displays of the same title in the first system searched.
- Disregarding multiple displays of the same title in the second system searched.

When we made the above changes, there was one relevance assessment per unique title displayed in each of the two experimental systems. The one relevance assessment was the same across the two systems. If a user displayed the same title in both Blue and Pinstripe searches, there were two relevance assessments for the title — one assessment in the Blue System and one assessment in the Pinstripe System. Furthermore, the two assessments were the same and were based on the last relevance assessment users gave to the title in the second system searched.

Table 8.10 compares the number and percentage of titles users displayed in the Blue and Pinstripe Systems at UM-D and Earlham according to the three relevance assessments: (1) “not useful,” (2) “possibly useful,” and (3) “very useful.” Percentages of titles in these three relevance assessment categories are given before and after conflicting assessments were resolved.

**Table 8.10. Comparing Relevance Assessments Before and After Resolution of Conflicts**

| Institution/ system | No. of titles dis- played | No. Not useful | % Not useful | No. Possibly useful | % Possibly useful | No. Very useful | % Very useful |
|---|---|---|---|---|---|---|---|
| UM-D/ Blue (before) | 1,900 | 434 | 22.8 | 234 | 12.3 | 1,232 | 64.9 |
| UM-D/ Blue (after) | 1,682 | 413 | 24.6 | 193 | 11.4 | 1,076 | 64.0 |
| UM-D/ Pinstripe (before) | 2,051 | 595 | 29.0 | 249 | 12.1 | 1,207 | 58.9 |
| UM-D/ Pinstripe (after) | 1,911 | 579 | 30.3 | 223 | 11.7 | 1,109 | 58.0 |
| Earlham/ Blue (before) | 466 | 204 | 43.8 | 92 | 19.7 | 170 | 36.5 |
| Earlham/ Blue (after) | 384 | 193 | 50.3 | 68 | 17.7 | 123 | 32.0 |
| Earlham/ Pinstripe (before) | 308 | 82 | 26.6 | 64 | 20.8 | 162 | 52.6 |
| Earlham/ Pinstripe (after) | 282 | 88 | 31.2 | 59 | 20.9 | 135 | 47.9 |

At UM-D, percentages of titles in the three relevance-assessment categories varied by less than 2% after conflicts were resolved. At Earlham, such percentages varied by as little as 0.1% and as much as 6.5%.

We compared sets of titles in each of the three relevance-assessment rating categories using t-tests to determine if the resolution of conflicting relevance assessments made a significant difference in the size of sets. The results of all t-tests were not significant. The t-test on titles rated “not useful” at Earlham came closest to producing significant results ($p=0.0424$). The closest any other test came to producing significant results was the t-test on titles rated “very useful” at Earlham ($p=0.1100$). Generally speaking, the resolution of conflicting relevance assessments had no significant effect on sets of rated titles.

## 8.6Chapter Summary

Chapter 8 examined the characteristics of system users and searches. Characteristics of system users were gleaned from pre-search questionnaire responses (see section 8.2). Search characteristics came from an analysis of "usable" search administrations (see sections 8.3–8.5).

The majority of users at both UM-D and Earlham were searching ASTUTE for the first time (see figure 8.1). The majority were also daily users of computer systems. (see figure 8.2) Large percentages of UM-D respondents were computer science or engineering majors (29.1% and 46.1%, respectively). Thus, the majority of UM-D respondents were performing searches for topics in their major field of study. This was not true of respondents at Earlham College where ASTUTE's database focused on American history. Only one-quarter of Earlham respondents were history majors; consequently, the majority of Earlham respondents were not performing searches for topics in their major field of study (see Table 8.1).

Searchers did not spend much time using the ASTUTE experimental online catalog (see section 8.3). They spent less than a minute completing the pre-search questionnaire (see figures 8.3 and 8.4). Time spent searching was quite low for search administrations bearing suspect post-search questionnaires — a little more than three minutes at UM-D and two minutes at Earlham. Time spent searching was the longest for full search administrations — over four minutes at both UM-D and Earlham. Completing post-search questionnaires took less than a minute in search administrations with suspect questionnaire responses and took a minute and a half in full search administrations.

ASTUTE was designed to handle one query per search experiment administration (see section 8.4). Thus, the average number of queries per search (1.1 or 1.3) was much lower than the average number in operational online catalogs. A statistical test demonstrated that the length of queries had a significant effect on the ability of the experimental online catalog to make matches of its subject vocabulary (see Table 8.5). Long user queries (three or more words) were less likely to make matches than short queries (two or fewer words).

The first comparison of system performance was given in an analysis of system responses that designated matches of user queries (see Table 8.6). Matches gave users the opportunity to pursue the results of subject searching approaches. Generally, the Blue System was more likely to provide more opportunities for guiding users to useful information on their topics of interest than the Pinstripe System.

Providing users the opportunity to display titles did not guarantee that users would choose seemingly relevant subject headings from a list that would take them one step

closer to retrieving titles connected with the subject heading. Furthermore, it did not guarantee that users would browse more than the first title displayed in free-text approaches. Users seldom displayed all the titles they retrieved, in fact, the statistics in Table 8.8 showed that they displayed less than 20% of the titles they retrieved. At UM-D, the Pinstripe System retrieved more than two times as many titles as the Blue System. High-posted searches in the Pinstripe System were likely to come from keyword-in-record searches; this search made retrievals based on matches in all subject-rich fields. Pinstripe-system users at UM-D displayed only about an eighth of the titles they retrieved.

At UM-D, Blue- and Pinstripe-system users displayed a total of 470 titles more than once. Of this total, 82.8% of titles were displayed twice across systems, 11.5% were displayed three times across systems, and 5.7% were displayed more than twice across systems. At Earlham, Blue- and Pinstripe-system users displayed a total of 183 titles more than once. Of this total at Earlham, 80.1% of titles were displayed twice across systems, 12.5% were displayed three times across systems, and 7.4% were displayed more than three times across systems. Section 8.5.5 tells how conflicting relevance assessments were resolved. The resolution of such conflicts did not have a significant effect on sets of rated titles. Resolution was necessary, however, to submit the data to statistical procedures to measure precision and estimated recall which was the focus of the data analyses in Chapter 9.

## References

Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

Markey, Karen. 1984. *Subject searching in library catalogs: before and after the introduction of online catalogs.* Dublin, Ohio: OCLC.

# 9 Precision and Estimated Recall in the Blue and Pinstripe Systems

## 9.1Introduction

Chapter 9 features a discussion of relevance ratings in Blue- and Pinstripe-system searches and uses relevance ratings to measure and compare the search performance of the two experimental systems. This chapter also features end-user responses to the post-search questionnaire. Post-search questions probed user assessments of system performance, ease of system use, and importance of system features.

## 9.2Precision in Experimental Catalog Searches

### 9.2.1 Overall Precision Scores

Precision scores were based on relevance assessments users gave to titles the last time they saw them in Blue- or Pinstripe-system searches. Precision scores for individual searches were calculated by dividing the number of titles retrieved and displayed and given "very useful" ratings in a search by the total number of titles retrieved and displayed in a search.

An example would help explain how precision is calculated. Let's say a user searched the Blue System first and displayed ten titles. She gave displayed titles five "not useful" ratings, three "possibly useful" ratings, and two "very useful" ratings. The user continued searching in the Pinstripe System and displayed six titles. She gave displayed titles two "not useful" ratings, two "possibly useful" ratings, and two "very useful" ratings. Two of the sixteen titles displayed were displayed more than once. One of the ten titles displayed in the Blue System was the same as one of the titles displayed in the Pinstripe System. The user rated this title "possibly useful" in the Blue System and "very useful" in the Pinstripe System. Since the user displayed this title last in the Pinstripe search, the relevance assessment for this title in the Blue System was changed to "very useful."

Of the six titles displayed in the Pinstripe System, one title was displayed twice. The first time the user displayed this title, she gave it a "not useful" rating; the second time the user displayed this title, she gave it a "possibly useful" rating. The relevance assessment for the first time this title was displayed was disregarded. Thus, this title retained the second relevance assessment for "possibly useful." Precision in the Blue System would be 30% (two "very useful" ratings and one "possibly useful" rating converted to "very useful" divided by ten displayed titles). Precision in the Pinstripe System would be 40% (two "very useful" divided by five displayed titles).

Figures 9.1 and 9.2 give average precision scores for the Blue and Pinstripe Systems according to search administration type at UM-D and Earlham.

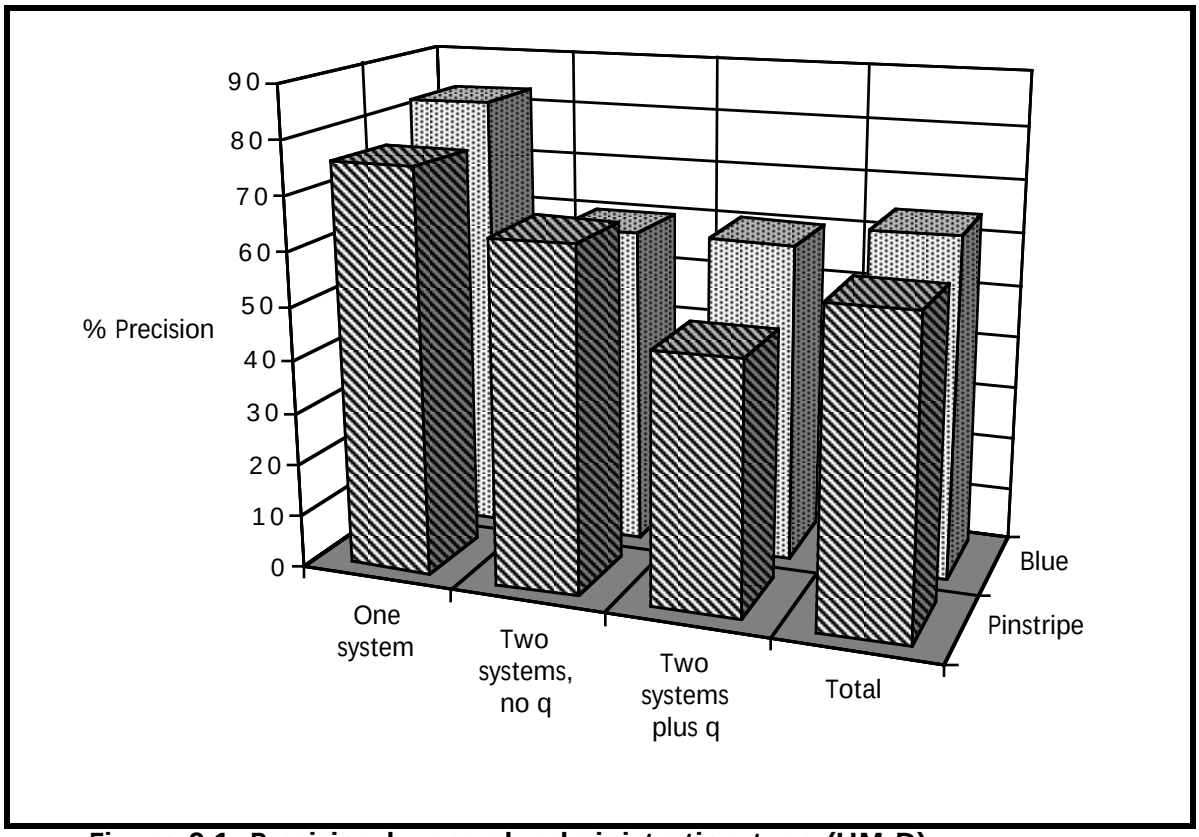


**Figure 9.1. Precision by search administration type (UM-D)**

At UM-D, precision scores averaged 64% in the Blue System and 58% in the Pinstripe System. Except for precision scores for search administrations in which users failed to complete questionnaires or gave suspect responses, precision scores in the Blue System exceeded precision scores in the Pinstripe System.

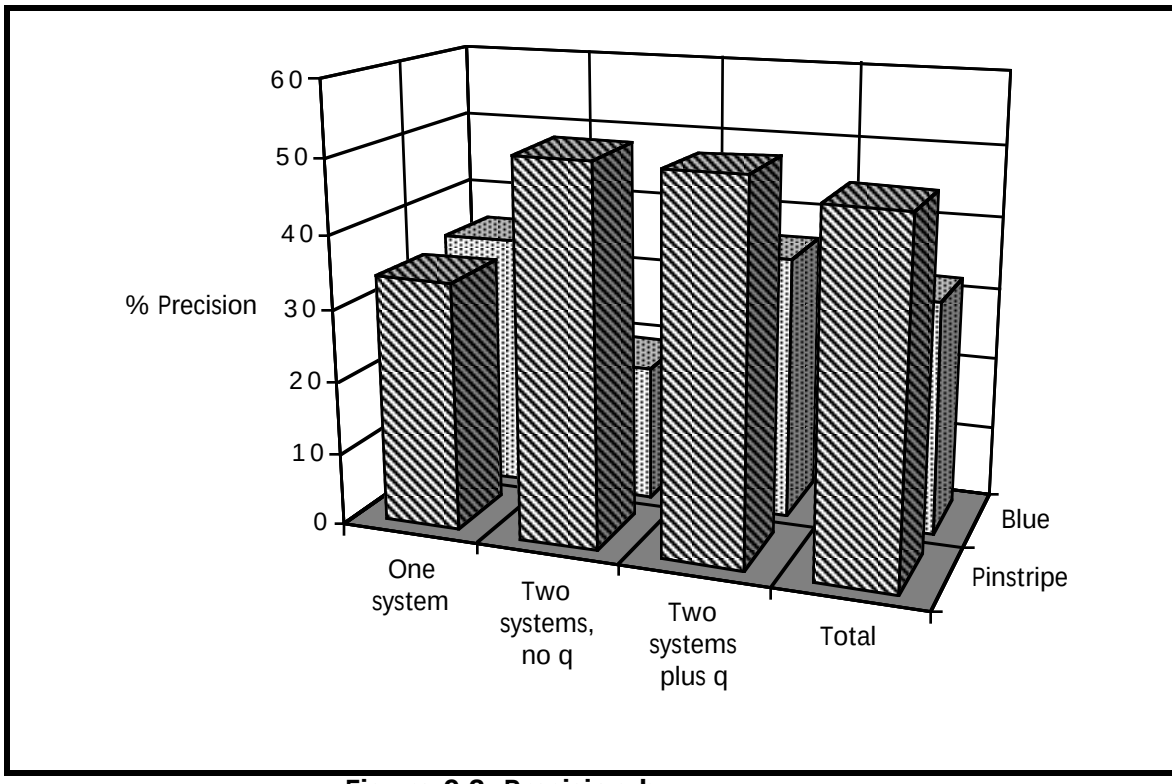


**Figure 9.2. Precision by search administration type (Earlham)**

The results were almost opposite at Earlham where precision scores averaged 32.0% in the Blue System and 47.9% in the Pinstripe System. The difference in precision scores was most evident for search administrations in which users failed to complete questionnaires or gave suspect responses; however, these precision scores were computed from "very useful" retrievals in 29 pairs of searches. The number of searches contributing "very useful" retrievals to "one-system" search administrations was 18 searches in the Blue System and 12 searches in the Pinstripe System. Thus, the numbers of cases for these two types of search administrations were quite small.

We compared precision scores for the two types of search administrations in which users conducted searches for the same topics in Blue and Pinstripe Systems to determine if the difference between scores was significant. The results of t-tests comparing precision scores are given in Tables 9.1 and 9.2. These tables limit t-test results to search administrations in which the same users conducted searches for the same topics in both Blue and Pinstripe Systems.

**Table 9.1. T-test Results for Precision (UM-D)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***T-test results for two systems, no or suspect questionnaire*** | | | |
| Blue System | 653 | 0.6432 | 0.4794 |
| Pinstripe System | 430 | 0.6000 | 0.4905 |
| T = 1.4372 | Degrees of freedom = 1,081 | | Probability = 0.1510 |
| ***T-test results for two systems plus questionnaire*** | | | |
| No match | 966 | 0.6014 | 0.4899 |
| Successful match | 954 | 0.4706 | 0.4994 |
| T = 5.7929 | Degrees of freedom = 1,918 | | Probability = 0.0000 |
| ***Total (T-test results for two-system administrations)*** | | | |
| No match | 1,396 | 0.6010 | 0.4899 |
| Successful match | 1,607 | 0.5408 | 0.4985 |
| T = 3.3298 | Degrees of freedom = 3,001 | | Probability = 0.0009 |

At UM-D, significant results were obtained for full search administrations and for the total of the two search administrations in which the same users conducted searches for the same topic in both Blue and Pinstripe Systems. Generally, precision scores in UM-D's Blue System were significantly greater than precision scores in UM-D's Pinstripe System.

At Earlham, significant results were obtained for all three analyses. Generally, precision scores in Earlham's Pinstripe System were significantly greater than precision scores in Earlham's Blue System.

**Table 9.2. T-test Results for Precision (Earlham)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***T-test results for two systems, no or suspect questionnaire*** | | | |
| Blue System | 85 | 0.1882 | 0.3932 |
| Pinstripe System | 80 | 0.5125 | 0.3932 |
| T = -4.6282 | Degrees of freedom = 163 | | Probability = 0.0000 |
| ***T-test results for two systems plus questionnaire*** | | | |
| Blue System | 225 | 0.3600 | 0.4811 |
| Pinstripe System | 152 | 0.5066 | 0.5016 |
| T = -2.8524 | Degrees of freedom = 375 | | Probability = 0.0046 |
| ***Total (T-test results for two-system administrations)*** | | | |
| Blue System | 310 | 0.3129 | 0.4644 |
| Pinstripe System | 232 | 0.5086 | 0.5010 |
| T = -4.6929 | Degrees of freedom = 540 | | Probability = 0.0000 |

### 9.2.2 Precision of Controlled Vocabulary Searches

The Blue and Pinstripe Systems featured controlled vocabulary and free-text approaches to subject searching. The Blue System's search trees favored controlled vocabulary approaches over free-text approaches in response to user queries. The Pinstripe System selected randomly amongst two controlled vocabulary and one free-text approaches.

Controlled vocabulary approaches to subject searching in the Blue System were: (1) exact, (2) alphabetical (for subjects and for personal names), (3) keyword-in-main-heading, and (4) keyword-in-subdivided-heading searches. The Pinstripe System featured the second and fourth controlled vocabulary approaches listed above. Free-text approaches to subject searching in the Blue System were: (5) title-keyword, (6) keyword in subject heading fields, and (7) keyword-in-record (for subjects and for personal names) searches. The Pinstripe System featured the seventh free-text approach listed above.

We categorized relevance ratings as to the particular approach (controlled vocabulary or free-text) responsible for their retrieval and display. We then calculated precision scores. Figures 9.3–9.5 compare precision scores achieved through controlled vocabulary searches to precision scores achieved through free-text approaches.

Figure 9.3 shows precision scores for UM-D's Blue System.

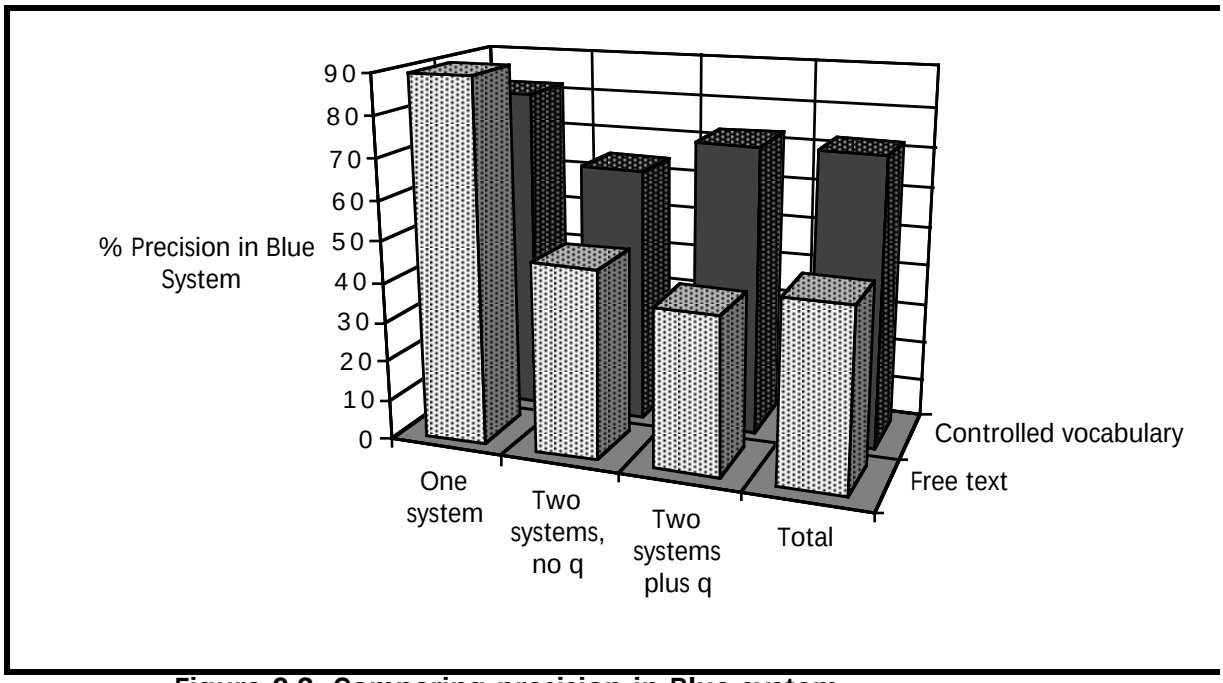


**Figure 9.3. Comparing precision in Blue-system controlled vocabulary and free-text searches (UM-D)**

In UM-D's Blue System, precision scores for controlled vocabulary approaches averaged 71.9% and precision scores for free-text approaches averaged 45.0%. Except for precision scores for one-system search administrations, precision in controlled vocabulary searches exceeded precision in free-text approaches.

We compared precision scores to determine if the difference between the scores for controlled vocabulary and free-text approaches was significant. The results of t-tests comparing precision scores in UM-D's Blue System are given in Table 9.3.

**Table 9.3. T-test Results for Precision in the Blue System (UM-D)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***T-test results for Blue System, no or suspect questionnaire*** | | | |
| Controlled vocabulary | 328 | 0.6402 | 0.4807 |
| Free text | 102 | 0.4706 | 0.5016 |
| T = 3.0812 | Degrees of freedom = 428 | | Probability = 0.0002 |
| ***T-test results for Blue System plus questionnaire*** | | | |
| Controlled vocabulary | 612 | 0.7222 | 0.4483 |
| Free text | 354 | 0.3927 | 0.4890 |
| T = 10.6458 | Degrees of freedom = 964 | | Probability = 0.0000 |
| ***Total (T-test results for Blue System administrations)*** | | | |
| Controlled vocabulary | 940 | 0.6936 | 0.4612 |
| Free text | 456 | 0.4101 | 0.4924 |
| T = 10.5341 | Degrees of freedom = 1,394 | | Probability = 0.0000 |

Significant results were obtained for all three analyses. Generally, precision scores for controlled vocabulary approaches in UM-D's Blue System were significantly greater than precision scores for free-text approaches in UM-D's Blue System.

Figure 9.4 shows precision scores for UM-D's Pinstripe System.

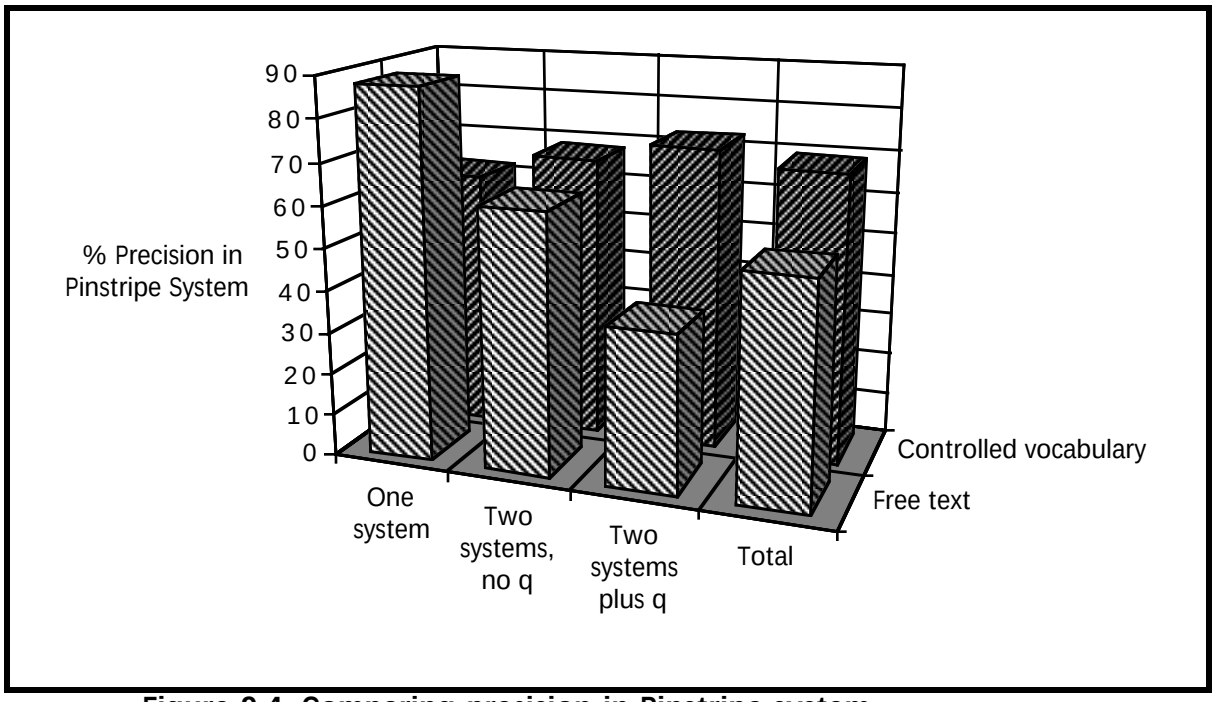


**Figure 9.4. Comparing precision in Pinstripe-system controlled vocabulary and free-text searches (UM-D)**

In UM-D's Pinstripe System, precision scores for controlled vocabulary approaches averaged 68.8% and precision scores for free-text approaches averaged 52.8%. Except for precision scores for search administrations in one system, precision in controlled vocabulary searches exceeded precision in free-text approaches.

We compared precision scores to determine if the difference between the scores for controlled vocabulary and free-text approaches was significant. The results of t-tests comparing precision scores in UM-D's Pinstripe System are given in Table 9.4.

**Table 9.4. T-test Results for Precision in the Pinstripe System (UM-D)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***T-test results for Blue System, no or suspect questionnaire*** | | | |
| Controlled vocabulary | 257 | 0.6770 | 0.4685 |
| Free text | 396 | 0.6212 | 0.4857 |
| T = 1.4551 | Degrees of freedom = 651 | | Probability = 0.1461 |
| ***T-test results for Blue System plus questionnaire*** | | | |
| Controlled vocabulary | 265 | 0.7245 | 0.4476 |
| Free text | 689 | 0.3730 | 0.4840 |
| T = 10.2564 | Degrees of freedom = 952 | | Probability = 0.0000 |
| ***Total (T-test results for Blue System administrations)*** | | | |
| Controlled vocabulary | 522 | 0.7011 | 0.4582 |
| Free text | 1,085 | 0.4636 | 0.4989 |
| T = 9.1752 | Degrees of freedom = 1,605 | | Probability = 0.0000 |

Significant results were obtained for two of the three analyses. Results were not significant for search administrations in which users failed to complete post-search questionnaires or gave suspect responses; however, precision scores for controlled vocabulary approaches were higher than such scores for free-text approaches. Generally, precision scores for controlled vocabulary approaches in UM-D's Pinstripe System were significantly greater than precision scores for free-text approaches in UM-D's Pinstripe System.

Figure 9.5 shows precision scores for Earlham's Blue System.

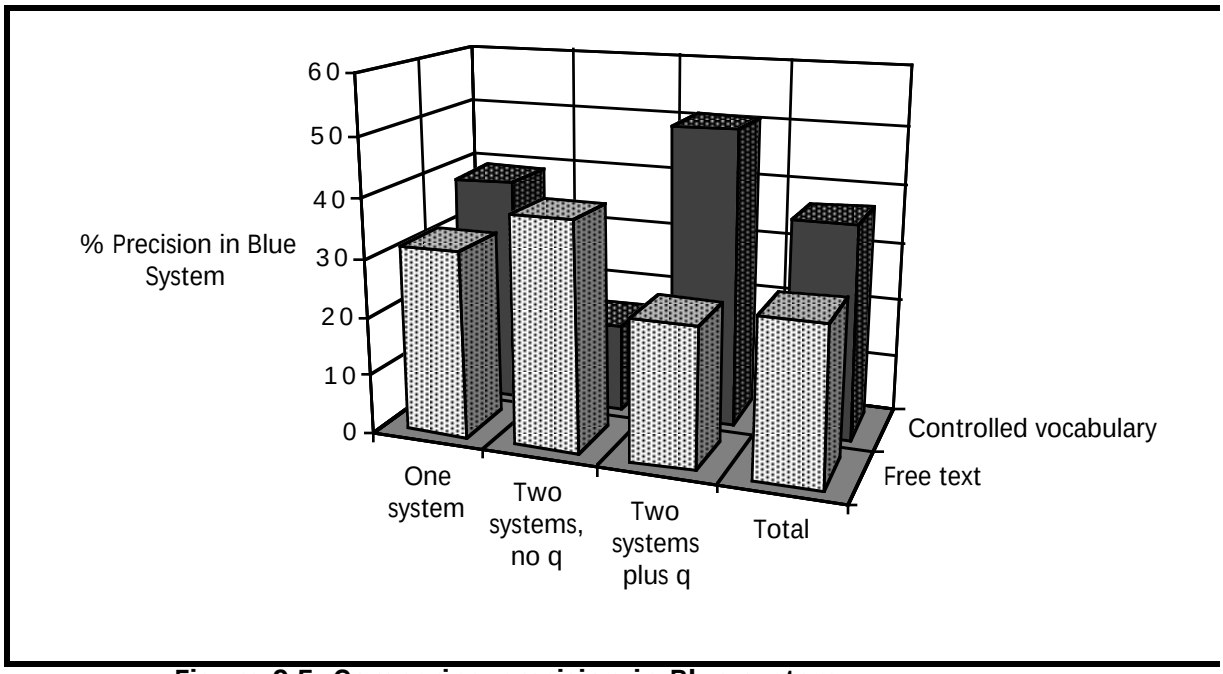


**Figure 9.5. Comparing precision in Blue-system controlled vocabulary and free-text searches (Earlham)**

In Earlham's Blue System, precision scores for controlled vocabulary approaches averaged 36.5% and precision scores for free-text approaches averaged 26.6%. Except for precision scores for search administrations in which users failed to complete post-search questionnaires or gave suspect responses, precision in controlled vocabulary searches exceeded precision in free-text approaches.

We compared precision scores to determine if the difference between the scores for controlled vocabulary and free-text approaches was significant. The results of t-tests comparing precision scores in Earlham's Blue System are given in Table 9.5. Significant results were obtained for only one analysis — for full search administrations.

**Table 9.5. T-test Results for Precision in the Blue System (Earlham)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***T-test results for Blue System, no or suspect questionnaire*** | | | |
| Controlled vocabulary | 72 | 0.1528 | 0.3623 |
| Free text | 13 | 0.3846 | 0.5064 |
| T = -1.9907 | Degrees of freedom = 83 | | Probability = 0.0498 |
| ***T-test results for Blue System plus questionnaire*** | | | |
| Controlled vocabulary | 103 | 0.5049 | 0.5024 |
| Free text | 122 | 0.2377 | 0.4274 |
| T = 10.6458 | Degrees of freedom = 223 | | Probability = 0.0000 |
| ***Total (T-test results for Blue System administrations)*** | | | |
| Controlled vocabulary | 175 | 0.3600 | 0.4814 |
| Free text | 135 | 0.2519 | 0.4357 |
| T = 2.0433 | Degrees of freedom = 308 | | Probability = 0.0419 |

Users conducted few searches in Earlham's Pinstripe System. Overall, precision for controlled vocabulary searches was 61.9% and precision for free-text searches was 45.4%. Table 9.6 gives the results of a single t-test for precision scores in search administrations in which users completed post-search questionnaires or gave suspect responses.

**Table 9.6. T-test Results for Precision in the Pinstripe System (Earlham)**

| Administration type | No. of titles displayed | Av. precision | Standard deviation |
|---|---|---|---|
| ***Total (T-test results for Pinstripe System administrations)*** | | | |
| Controlled vocabulary | 42 | 0.6190 | 0.4915 |
| Free text | 240 | 0.4542 | 0.4989 |
| T = 1.9800 | Degrees of freedom = 280 | | Probability = 0.0487 |

Results of the analysis were promising but not significant. For the most part, precision scores for controlled vocabulary approaches in Earlham's Pinstripe System were greater than precision scores for free-text approaches in Earlham's Pinstripe System.

With few exceptions, controlled vocabulary searches resulted in greater precision than free-text searches. This result supported the notion of search trees because the Blue System's search trees favored controlled vocabulary searches over free-text searches.

## 9.3Estimated Recall Scores

Estimated recall scores were based on relevance assessments users gave to titles the last time they saw them in Blue- and Pinstripe-system searches. Recall scores for individual searches were calculated by dividing the number of titles retrieved and displayed and given "very useful" ratings in a search by the total number of titles retrieved and displayed and given "very useful" ratings" in the two searches the user conducted for his topic in the Blue and Pinstripe Systems.

Let's use the same example as we used in section 9.3.1 in a discussion of calculating precision to explain how recall was calculated. The user searched the Blue System first and displayed ten titles. He gave displayed titles five "not useful" ratings, three "possibly useful" ratings, and two "very useful" ratings. The user continued searching in the Pinstripe System and displayed six titles. She gave displayed titles two "not useful" ratings, two "possibly useful" ratings, and two "very useful" ratings. Two of the sixteen titles displayed were displayed more than once. One of the ten titles displayed in the Blue System was the same as one of the titles displayed in the Pinstripe System. The user rated this title "possibly useful" in the Blue System and "very useful" in the Pinstripe System. Since the user displayed this title last in the Pinstripe search, the relevance assessment for this title in the Blue System was changed to "very useful." Of the six titles displayed in the Pinstripe System, one title was displayed twice. The first time the user displayed this title, she gave it a "not useful" rating; the second time the user displayed this title, she gave it a "possibly useful" rating. The relevance assessment for the first time this title was displayed was disregarded. Thus, this title retained the second relevance assessment for "possibly useful."

After conflicting relevance assessments were resolved, the search in the Blue System resulted in five titles deemed "not useful," two titles deemed "possibly useful," and three titles deemed "very useful;" the search in the Pinstripe System resulted in one title deemed "not useful," two titles deemed "possibly useful," and two titles deemed "very useful." The number of titles given "very useful" ratings was four because one of the titles deemed "very useful" was displayed in both systems. Thus, recall in the Blue System was 75% (3 "very useful" titles displayed in the Blue System divided by the total of 4 titles deemed "very useful" in searches of the two experimental systems) and recall in the Pinstripe System was 50% (2 "very useful" titles displayed in the Pinstripe System divided by the total of 4 titles deemed "very useful" in searches of the two experimental systems).

Recall scores reported in this study were estimates of this measure. To figure recall scores that were not estimates, the researchers would have to perform a comprehensive review of every title in the two experimental online catalogs for every search topic and

enlist subject experts for relevance assessments (Lancaster and Warner 1993, 181). Such a review was not within the purview of this study. The method of calculating recall employed in this study has been suggested as a way of estimating recall (Lancaster and Warner 1993, 181–3).

Figure 9.6 gives recall scores for experimental system users at UM-D who conducted searches in both Blue and Pinstripe Systems. This figure does not include scores of users who conducted a search of one of the two experimental systems because there was no second search for the same topic from which to extract relevance ratings for the calculation of estimated recall scores.

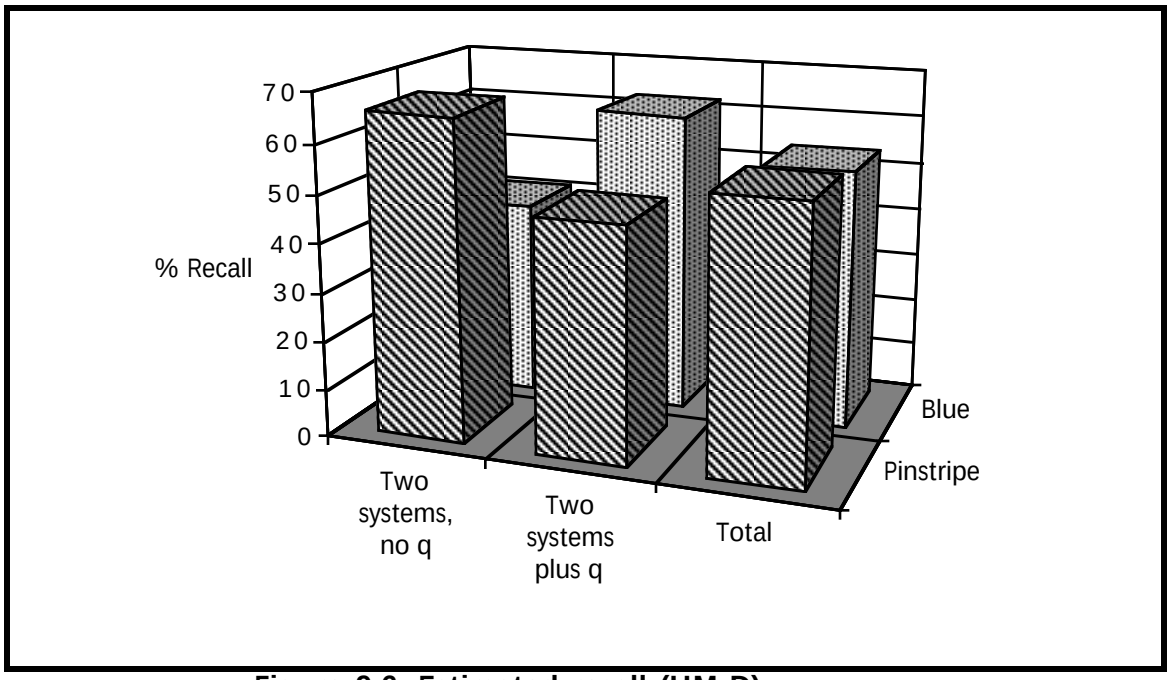


**Figure 9.6. Estimated recall (UM-D)**

Overall, recall scores in UM-D's experimental online catalog were about the same; recall in the Blue System was 53.4% and recall in the Pinstripe System was 55.4%. Recall scores in the two search administrations types flip flopped.

Figure 9.7 gives recall scores for experimental system users at Earlham who conducted searches in both Blue and Pinstripe Systems. This figure does not include scores of users who conducted a search of one of the two experimental systems because there was no second search for the same topic from which to extract relevance ratings for the calculation of estimated recall scores.

Overall, recall in Earlham's Pinstripe System (61.8%) was somewhat higher than recall in the Blue System (50.8%). Recall scores in the two search administration types

flip flopped. Recall scores (74.5%) in the Pinstripe System for search administrations in which users did not complete post-search questionnaires or gave suspect responses were considerably higher than such scores in the Blue System (29.1%); however, recall scores for this search administration type were based on only eighteen search administrations.

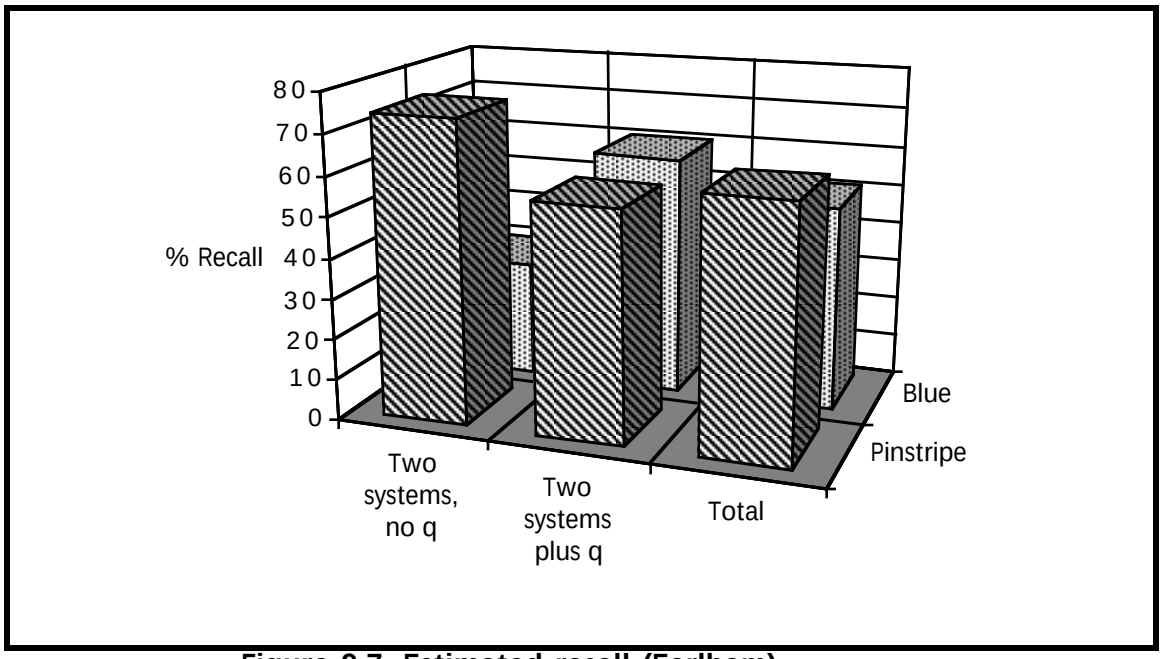


**Figure 9.7. Estimated recall (Earlham)**

Estimated recall scores varied greatly depending on the type of search administration. In the most complete search administrations (when users completed both searches in the Blue and Pinstripe Systems and the post-search questionnaire), recall scores were higher in the Blue System than in the Pinstripe System.

We did not perform t-tests on recall scores because t-tests, by definition, are used to determine whether the means of two *independent* samples differ significantly (Roscoe 1975, 224). In online retrieval tests, data were gathered and manipulated in such a fashion that the two samples were not independent.

The analysis of rated titles enlisted a relative recall calculation that required adding the total number of relevant documents found in one system to the total number of relevant documents found in a second system. For this experiment, the relative recall was calculated by dividing the number of titles rated “very useful” in the Blue or Pinstripe System by the total number of titles rated “very useful” in both searches for the same topic in the Blue and Pinstripe Systems.

System users performed searches and made relevance assessments on the titles they displayed. Ideally, users performed relevance assessments of titles retrieved and

displayed in both Blue and Pinstripe Systems. Users occasionally rated titles retrieved in both Blue and Pinstripe Systems. Because titles could be displayed and rated in both systems, the samples were not independent. Thus, a t-test of two independent samples was not a valid statistical test for these data.

## 9.4Post-search Questionnaire Responses

### 9.4.1 System Performance

Of the eleven questions in the post-search questionnaire, three questions specifically asked end users about the performance of the two experimental systems in finding useful titles for their topics of interest. These questions asked users to compare the two experimental systems in terms of the number of useful titles they retrieved, their satisfaction with the two systems, and their system preference for future searches.

Users who completed post-search questionnaires were a subset of the users who searched the experimental online catalogs and whose precision and estimated recall scores were given in several tables and figures of section 9. Users who completed post-search questionnaires searched both Blue and Pinstripe Systems to find titles on their topics of interest. In table and figure captions, precision and estimated recall scores that these users' searches achieved were designated by the caption "Search in two systems, questionnaire" or "Two systems plus q." Totals of 171 users at UM-D and 57 users at Earlham answered post-search questions.

The first post-search question asked users to compare the two systems based on the number of useful titles they retrieved. The first question and four responses categories were:

**"Comparing the Blue Test System with the Pinstripe Test System, did you find:**

- A greater number of useful titles in the Blue Test System
- About the same number of useful titles in both systems
- A smaller number of useful titles in the Blue System
- No useful titles in either system

Figure 9.8 summarizes end-user responses in search administrations at UM-D and Earlham.

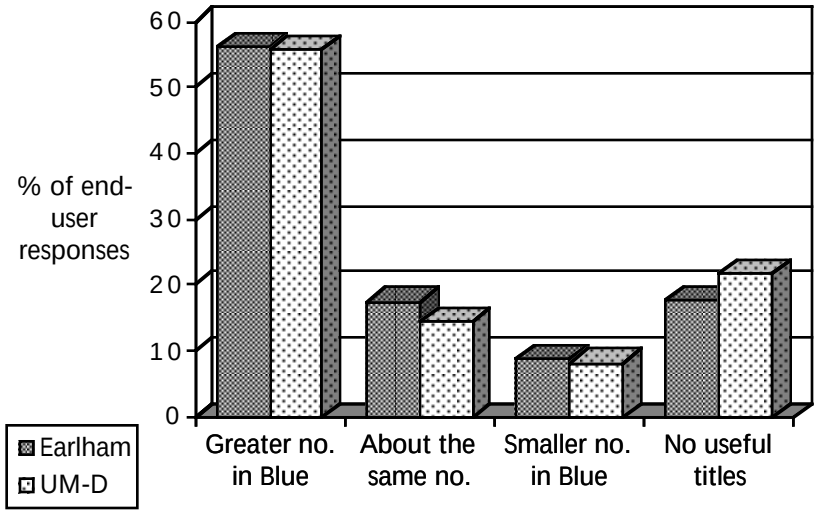


**Figure 9.8. Comparing the number of useful titles retrieved**

Almost three-quarters of users felt that the Blue System retrieved a greater number of useful titles than or about the same number of useful titles as the Pinstripe System.

The second post-search question asked users about their satisfaction with their searches. This question and three responses categories were:

**"In relation to what you were looking for, would you say your search was more satisfactory in:**

- Blue Test System
- Pinstripe Test System
- No difference

Figure 9.9 summarizes end-user responses to this question at UM-D and Earlham.

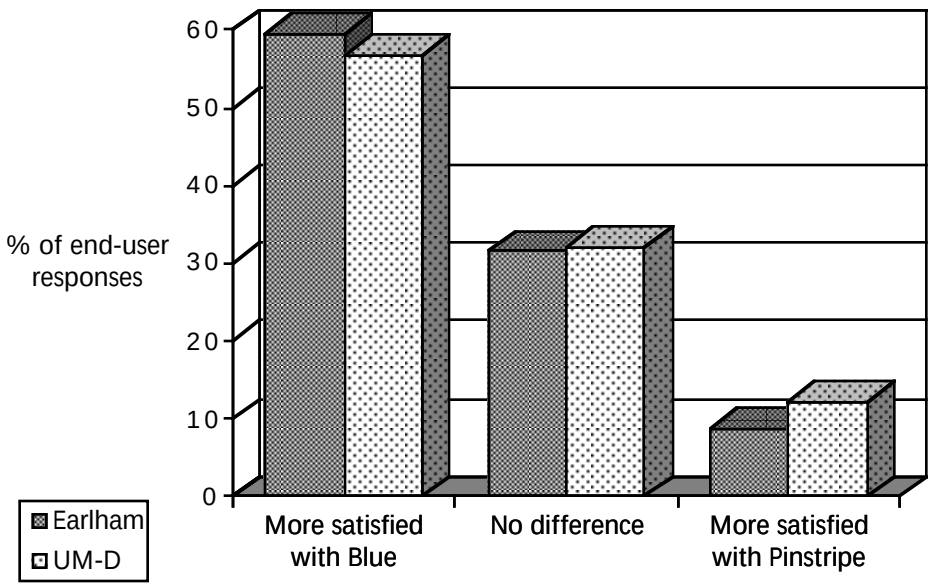


**Figure 9.9. Satisfaction with the search**

The vast majority of end users felt their search was more satisfactory in the Blue System than in the Pinstripe System. At UM-D, 56.7% of end users chose the Blue System. At Earlham, 59.6% of end users chose the Blue System. About a third of users at both institutions felt no difference in satisfaction with their search in the two experimental systems. The percentages of users who were more satisfied with their search of the Pinstripe System were quite small (12.3% of UM-D users and 8.8% of Earlham users). The low percentage of Earlham users who were more satisfied with their search in the Pinstripe System was rather surprising in view of the higher overall precision scores for searches of the Pinstripe System than for searches of the Blue System (see figure 9.2). However, other precision scores for Pinstripe-system searches were lower than such scores for Blue-system searches (see Table 9.6).

The third post-search question asked users which system they preferred. This question and three responses categories were:

**"The next time you need to conduct a computer search, will you want to use:**

- Blue Test System
- Pinstripe Test System
- No difference

Figure 9.10 summarizes end-user responses to this question on system preference.

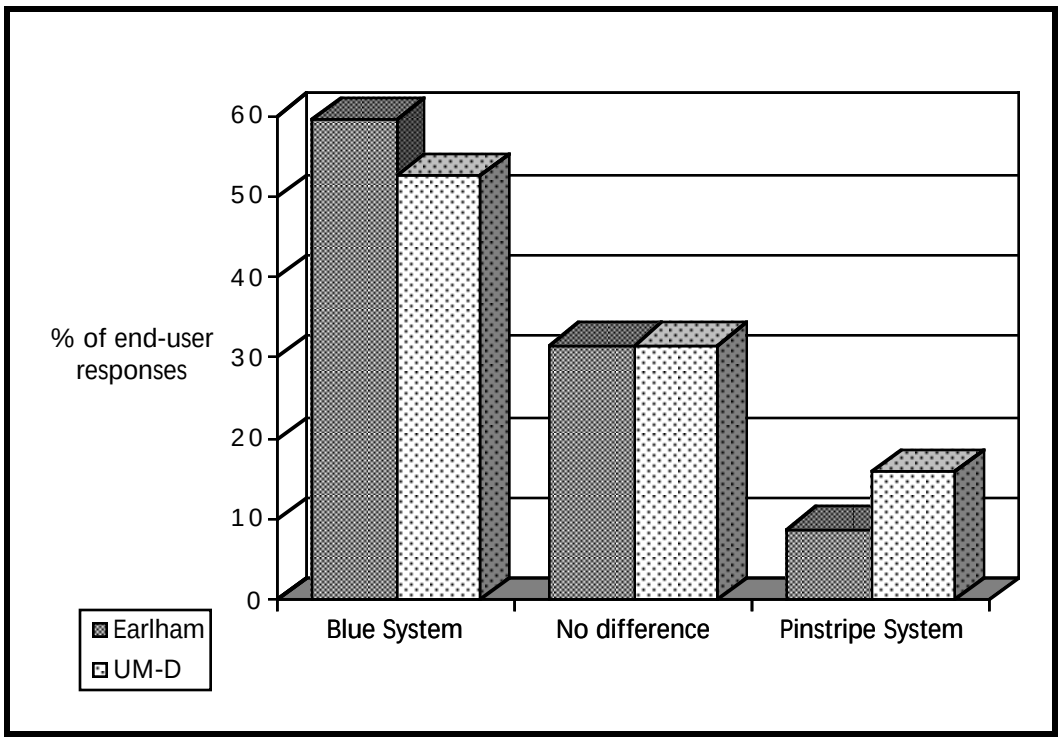


**Figure 9.10. System preference**

The majority of end users preferred the Blue System over the Pinstripe System. At UM-D, 52.6% of end users chose the Blue System. At Earlham, the percentages were unchanged from the previous question. Clearly, end users at Earlham preferred the Blue System. About a third of users at both institutions did not have a preference for one system over another.

### 9.4.2 Ease of System Use

Measures such as precision and estimated recall that information retrieval researchers have used over the years to evaluate systems were important to the end users who participated in the retrieval tests in this study because their responses to post-search questions regarding retrieving useful titles, search satisfaction, and system preference favored the Blue System which gave them high precision and recall. The post-search questionnaire probed other factors — ease of system use, system efficiency in retrieving useful titles, system helpfulness in giving users new subject searching ideas — in the comparison of Blue and Pinstripe Systems.

The next five questions on the post-search questionnaire dealt with ease of system use. The first question asked users to compare the ease of giving instructions to the two systems. The question was "Which system do you find easier to give instructions to, for example, to type in subjects, continue searches, make selections from lists?"

Figure 9.11 summarizes end-user responses to this question.

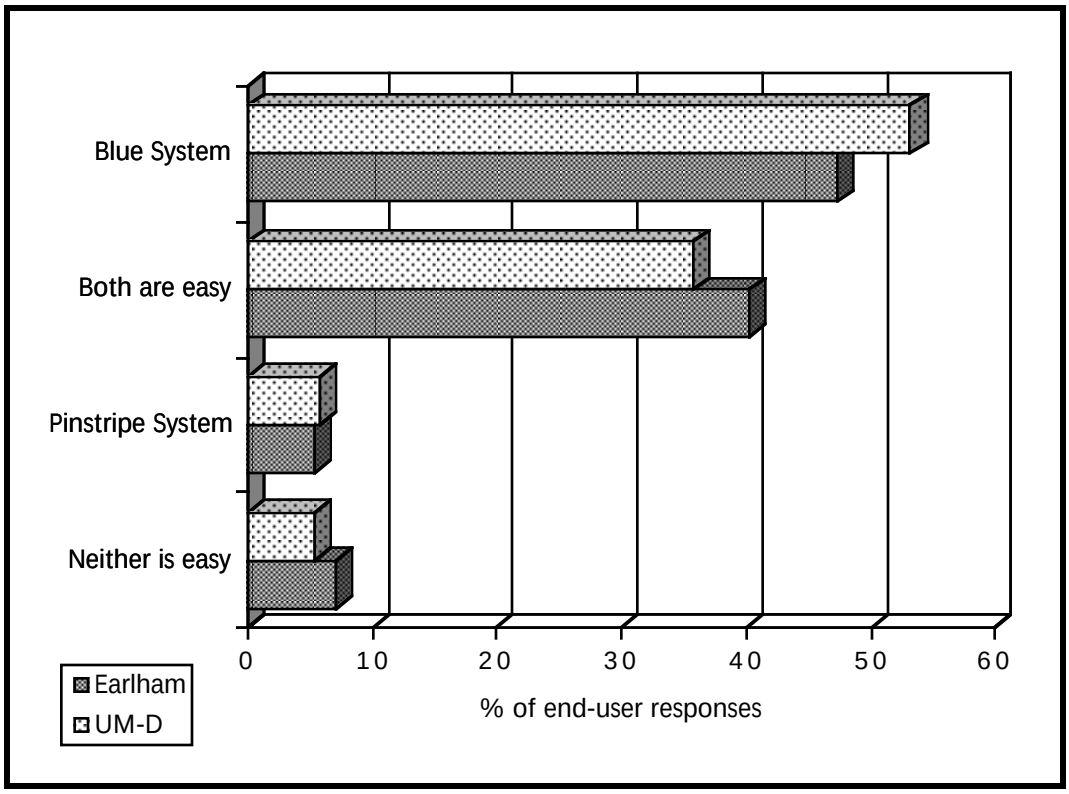


**Figure 9.11. Ease of giving instructions to the experimental systems**

Generally, users at both institutions chose the Blue System over the Pinstripe System as the easier of the two systems to which to give instructions. Large percentages of users (35.7% and 40.3% of UM-D and Earlham users, respectively) felt both systems were easy to give instructions to.

Responses to the next four questions on ease of system use fell into the same pattern, that is, large percentages of users chose the Blue System over the Pinstripe System and sizable percentages of users felt positively about both systems about the particular ease-of-use aspect under scrutiny.

Figure 9.12 summarizes user responses to a question about getting instructions from the system, viz. "Which system gives you more useful instructions in response to the subjects you type in?"

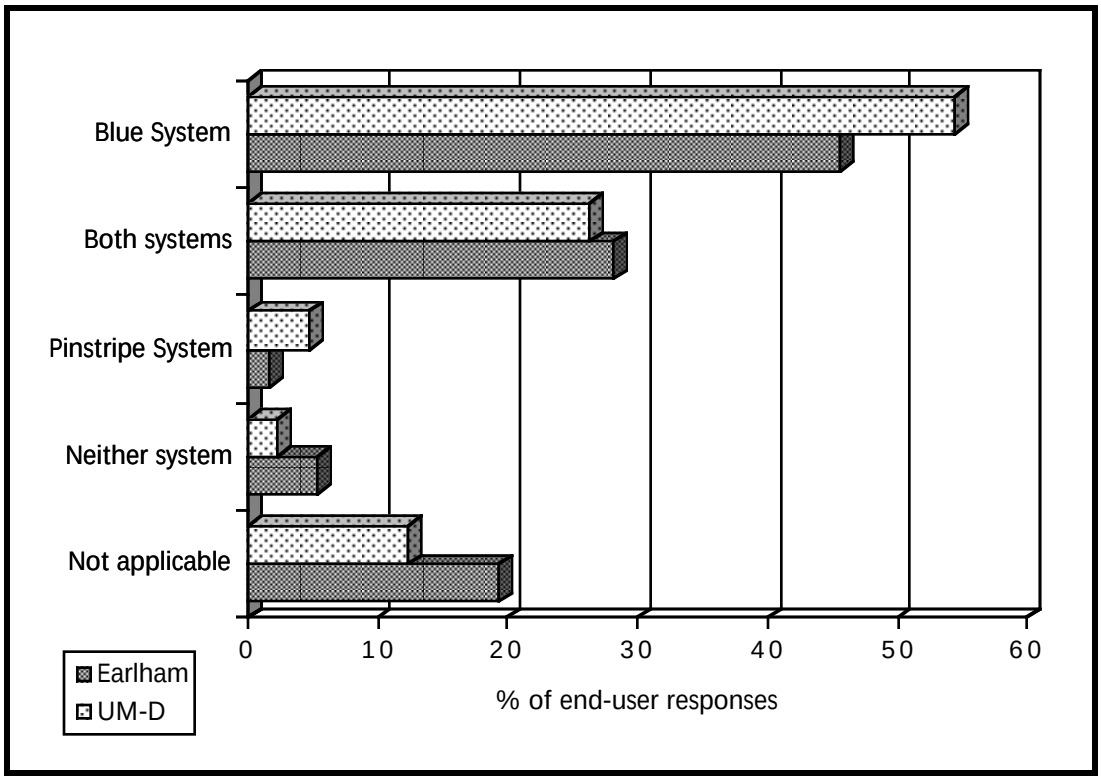


**Figure 9.12. Ease of getting instructions from the experimental systems**

Besides fitting into the response pattern described above, between 12% and 20% of users selected the "not applicable" category. Many of these users had conducted searches for subjects that failed to produce retrievals. Following a system message informing users that the system found nothing in the system, the system continued with the post-search questionnaire. Thus, the experimental systems gave users little or no opportunity to assess whether system instructions were useful in response to user queries because they retrieved no titles for user queries.

Figure 9.13 summarizes user responses to a question about the systems' clarity, viz. "Which system tells you what to do more clearly?"

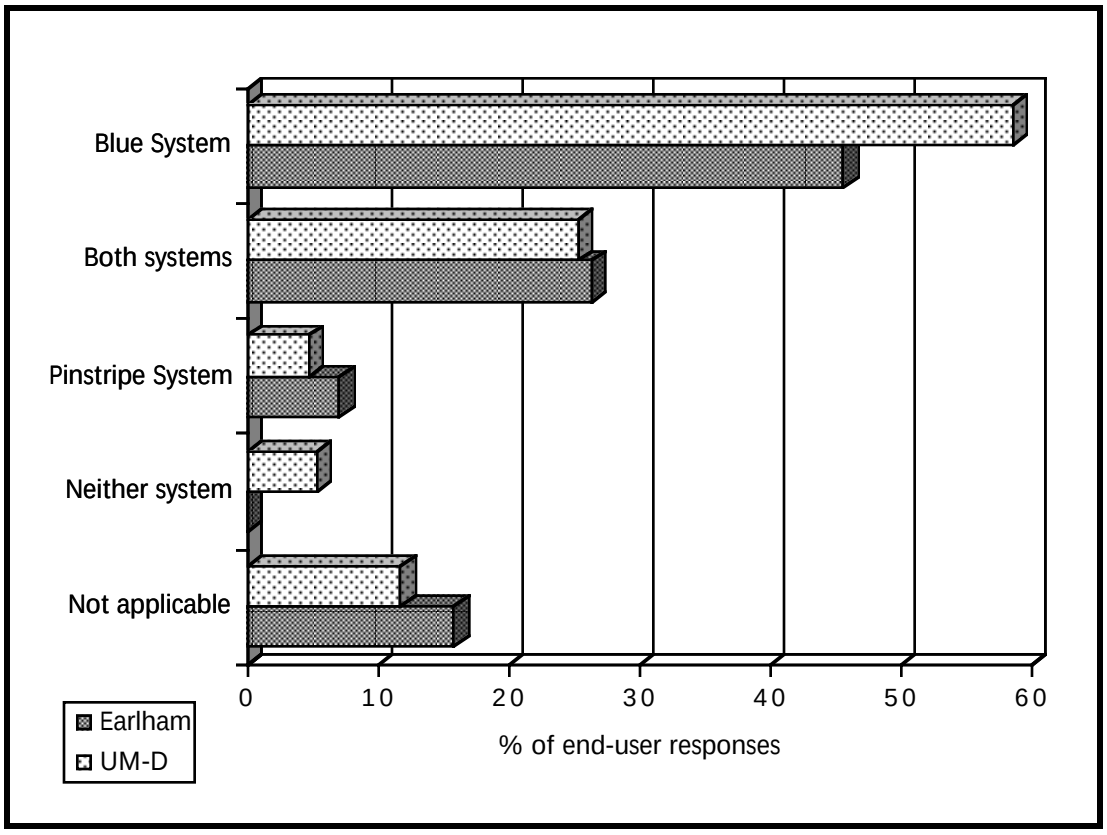


**Figure 9.13. Clarity of the experimental systems**

User responses followed the same pattern as the previous question. That is, many users chose the Blue System over the Pinstripe System in terms of clarity. About a quarter of users felt both systems were clear. Sizable percentages of users (11.7% at UM-D and 15.8% at Earlham did not feel the question was applicable to their experience using the experimental systems. The reason could have been attributed to searches in which no titles were retrieved because the systems did not give users the opportunity to assess their respective clarity.

Figure 9.14 summarizes user responses to a question about the systems' efficiency, viz. "Which system is more efficient in guiding you to useful information?"

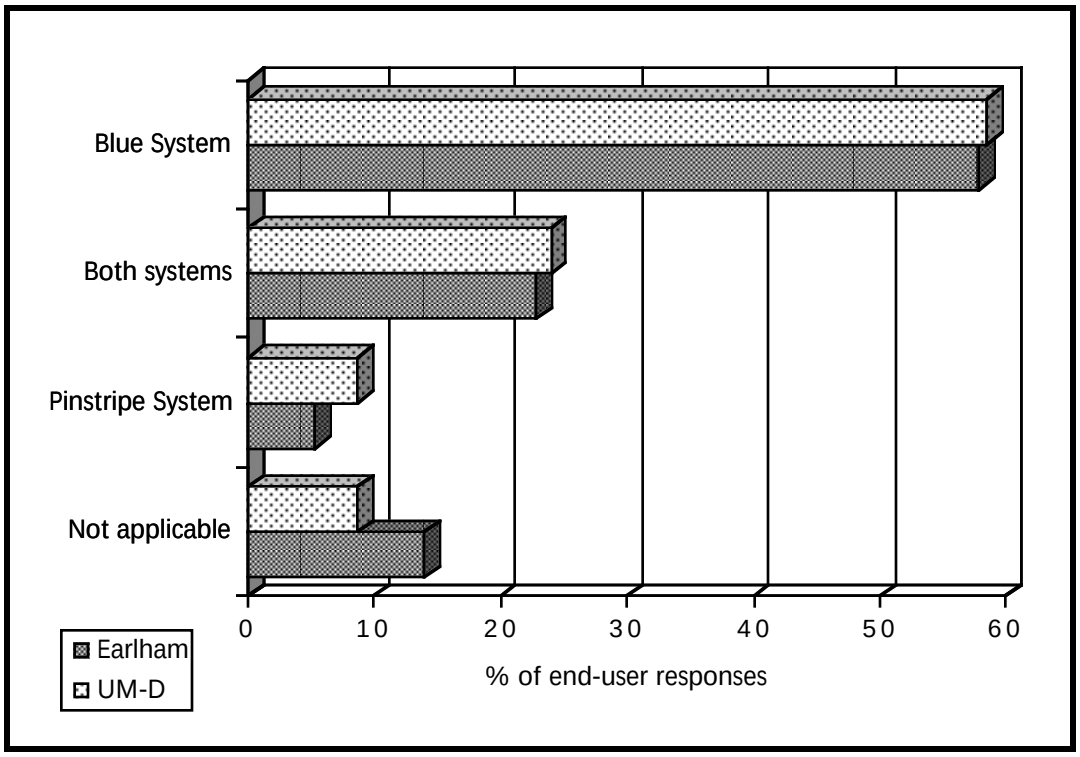


**Figure 9.14. Efficiency of the experimental systems**

The familiar pattern of favoring the Blue over the Pinstripe System was evident in user responses to the question on efficiency. About a quarter of respondents felt both systems were efficient in guiding them to useful information

Figure 9.15 summarizes user responses to a question about the systems' helpfulness in giving users new ideas for subject searching, viz. "Which system is more helpful for giving you new ideas for searching for subjects?"

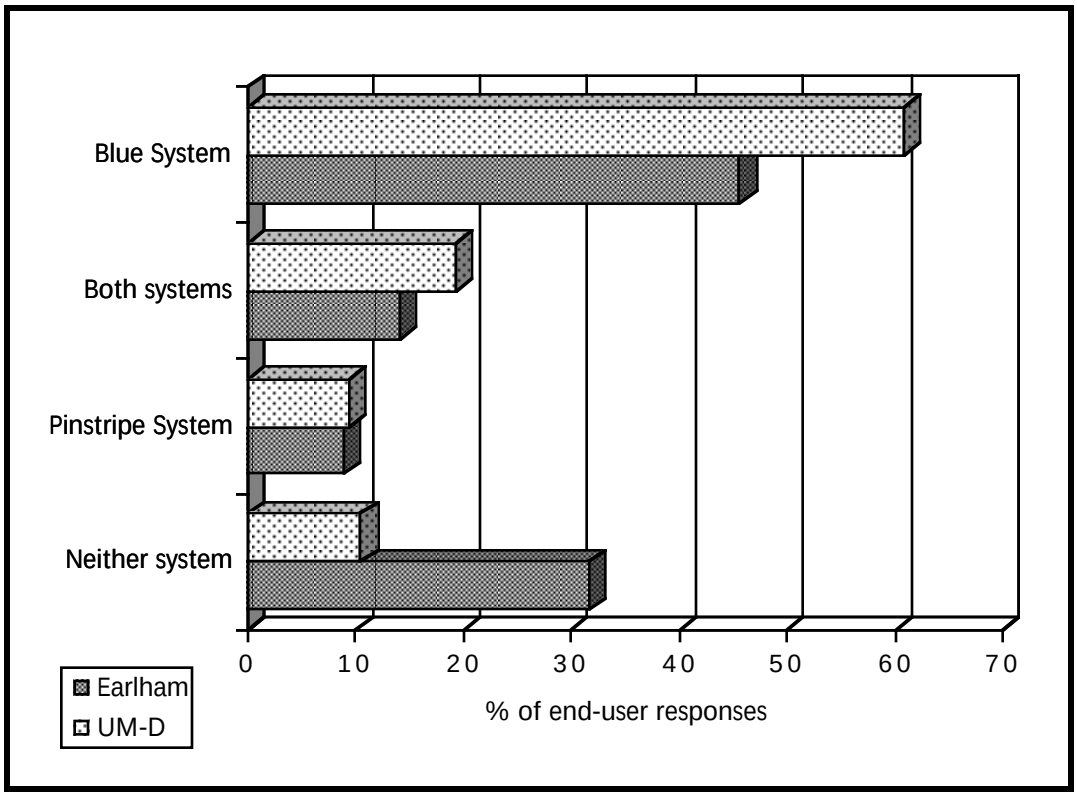


**Figure 9.15. Helpfulness of giving new ideas for subject searching**

The response pattern was a little different for this question. Respondents at UM-D (60.8%) clearly favored the Blue System over the Pinstripe System. Only about ten percent of UM-D users felt neither system was helpful in giving new ideas for subject searching. Respondents at Earlham (45.6%) were somewhat less exuberant about the helpfulness of the Blue System. Almost a third of Earlham users felt neither system was helpful in this regard. Recall the lower estimated precision score that the Blue System achieved at Earlham (figure 9.2). Although Earlham users clearly favored the Blue System overall, their response to this question indicated that both systems need improvement in capabilities that help users get new ideas for subject searching.

### 9.4.3 Importance of System Features

Following the five questions on ease of system use, the questionnaire enumerated six system features and asked users to tell us how important each feature was to them. The question was "On a scale of 1 to 5 in which 1 means 'not at all important' and 5 means 'very important,' rate the following features as to their importance in helping you

search library computer systems like ASTUTE." The system prompted users to enter a number between 1 and 5. It listed six features for users to rate. The features were:

- System's ability to continue searching on its own to find useful information
- System's efficiency in guiding you to useful information
- System's clarity in telling you what to do
- New ideas that the system gives you for searching for subjects
- Usefulness of the system's instructions
- Ease of giving instructions to the system

All but the first-listed feature above were similar to the ease-of-use aspects that had been addressed in the previous five questions.

We included this question to determine what system features were especially important to users. Figure 9.16 shows the percentage of users who rated features "important" or "very important."

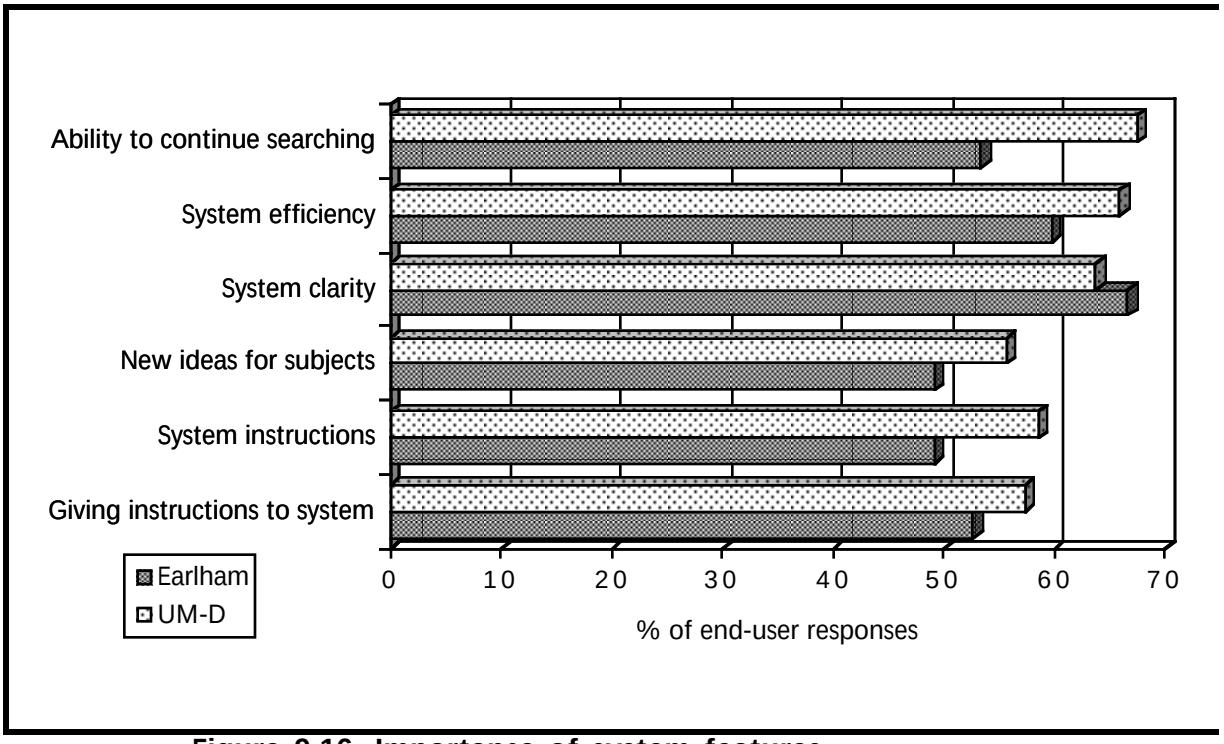


**Figure 9.16. Importance of system features**

End users at both institutions were positive about all listed features. End users at UM-D were especially keen on system efficiency and a system's ability to continue searching on

its own to find useful information. End users at Earlham were more positive about system clarity and efficiency.

### 9.4.4 Concluding Questions

The post-search questionnaire concluded with two questions. The first questioned users about their familiarity with the subject they searched. The question was "How would you rate your familiarity with the subject you searched today?" Figure 9.17 summarizes end-user responses at UM-D and Earlham.

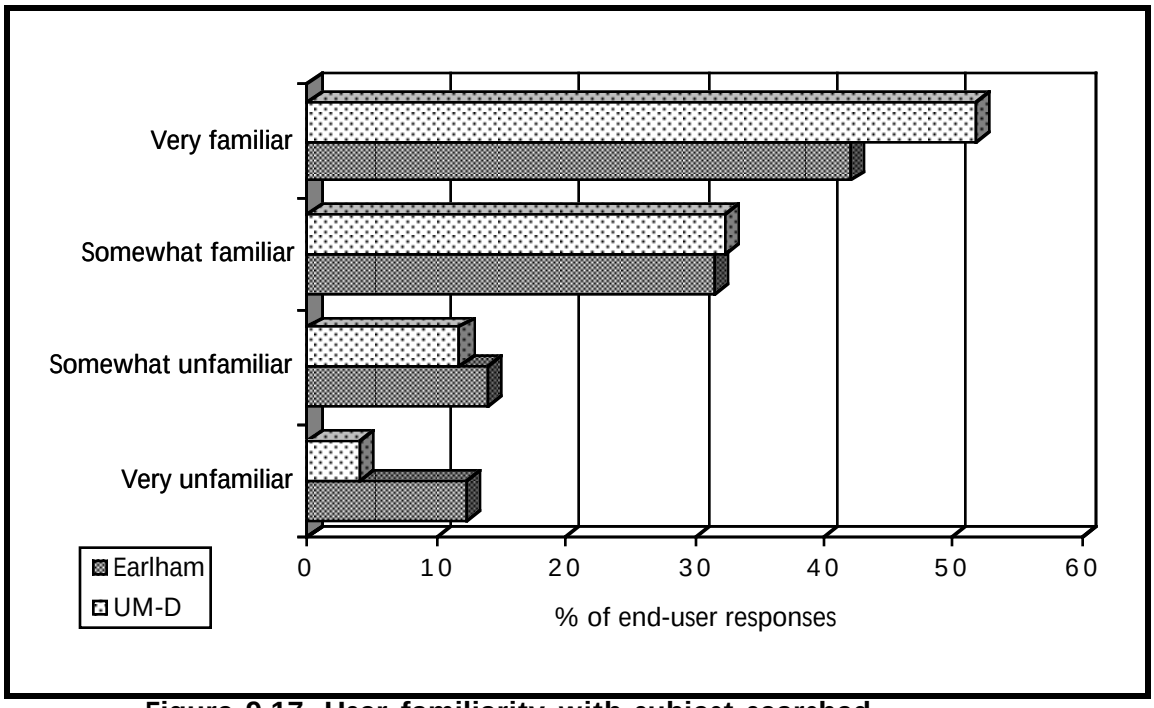


**Figure 9.17. User familiarity with subject searched**

Large percentages of users were very familiar with the subject they searched in the experimental online catalog. Although Earlham users reported a lesser degree of familiarity than UM-D users, almost three-quarters of Earlham users said they were "very familiar" or "somewhat familiar" with their topics of interest.

Only a quarter of Earlham users reported history as their major field of study (see Table 8.1). The experimental online catalog's database contained library material on American history. Thus, the lesser degree of familiarity reported by Earlham users was expected in view of the major fields of study they reported.

The final question asked users how interested they were in seeing the capabilities of the experimental online catalogs extended to topics besides Engineering and Computer

Science (at UM-D) and American history (at Earlham). Figure 9.18 gives user responses to the question.

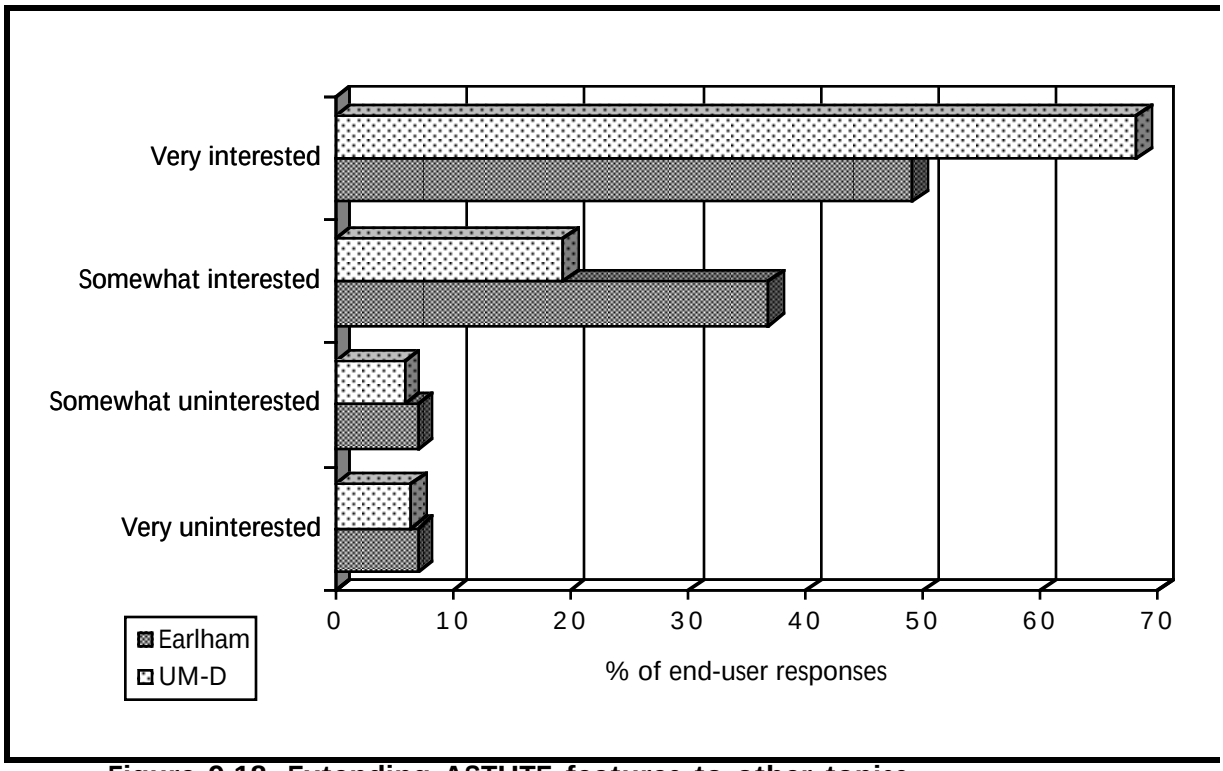


**Figure 9.18. Extending ASTUTE features to other topics**

We interpreted positive responses to this question as an indication that users were generally pleased with the capabilities of the experimental ASTUTE system. Almost seven of every eight users at UM-D were interested in seeing ASTUTE features extended to other topics. Earlham users were less enthusiastic than UM-D users but they were still quite positive. ASTUTE prompted users who selected the "very interested" response category to type in the topics to which they would ASTUTE features extended. Table 9.7 lists user-suggested topics.

**Table 9.7. Topics to Extend to ASTUTE Features**

| Topics Suggested by UM-D Users | Topics Suggested by Earlham Users |
|---|---|
| Business | Art |
| Politics | Art history |
| Arts and humanities | Philosophy |
| Physics | Sociology |
| Biology | Social sciences |
| Psychology | World War 2 |
| More than just engineering | African history |
| Natural sciences | Civil War |
| Geometry | Politics |
| History | African American poetry |
| Foreign languages | Religion |
| Law | Child abuse |
| Sociology | Supreme Court decisions |
| Flowcharting | African American women writers in Harlem Renaissance |
| Digital signal processing | Women in Reconstruction |
| Solar powered automobiles | |
| Object oriented programming | |

Many suggested topics named broad fields of study, e.g., geometry, physics, biology. Some topics described specific subjects that had limited coverage in the UM-D and Earlham databases, e.g., “object oriented programming,” “digital signal processing,” “African American women writers in Harlem Renaissance,” and “Women in Reconstruction.”

## 9.5Chapter Summary

This chapter used the traditional measures of precision and estimated recall to compare the performance of the Blue and Pinstripe Systems. Statistical results were supported by end-user responses to questions in post-search questionnaires.

At UM-D, precision scores averaged 64% in the Blue System and 58% in the Pinstripe System. In full search administrations at UM-D (i.e., administrations in which users conducted searches in both Blue and Pinstripe Systems and completed a post-search questionnaire), precision scores in the Blue System were significantly higher than precision scores in the Pinstripe System (see figure 9.1. and Table 9.1). The result was exactly the opposite at Earlham. That is, precision scores in the Pinstripe System were

significantly higher than precision scores in the Blue System (see figure 9.2. and Table 9.2).

We extended the analysis of precision scores to comparisons for controlled vocabulary and free-text searches because the Blue System's search trees favored controlled vocabulary approaches over free-text approaches in response to user queries. We expected the precision of titles retrieved through controlled vocabulary approaches to exceed the precision of titles retrieved through free-text approaches. Such results would support the design of search trees in the experimental online catalog.

In UM-D's Blue and Pinstripe Systems, precision scores for controlled vocabulary approaches exceeded such scores for free-text approaches (see figures 9.3 and 9.4). Furthermore, the differences between scores were significant (see Tables 9.3 and 9.4). Especially large differences between precision scores were registered for full search administrations. In Earlham's Blue System, precision scores for controlled vocabulary approaches exceeded such scores for free-text approaches in full search administrations (see figure 9.5). There was an insufficient number of searches for statistical comparisons of precision in Earlham's Pinstripe System; however, overall precision scores for controlled vocabulary approaches were higher than such scores for free-text approaches at Earlham (see Table 9.6). These findings about precision and search approaches support the design of search trees that favor controlled vocabulary over free-text approaches for searching bibliographic databases.

Overall, recall scores in UM-D's experimental online catalog were about the same; recall in the Blue System was 53.4% and recall in the Pinstripe System was 55.4% (see figure 9.6). Recall scores in the two search administrations flip flopped. That is, recall scores in the Blue System exceeded such scores in the Pinstripe System in full search administrations and vice versa in partial administrations. Type of search administration results were the same at Earlham — recall scores in the two search administrations flip flopped (see figure 9.7). That is, recall scores in the Blue System exceeded such scores in the Pinstripe System in full search administrations and vice versa in partial administrations. Overall, recall in Earlham's Pinstripe System (61.8%) was somewhat higher than recall in the Blue System (50.8%). No statistical tests were performed on recall scores because the design of the comparison search experiment violated the t-test's principle of independence.

End-user responses to the post-search questionnaire were available for full search administrations. Generally, responses reflected statistical results of this type of search administration. That is, users rated the performance of the Blue System higher than the Pinstripe System in terms of retrieving greater numbers of useful titles (see figure

9.8), search satisfaction (see figure 9.9), and they preferred the Blue System over the Pinstripe System (see 9.10).

The post-search questionnaire probed other factors — ease of system use, system efficiency in retrieving useful titles, system helpfulness in giving users new subject searching ideas — in the comparison of Blue and Pinstripe Systems. The majority of end users rated the Blue System or both systems highly in terms of the ease of giving instructions to the systems (see figure 9.11), ease of getting instructions from the system (see figure 9.12), system clarity (see figure 9.13), and system efficiency in guiding them to useful information (see figure 9.14).

At Earlham, users were a little less enthusiastic about the Blue System's helpfulness in giving new ideas for subject searching (see figure 9.15). Since the design of the Blue System was intended to give users new subject searching ideas through exact-approach browsing and expanding opportunities, we were somewhat concerned about this response. The failure analysis of experimental catalog searches examines searches closely to determine the reasons for failure.

## References

Lancaster, F. Wilfred, and Amy J. Warner. 1993. *Information retrieval today*. Arlington, Va.: Information Resources Press.

Roscoe, John T. 1975. *Fundamental research statistics for the behavioral sciences.* New York: Holt, Rinehart, and Winston.

# 10 Failure Analysis of User Searches

## 10.1 Introduction

Recall and precision are measures that information retrieval researchers have used for many years to evaluate searches (Lancaster and Warner 1993). While these measures disclose important information about the number of useful titles searches produce, they do not reveal why systems are successful or unsuccessful in retrieving useful titles.

This and subsequent chapters feature a failure analysis of end-user searches to determine why the experimental ASTUTE online catalog was successful or unsuccessful in retrieving useful titles. The project director, an expert searcher of information retrieval systems, repeated end-user searches that were successful and unsuccessful in retrieving useful titles to identify the reasons why they succeeded or failed. This section features the results of the failure analysis.

## 10.2 Successful and Unsuccessful Searches at UM-Dearborn

The first step in the failure analysis of searches performed at the University of Michigan-Dearborn and Earlham College was to characterize searches that users performed in the Blue and/or Pinstripe Systems as successful or unsuccessful:

- Successful search — users rated one or more titles “very useful” in a search performed in the Blue or in Pinstripe System.
- Unsuccessful search — users did not rate one or more titles “very useful” in a search performed in the Blue or in Pinstripe System.
- Unsuccessful pairs of searches — users did not rate titles “very useful” in searches performed in the Blue *and* Pinstripe Systems for the *same* topic.

Table 10.1 enumerates the number and percentage of successful and unsuccessful searches in the Blue and Pinstripe Systems at UM-D.

**Table 10.1. Successful and Unsuccessful Searches at UM-D**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| 1 useful titles retrieved | 156 | 44.2 | 150 | 42.1 |
| No useful titles retrieved in one of two systems | 69 | 19.5 | 78 | 21.9 |
| No useful titles retrieved in searches of both systems | 128 | 36.3 | 128 | 36.0 |
| Total | 353 | 100.0 | 356 | 100.0 |

Percentages of successful and unsuccessful searches were about the same for the Blue and Pinstripe Systems. One or more useful titles were retrieved in 44.2% of searches in the Blue System; one or more useful titles were retrieved in 42.1% of searches in the Pinstripe System. The Blue System had a lower percentage (19.5%) of searches in which no useful titles were retrieved than in the Pinstripe System (21.9%) but the percentages were not much different from one another. Generally, the Blue and Pinstripe Systems performed about the same in terms of retrieving useful titles in searches.

Subsections of section 10.2 focus on unsuccessful searches in the Blue and Pinstripe Systems at UM-D.

## 10.2.1 Failure of Both Systems

At UM-D, users performed a total of 297 pairs of searches. That is, they performed one search for their topic of interest in the Blue Systems followed by a search for the same topic in the Pinstripe System. Of the 297 pairs, 128 search pairs failed to yield useful titles in searches of both experimental systems — about 43% of the search pairs. The project director, an expert searcher of information retrieval systems, repeated these 128 searches in both Blue and Pinstripe Systems to determine why end users were unsuccessful in retrieving useful titles.

### 10.2.1.1 User Perseverance

User perseverance was a common reason for the failure of searches to retrieve useful titles. For example, a user entered the subject query “fiber optics” into the Pinstripe System. The keyword-in-record search responded by displaying the first of fifteen retrieved titles to the user. The user gave this first title a “not useful” rating and terminated the search. Examples of titles this user missed by failing to continue the display of retrieved titles were *Fiber optics engineering, Fundamentals of optical fiber communication, Fiber optics,* and *Optical fiber systems: technology, design, and*

*applications.* This and many other searches like it should not be counted as unsuccessful searches. The catalog produced titles on the subject query that the user entered. The user chose to terminate his search rather than display additional titles.

Perseverance was also the reason why the user's search of the Blue System for the same topic failed to produce useful titles. In response to the query "fiber optics," the Blue System made an exact match of a subject heading and responded with exact-search options for displaying general works, browsing broader terms, displaying form subdivisions, and expanding the search using a different subject searching approach. Instead of pursuing one or more of the exact-search options, the user terminated the search.

Table 10.2 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to user perseverance.

**Table 10.2. User Perseverance**

| User query | Explanation |
|---|---|
| ceramics | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, notes, form subdivisions, and for expanding the search are enabled. The user ends the search. |
| probability | The Blue System's title-keyword search displays the first of 24 retrieved titles. The user gives it a negative relevance assessment and ends the search. By failing to display additional titles, the user misses potentially useful titles such as *Probability and statistics for engineers and scientists, Probability and statistics for engineers,* and *Probability and statistics for modern engineering.* |
| operations research | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, subject and form subdivisions, and for expanding the search are enabled. The user ends the search. Had she selected the general works option, she would have retrieved titles such as *Operations research: An introduction, Introduction to operations research,* and *Applied operations research and management science.* |
| chemical engineering | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, subject and form subdivisions, and for expanding the search are enabled. The user selects the general works option. The system responds by displaying the first of 32 retrieved titles. The user rates this title "not useful" and ends the search. |

If users had continued searching for the searches described in Table 10.2, these searches would have most likely succeeded in retrieving useful titles.

The researchers did not conduct personal interviews with end users following their search(es). Had we conducted such interviews, we would have had explanations from users as to their reasons for terminating their searches. In the absence of interview data, we can speculate that the results of alphabetical or exact searches could have been too complex for users to understand. That is, they could have been lost or confused about which of the several options to pursue for continuing the search. This does not explain what was in the mind of the user who conducted the search for "chemical engineering." The Blue System retrieved a manageable number of titles (32) and the user failed to display more than the first title.

In most controlled vocabulary searches that failed due to perseverance, users failed to make actions (e.g., select an exact-search option, select a subject heading from an alphabetical list) that would enable them to retrieve and display titles. In several free-text searches, users failed to display more than the first of several retrieved titles.

Users conducted two-thirds of the 39 searches that failed due to perseverance in the Blue System. In several searches, users made a few actions to find useful titles. Two examples of such searches in the Blue and Pinstripe Systems follow. One user entered the query "c programming language" into the Pinstripe System which responded with the alphabetical approach and placed the user in an alphabetical list bearing subject headings such as "C (Computer program language)," "C Basic (Computer program language)," and "C++ (Computer program language)." The user selected the former, gave two of 42 retrieved titles negative relevance assessments, and ended the search.

Another user entered the query "mathematics" into the Blue System which made an exact match of a subject heading and responded with the exact approach. The user browsed narrower terms, selected the narrower term "Logic, Symbolic and mathematical," displayed three of nine retrieved titles, and gave them negative relevance assessments. The user had several options for continuing the search. He could have browsed titles assigned to subdivided forms of the selected narrower term; he could have returned to the exact-approach main menu for the original matching subject heading "Mathematics" and browsed titles assigned to this or subdivided forms of this subject heading. Instead, the user terminated the search without browsing or exploring terminology or titles connected to mathematics.

The failure analysis of searches that failed due to user perseverance demonstrates that these searches were really successful searches in terms of leading users to titles on their topics of interest. Despite the several opportunities the systems gave to users for browsing terminology or displaying titles, users terminated their searches.

#### 10.2.1.2 Specificity of User Queries

On occasion, users entered straightforward queries into the experimental online catalog, retrieved seemingly useful titles, but rated them "possibly useful" or "not useful." We have used the rubric "Specificity of user queries" to describe these queries. We think that users might have had a different (and probably related) topic in mind but, for whatever reason, they entered a query into the experimental online catalog that was not indicative of their information needs. The project director encountered comparable queries in a previous research project (Markey and Demeyer 1986, 157–8).

Table 10.3 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to specificity of user queries.

**Table 10.3. Specificity of User Queries**

| User query | Explanation |
|---|---|
| computer user interface | Blue System retrieves one title in a keyword-in-record search. The user gives the title (*Designing the user interface: strategies for effective human-computer interaction*) a "not useful" rating. |
| composite materials | Pinstripe System retrieves one subject heading, i.e., "Composite materials," in a keyword-in-subdivided-heading search. User selects the heading, displays all 13 retrieved titles, and gives them negative relevance assessments. Examples of these titles are *Mechanics and composite materials, Composite materials,* and *Composite materials and their use in structures.* |
| material handling | Blue System informs the user that the entered term is a cross reference to "Materials handling." Exact-approach options for displaying general works, broader terms, narrower terms, subject and form subdivisions, and for expanding the search are enabled. The user displays several titles (e.g., *Integrated materials handling in manufacturing, The principles and practice of materials handling*) under subdivided forms of this subject heading and gives them negative relevance assessments. |

Had we conducted post-search interviews with users, we would have had explanations from users regarding their reasons for giving negative relevance assessments to retrieved titles. In the absence of interview data, we can speculate that the topic that users had in mind was not quite the same as the subjects of the queries they entered into the experimental online catalog.

A total of seventeen searches failed because of query specificity. Twelve searches were conducted in the Blue System and five searches were conducted in the Pinstripe System. The failure analysis demonstrates that these searches were really successful searches in terms of leading users to titles on the topics they expressed in the queries they entered into the catalogs. In view of user relevance assessments, the searches were not successful in terms of retrieving titles that fulfilled the specifications of queries users had in their minds but did not make explicit in their queries.

#### 10.2.1.3 Conflicting Relevance Assessments

Section 8.5.5 described how we resolved conflicting relevance assessments. Basically, we used the last relevance assessment that searchers gave to titles displayed more than once and disregarded other relevance assessments for the same title.

Conflicting relevance assessments could have affected a search in a single system. For example, a user could have given two different relevance assessments to the same title in a search of the Blue System or to the same title in a search of the Pinstripe System. Conflicting relevance assessments could have also affected searches of both experimental systems for the same topic. That is, a user could have given two different relevance assessments to the same title in separate searches of the Blue and Pinstripe Systems.

The researchers attributed the failure of eight searches (five in the Blue System and three in the Pinstripe System) to conflicting relevance assessments. An example of a search in which we attributed failure to conflicting relevance assessments was a search of the Blue System for "attenuation equalizers." The Blue System informed the user that this term was a cross reference to the subject heading "Equalizers (Electronics)" and responded with several exact-search options. The user displayed the title *Passive equalizer design data* three times, rated it "very useful" the first and second times, and "not useful" the third time.

#### 10.2.1.4 Database Failure

The failure of searches was attributed to database failure in the following circumstances: (1) users retrieved titles but displayed titles were much more specific than the original user query, (2) users retrieved too few titles on topics that were covered in the selected classification areas, and (3) users retrieved no titles on topics that were covered in the selected classification areas. The researchers were guided in their decision about attributing failure to database failure on the titles users displayed in their searches and the negative relevance assessments they gave to titles.

Table 10.4 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to database failure.

**Table 10.4. Database Failure**

| User query | Explanation |
|---|---|
| workstation | Blue System retrieves one title in a title-keyword search, *Workstation environment for wastewater treatment design using artificial intelligence and mathematical models,* which the user rates "not useful." In Pinstripe, the keyword-in-subdivided-heading search retrieves no headings. |
| virtual reality | Both Blue and Pinstripe Systems fail to retrieve titles. Blue System enlists subject searching approaches for queries composed of more than one word. Pinstripe System tries to retrieve titles through the keyword-in-record search. |
| design propeller | Pinstripe System invokes the alphabetical approach which places the user in an alphabetical list of subject headings at the subject headings "Design, Industrial" and "Desktop publishing." Blue System enlists subject searching approaches for queries composed of more than one word and fails to make a match. |

Examples of other queries that failed to retrieve useful titles in both Blue and Pinstripe Systems due to database failure were:

- linked lists
- oriented polymers
- time motion studies
- multiple integration
- modem programming
- wide area networks
- joint application development

The failure of 120 searches was attributed to database failure. Of these 120 searches, there were 59 pairs of searches in which the failure of searches in both the Blue and Pinstripe Systems was attributed to database failure. The failure of another two searches was attributed to database failure in the Pinstripe System. The failure of Blue-system searches for the same topics was attributed to the exact approach because users seemed to have difficulty navigating exact-search options in the Blue System. An example was a Blue-system search in which the user was interested in historic preservation and began the search with the query "architecture." The Pinstripe System responded with 16 subject headings retrieved in a keyword-in-subdivided-heading search. The user displayed titles assigned to "Landscape architecture" and "Architecture — Conservation and restoration" and gave them negative relevance

assessments. In the search of the Blue System for the same topic, the user seemed to tire of selecting the several exact-approach options that led to the display of the subject heading "Architecture — Michigan — Detroit — Conservation and restoration" because she terminated the search before displaying titles assigned to this subdivided subject heading. Section 10.2.1.8 focuses on navigational problems in experimental-system searches.

#### 10.2.1.5 Large Retrievals

We did not expect large retrievals to be a problem in this study because the databases of the experimental online catalog were quite small. Yet, a few topics were represented by large numbers of bibliographic records, e.g., computers, architecture, various computer languages, mathematics.

Table 10.5 details one Blue-system search and three Pinstripe-system searches that were plagued by too many retrievals.

**Table 10.5. Large Retrievals**

| User query | Explanation |
|---|---|
| fortran | Blue System responds with the results of an exact search in which options for displaying general works, narrower terms, form subdivisions, and for expanding the search are enabled. The user selects the general works option. The system responds by displaying the first of 110 retrieved titles. The user walks away from the system without even rating the first title. |
| computer | Pinstripe System responds by displaying the first of 2,324 titles retrieved in a keyword-in-record search. User displays three titles and rates them "not useful." |
| materials | Pinstripe System responds by displaying the first of 815 titles retrieved in a keyword-in-record search. User rates this title "not useful" and ends the search. |
| operations research | Pinstripe System retrieves 144 subject headings in a keyword-in-subdivided-heading search. The user selects one subject heading, displays one title, gives it a "not useful" rating, and ends the search. |

Of the 14 searches that failed due to the retrieval of too many titles, 13 searches were conducted in the Pinstripe System. Most of the time, large retrievals in the Pinstripe System were connected with the results of keyword-in-record searches; keyword-in-subdivided-heading searches occasionally produced too many titles.

As library catalogs have increased in size in recent years, several researchers have focused on the problem of too many retrievals (Markey and Demeyer 1986, 277;

Wiberley and Daugherty 1988; Lynch 1990). In a recent study, Wiberley, Daugherty, and Danowski (1993) concluded that few users displayed titles when the number of retrieved titles exceeded 200. They suggested that system designers provide features that help users cope with retrievals that exceed this number (Wiberley, Daugherty, and Danowski 1993, 5). In view of their research, we considered searches which resulted in large retrievals (titles or subject headings) as unsuccessful searches because the particular subject searching approach invoked by the experimental system did not provide features for managing the display of such large retrievals.

#### 10.2.1.6 Vocabulary of User Queries

The experimental online catalog contained information on the topics that interested several users. However, users phrased their queries in certain ways that prevented the experimental online catalogs from making retrievals in response to their queries. When users entered queries with such phrasing, the two experimental systems responded in different ways. The Blue System exhausted the subject-searching approaches on the search trees before responding to users that it was unable to find titles to satisfy their queries. The Pinstripe System performed the search using one of the three subject-searching approaches at its disposal before responding to users that it was unable to find titles to satisfy their queries. However, differences between the two systems in terms of subject-searching approaches were transparent to end users.

An example of a search characterized with an unfortunate phrasing of the query was for "mechanical statics." The Blue System exhausted the subject-searching approaches on the search trees and informed the user that it was unable to find titles to satisfy the query. The Pinstripe System submitted this query to a keyword-in-subdivided-heading search and made no retrievals. If the systems had automatically truncated the word "mechanical," they would have retrieved titles like *Engineering mechanics statics, Schaum's outline of theory and probabilities of engineering: mechanics, statics, and dynamics,* and *Engineering mechanics: statics and dynamics.*

Table 10.6 describes four of the 31 queries that failed due to the vocabulary of user queries. In the left column of Table 10.6 are listed queries that failed due to vocabulary problems. Under the right-hand column are suggested ways in which systems could process queries to retrieve titles and subject headings on users' topics of interest. Fourteen queries had vocabulary problems that adversely affected both searches in Blue and Pinstripe systems. Vocabulary problems adversely affected another three queries in the Blue System only. (The search for the same topic in the Pinstripe System failed for different reasons.)

**Table 10.6. Vocabulary Problems**

| User query | Explanation |
|---|---|
| flow through porous media | Blue System exhausts the search trees without retrieving titles or subject headings. Pinstripe System's keyword-in-record approach fails to produce retrievals. If this query were reduced to "porous media," the experimental system would retrieve *Fundamentals of transport phenomena in porous media* in a title-keyword search. |
| serial communication | Blue System exhausts the search trees without retrieving titles or subject headings. Pinstripe System's keyword-in-record approach fails to produce retrievals. If automatic truncation were applied to this query, the experimental system would retrieve *Mastering serial communications* in a title-keyword search. |
| factor of safety | Blue System exhausts the search trees without retrieving titles or subject headings. Pinstripe System's keyword-in-record approach fails to produce retrievals. If this query were reduced to "safety," the experimental system would respond with the alphabetical approach and display subject headings such as "Safety engineering," "Safety, Industrial," "Safety of systems," and "Safety of workers." |
| c.a.s.e. | Blue System retrieves seven titles bearing long contents notes which contain various initials for author first and middle names including the letters and punctuation "c.," "a.," "s.," and "e." Pinstripe System's keyword-in-subdivided-heading approach fails to produce retrievals. The experimental system should delete punctuation in user queries. If single letters result, the system should concatenate the letters to form an acronym. If two or more series of characters result, the system should treat the query as a phrase and submit it to the search tree for queries exceeding one word. |

If we could redesign the ASTUTE experimental online catalog, we would add system capabilities such as automatic stemming and/or best match to improve retrieval for the queries in this failure analysis category. We would also remove punctuation in user queries to effect matches of acronyms, subject headings, or title words. (Section 12.3 focuses on searches with vocabulary problems and makes suggestions for system improvements.)

#### 10.2.1.7 Subject Searching Approach Failures

The experimental online catalog contained information on topics that interested users. However, its failure to retrieve the desired information was occasionally attributed to the particular subject searching approach selected by the system. For example, in response to the query "chemical engineering," the Pinstripe System performed a

keyword-in-subdivided-heading search and failed retrieve subject headings. (The user's search in the Blue System for this query failed for a different reason, namely, perseverance.)

Nineteen subject searching approach failures occurred in Pinstripe-system searches. The Pinstripe System failed to produce potentially useful retrievals through the subject searching approach that it chose randomly. (The Blue-system searches for the same topics failed for different reasons, e.g., perseverance, conflicting relevance assessments, vocabulary problems.)

Table 10.7 lists user queries and gives explanations for subject searching approach failures in Pinstripe-system searches.

**Table 10.7. Subject Searching Approach Failures in the Pinstripe System**

| User query | Explanation |
|---|---|
| control system | Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (A vocabulary problem adversely affects the Blue-system search for this query.) |
| bar code | Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (Search approach also adversely affects the Blue-system search for this query.) Had the systems responded with the alphabetical approach, they would have placed the user close to the cross reference "Bar coding of products." |
| attenuation equalizers | Pinstripe System responds to this query with the keyword-in-record approach and is unsuccessful retrieving titles. (Conflicting relevance assessments adversely affect the Blue-system search for this query that leads to the retrieval of titles through the cross reference "Attenuation equalizers" for the subject heading "Equalizers (Electronics).") |
| material handling | Pinstripe System responds to this query with the keyword-in-record approach and retrieves two titles that the user rates "possibly useful." (A navigational problem adversely affects the Blue-system search for this query which features the exact approach due to the cross reference "Material handling" for the subject heading "Materials handling.") |

| | |
|---|---|
| cad | Pinstripe System responds to this query with the keyword-in-subdivided-heading approach and is unsuccessful retrieving subdivided headings. (Perseverance adversely affects the Blue-system search for this query in which the alphabetical approach displays an alphabetical list of subject headings bearing the cross reference "CAD" for the subject heading "Computer-aided design." Instead of selecting the reference, the user ends the search.) |

Of the 19 search approach failures in the Pinstripe System, all but two searches failed to produce retrievals through the keyword-in-subdivided-heading approach. The two remaining failures involved the keyword-in-record approach and were described in Table 10.7.

Search approach failures were not as common in the Blue System as in the Pinstripe System. Table 10.8 lists user queries and gives explanations for the three subject searching approach failures in Blue-system searches.

**Table 10.8. Subject Searching Approach Failures in the Blue System**

| User query | Explanation |
|---|---|
| dataflow | Blue System exhausts the search trees and fails to make a match. (Perseverance adversely affects the Pinstripe-system search for this query in which the user needs to browse forward 2 screens to find cross references such as "Data flow across national boundaries" and "Data flow, Transborder" that describe the user's topic of interest.) |
| flow charting | Blue System exhausts the search trees and fails to make a match. (Pinstripe System responds to this query with the keyword-in-subdivided-heading approach and is unsuccessful retrieving subdivided headings.) Had the systems responded with the alphabetical approach, they would have placed the user close to the subject heading "Flow charts." |
| bar code | Blue System exhausts the search trees and fails to make a match. (Pinstripe System responds to this query with the keyword-in-subdivided-heading approach and is unsuccessful retrieving subdivided headings.) Had the systems responded with the alphabetical approach, they would have placed the user close to the cross reference "Bar coding of products" for the subject heading "Product coding." |

In all three search approach failures, the Blue System exhausted the search trees without making a match that would have invoked a particular subject searching approach. Yet all three failures could have been successful searches had the Blue System responded with the alphabetical approach. Searchers would have still have to

browse listed cross references or subject headings and select them to retrieve titles on their topics of interest.

In searches in which systems exhaust search trees and fail to make retrievals, it is possible users would find a pertinent search term in lists of subject headings in the alphabetical neighborhood of the user-entered query. Thus, the alphabetical approach could be the recommended system response when systems exhaust search trees without finding a match.

#### 10.2.1.8 Navigational Problems

Subject searching approaches that featured matches of controlled vocabulary terms (e.g., exact, alphabetical, keyword-in-main-heading) offered users many options for furthering their search. On occasion, users pursued several potentially fruitful options but, despite their searching, they were unable to find useful titles. We retraced their actions in an effort to increase our understanding of what they experienced during their search. In the course of recreating their searches, we realized that users might have experienced difficulties navigating the experimental online catalog. In addition to navigational problems, users could have become frustrated with how long the system took to search selected subject headings, broader or narrower terms, or subdivided forms of selected subject headings.

An example of a search plagued by navigational problems and user frustration at system response time was a search for historic preservation. The user initiated this search with the query “architecture.” The Blue System responded with the results of an exact search in which options for displaying general works, broader terms, narrower terms, subject, place, and form subdivisions, and for expanding the search were enabled. The user browsed narrower terms and selected “Historic buildings.” It took the Blue System about 30 seconds to search for the subject heading “Historic buildings,” determine which exact-search options were enabled, and create the exact-search main menu (see figure 10.1).

The user selected the place-subdivision option and choose the subdivision “Michigan.” It took the Blue System another 30 seconds to search for “Michigan” under the main subject heading “Historic buildings.” Since there were subdivided forms of the subject heading “Historic buildings— Michigan,” the system also had to determine which exact-search options were enabled and create the exact-search main menu for the subject heading “Historic buildings— Michigan” (see figure 10.2).

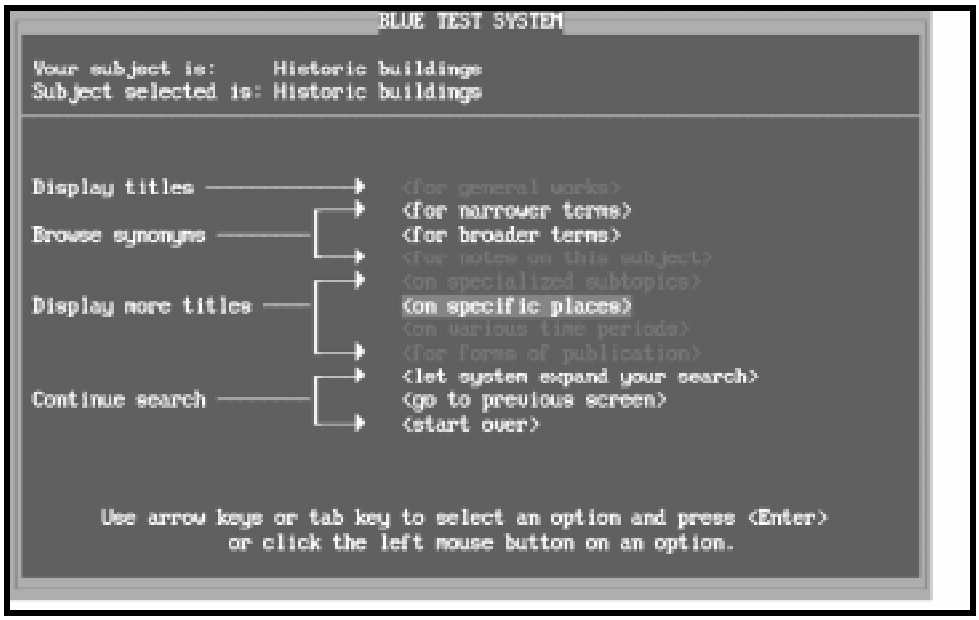


**Figure 10.1. Exact-search options for "Historic buildings"**

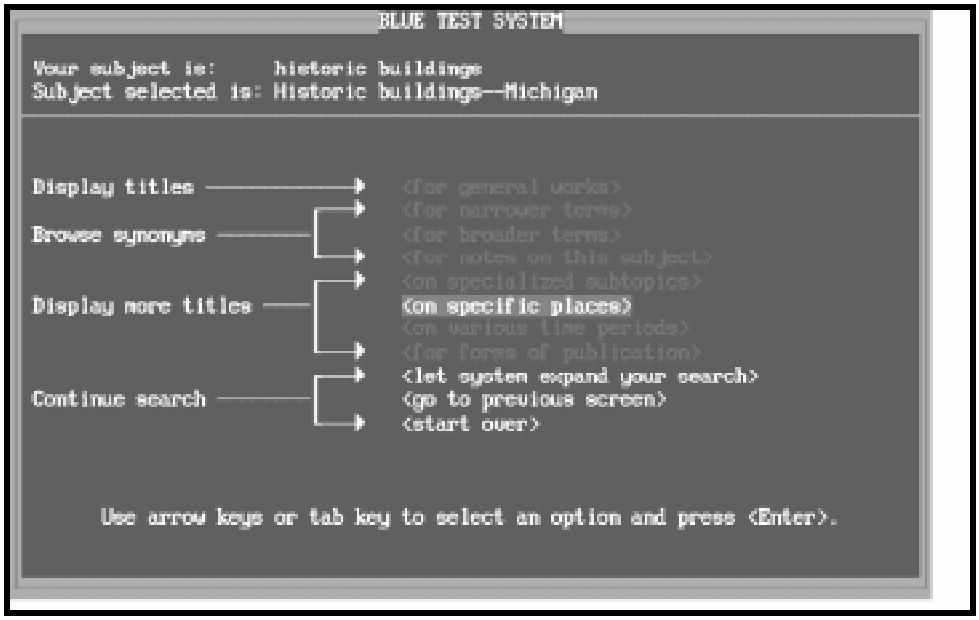


**Figure 10.2. Exact-search options for "Historic buildings— Michigan"**

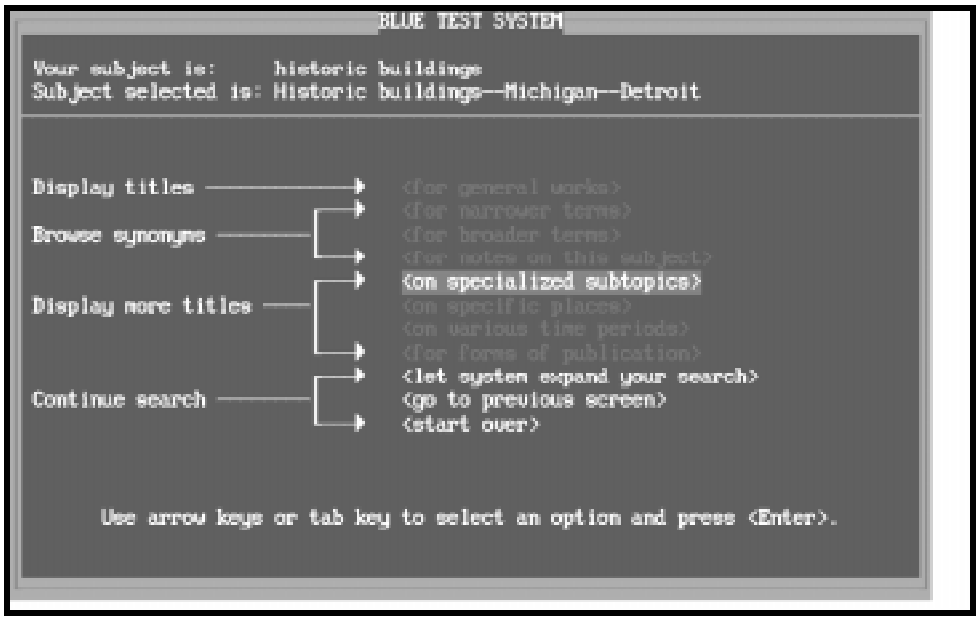


**Figure 10.3. Exact-search options for "Historic buildings— Michigan — Detroit"**

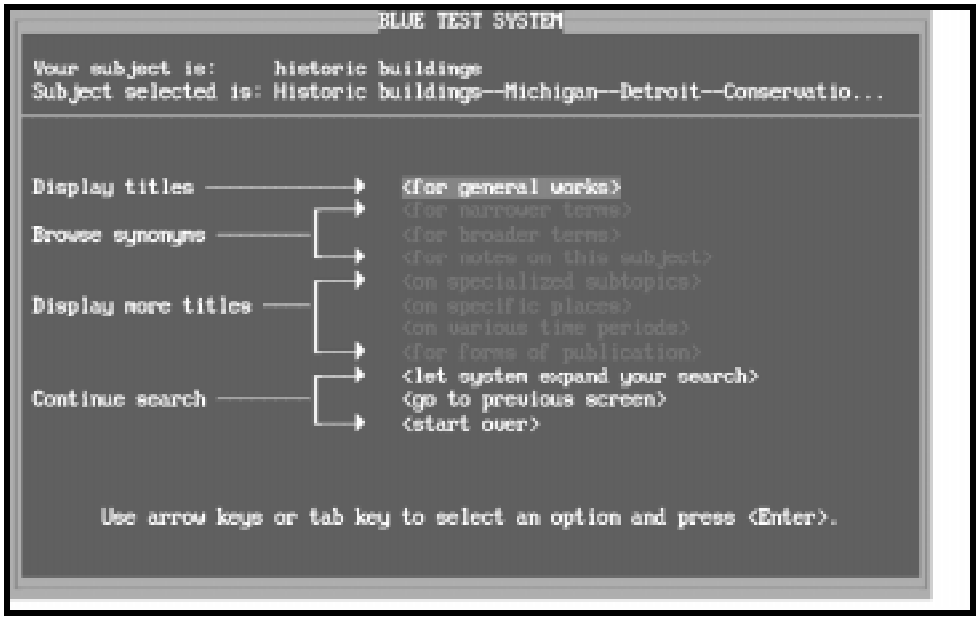


**Figure 10.4. Exact-search options for "Historic buildings— Michigan — Detroit — Conservation and Restoration"**

The Blue System responded once again with the results of an exact search in which options for displaying place subdivisions and for expanding the search were enabled.

The user selected the place-subdivision option again and chose the subdivision "Detroit." It took the Blue System another 30 seconds to search for "Detroit" under the subdivided subject heading "Historic buildings— Michigan." The Blue System responded once again with the results of an exact search in which options for displaying subject subdivisions and for expanding the search were enabled (see figure 10.3).

The user selected the subject-subdivision option and chose the subdivision "Conservation and restoration." It took the Blue System another 30 seconds to search for "Conservation and restoration" under the subdivided subject heading "Historic buildings— Michigan — Conservation and restoration." The Blue System responded once again with the results of an exact search in which options for displaying general works and for expanding the search were enabled (see figure 10.4). At long last, the system presented the user with the opportunity to display a title. For some reason, the user did not select the general-works option and terminated the search.

In the absence of user interview data, we speculated on the reasons why the user terminated the search. Perhaps she was frustrated by the length of time it took to select an exact-search option and obtain a useful response. Perhaps she lost her bearings in this lengthy, complex search and preferred to start the search over again.

In Table 10.9, two more user queries are described. The failure of these two searches was attributed to navigational problems.

The failure of five Blue-system searches was attributed to navigational problems. Three of the five searches were described in this section. Scenarios of the two remaining failed searches were quite similar to the descriptions of the two searches given in Table 10.9.

The Blue System's exact search played a major role in the failure of these five searches. Several improvements are needed to reduce user navigational problems in operational systems. First, the response time for searching selected terms and creating exact-search option menus should be reduced. Second, exact-search option menus should be used to summarize large retrievals only. If the selection of an exact-search option in the original main menu retrieves a few subject headings or titles, the system should go directly to the titles display rather than inserting exact-search option menus or subject heading lists as intermediary steps. Third, when the system matches a cross reference, it should check to determine whether the "expand search" option produces additional matches and disable the option if it does not produce such matches.

**Table 10.9. Navigational Problems**

| User query | Explanation |
|---|---|
| material handling | Blue System informs user that the entered term is a cross reference for "Materials handling." Blue System responds with enabled exact-search options for displaying general works, narrower terms, broader terms, subject and form subdivisions, and expanding the search. User chooses form-subdivisions option. System responds with enabled exact-search options for displaying general works and expanding the search. Instead of displaying general works, the user expands the search. The system exhausts the search trees and does not find another match. (The initial match was made on the cross reference "Material handling" for the subject heading "Materials handling.") |
| audio amplifier | Blue System informs user that the entered term is a cross reference for "Amplifiers, Audio." Blue System responds with enabled exact-search options for displaying subject subdivisions and expanding the search. User expands the search. Blue System responds with alphabetical list of subject headings. User chooses a cross reference for "Equalizers (Electronics)." User expands again and system performs a keyword-in-main-heading search. The result is the subject heading "Amplifiers, Audio" which the user selects. Despite the availability of a general-works option in the exact-approach main menu, the user does not display titles connected with this subject heading. |

The ASTUTE experimental online catalog is the first system in which the exact approach has been implemented. This approach was far more successful in terms of delivering useful titles to users (see section 11.2) than it was unsuccessful. However, the failure analysis in this chapter provides an excellent opportunity to study the performance of this approach, identify problems, and suggest improvements.

#### 10.2.1.9 Failure of Both Systems Revisited

Pairs of searches that failed to retrieve useful retrievals in both systems numbered 128 — a little over two-fifths of the pairs of searches performed in the experimental online catalog at UM-D. A failure analysis of these searches demonstrated the following problems that adversely affected the retrieval of potentially useful information from the experimental online catalogs: (1) user perseverance, (2) specificity that was not represented in user queries, (3) conflicting relevance assessments, (4) database failures, (5) large retrievals, (6) user query vocabulary, (7) search approach failures, and (8) navigational problems.

Problems with user perseverance, specificity of user queries, and conflicting relevance assessments had little to do with the ability of systems to retrieve useful information. Sections 10.2.1.1 to 10.2.1.3 gave several examples of searches that would have retrieved potentially useful titles but for user actions that prevented retrieval. In this failure analysis, we considered such searches successful.

Database failures are bound to occur in searches of any database. And users are likely to search databases for information on topics that the database does not cover or covers only marginally. Database failures are a type of system failure. Additional research is needed to determine the characteristics of queries that are likely to fall victim to database failures so that systems can redirect such queries to database(s) that contain the desired material. Other failures connected with systems were due to large retrievals, user query vocabulary, subject searching approach failures, and navigational problems. Throughout the subsections that described searches that failed for these reasons, we offered suggestions for making improvements to systems to reduce failure in future experimental and operational systems.

Table 10.10 reclassifies unsuccessful searches in the Blue and Pinstripe Searches. The reclassification is limited to the 128 pairs of unsuccessful searches conducted in the Blue and Pinstripe Systems and featured in subsections of section 10.2.1.

**Table 10.10. Reclassified Searches (Both systems, UM-D)**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| Successful searches† | 43 | 33.6 | 21 | 16.4 |
| Database failures | 56 | 43.8 | 58 | 45.3 |
| System failures* | 29 | 22.6 | 49 | 38.3 |
| Total | 128 | 100.0 | 128 | 100.0 |

†Barring perseverance, query specificity, conflicting relevance assessments.

*Including problems with large retrievals, subject searching approaches, user query vocabulary, navigation.

Over forty percent of unsuccessful searches in the Blue and Pinstripe Systems were attributed to database failure. A larger percentage (38.3%) of system failures was connected with the Pinstripe System than with the Blue System (22.6%). Many system failures were search approach failures in the Pinstripe System. The failure analysis promoted some unsuccessful searches to successful searches. One-third of Blue-system searches and one-sixth of Pinstripe-system searches were successful in terms of guiding users to titles on the topics their queries described. Many of the searches promoted to successful status were due to perseverance in the Blue System.

Figure 10.5 is a graphical depiction of the statistics in Table 10.10. It shows obvious differences in the percentages of successful searches and system failures in Blue and Pinstripe Systems.

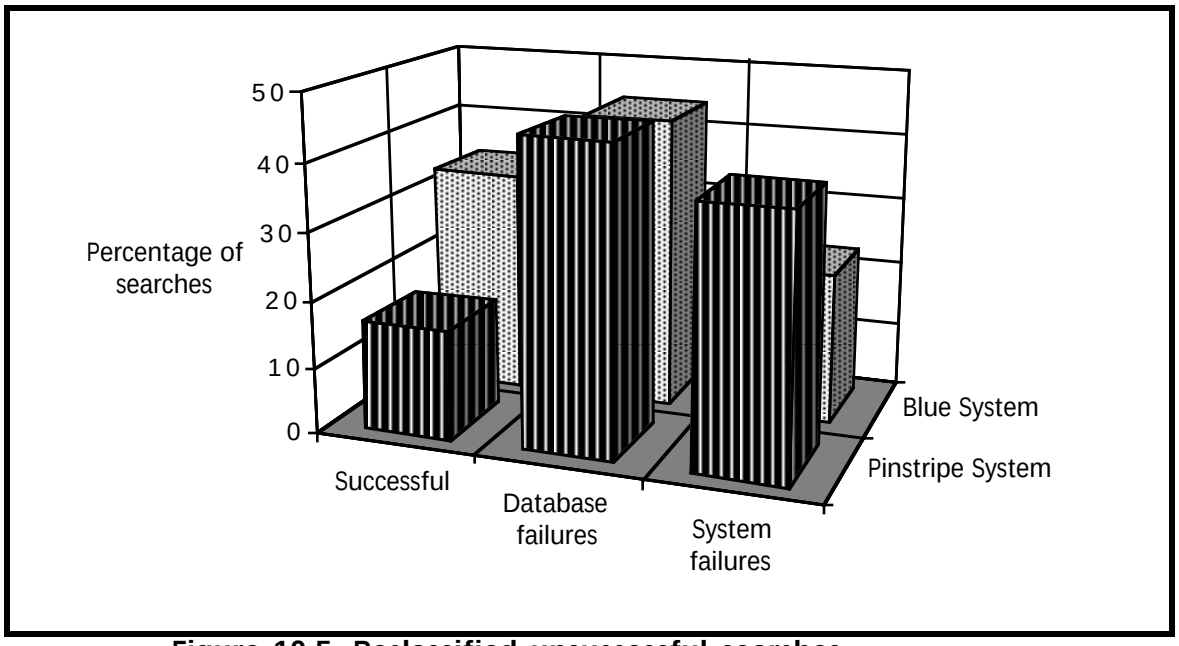


**Figure 10.5. Reclassified unsuccessful searches (both systems, UM-D)**

### 10.2.2 Failure of One System

A total of 147 searches failed to yield at least one useful title in one of the two experimental systems (see Table 10.1). Users performed 69 searches in the Blue System and 78 searches in the Pinstripe System. Of the total of 147 unsuccessful searches, 95 searches were successful in retrieving at least one useful title in one of the two experimental online catalogs. The failure analysis involved recreating the search in both systems — the system that failed to yield useful titles and the system that was successful in producing useful titles — to determine the reason for failure in the former systems. Users performed the remaining 52 (of 147) searches in only one of the two experimental online catalogs. The failure analysis involved recreating the search in the one system searched by the user. It also involved performing a search for the same topic in the other system to determine whether users would have been successful retrieving useful titles had they searched two systems for their topics of interest.

This failure analysis of searches resulted in a categorization of searches based on the problems that adversely affected the retrieval of potentially useful information from the experimental systems. Problems were the same as problems for the failure of searches in both systems except for the addition of a new category for spelling problems.

The nine categories were as follows: (1) user perseverance, (2) specificity that was not represented in user queries, (3) conflicting relevance assessments, (4) database failures, (5) large retrievals, (6) user query vocabulary, (7) subject search approach failures, (8) navigational problems, and (9) spelling problems. Subsequent sections discuss these problems as they pertain to unsuccessful searches in one of the two experimental systems.

#### 10.2.2.1 User Perseverance

User perseverance was a common reason for the failure of searches to retrieve useful titles. For example, a user entered the subject query "turbo prolog" into the Blue System. It responded with an exact-approach main menu with enabled options for displaying titles, broader terms, and expanding the search. The user ended the search. Had the user selected the titles-display option, he would have retrieved the title *Introduction to Turbo Prolog.* This and the many other searches like it should not be counted as unsuccessful searches. The catalog retrieved titles on the subject queries that users entered. Users chose to terminate their searches rather than display titles.

Perseverance was an especially frequent reason for the failure of searches in the Blue System. Perseverance was attributed to the failure of over two-thirds (68.1%) of the 69 failed searches in the Blue System. In contrast, perseverance was attributed to the failure of a little over one-third (37.2%) of the 78 failed searches in the Pinstripe System.

Table 10.11 recounts user and system activity that resulted in attributing the failure of three more searches conducted in the Blue System to user perseverance.

Two of the three user perseverance failures described in Table 10.11 were connected with controlled vocabulary search approaches in the Blue System. Of the 47 failures due to user perseverance in the Blue System, 38 failures involved controlled vocabulary search approaches. Subject searching approaches that featured matches of controlled vocabulary terms (e.g., exact, alphabetical, keyword-in-main-heading) offered users many options for furthering their search. Perhaps users were overwhelmed with the array of available options and made actions to terminate their searches or start them over in a different way. Failure due to perseverance was attributed to 29 searches in the Pinstripe System but there was not the preponderance of such failures for controlled vocabulary approaches in the Pinstripe System as there was in the Blue System.

**Table 10.11. User Perseverance**

| User query | Explanation |
|---|---|
| flowcharting | Blue System retrieves 8 titles in a title-keyword search. The user walks away without rating even the first title. |
| computer aided design | Blue System responds with the results of an exact search in which options for displaying general works, narrower terms, subject and form subdivisions, and for expanding the search are enabled. The user ends the search. |
| auto | Blue System responds with the alphabetical approach bearing subject headings beginning with the word "auto," e.g., "Auto mechanics." User browses forward to "Automobile maintenance" which is a cross reference for "Automobiles — Motors — Maintenance and repair," selects it, but does not display retrieved titles such as *About your car: a basic guide to understanding and maintaining automobiles*, *Automotive diagnosis and tuneup,* and *The back-yard mechanic*. |

Section 10.2.2.8 discusses navigational problems connected with controlled vocabulary search approaches in the Blue System and makes suggestions for improving the design of these approaches. These suggestions could also address problems associated with user perseverance, and, thus, reduce the numbers of users who walk away from the results of controlled vocabulary search approaches.

#### 10.2.2.2 Specificity of User Queries

The problem of specificity of user queries adversely affected the retrieval of useful titles in four Blue-system searches and in four Pinstripe-system searches. The users whose searches were classified in this category probably had a different (and maybe related) topic in mind but, for whatever reason, they entered a query into the experimental online catalog that was not indicative of their information needs.

Table 10.12 recounts user and system activity that resulted in attributing the failure of two searches conducted in the Blue and Pinstripe Systems to specificity of user queries.

**Table 10.12. Specificity of User Queries**

| User query | Explanation |
|---|---|
| environment | Blue System informs user that the entered term is a cross reference for "Ecology" or "Man — Influence of environment." User chooses the former. Blue System responds with enabled exact-search options for displaying general works, narrower terms, broader terms, notes, subject and form subdivisions, and expanding the search. User browses notes, narrower terms, and subject subdivisions. User displays one title assigned the narrower term "Ecology — Vocational guidance" and rates it "not useful." (In a Pinstripe-system search for this topic, the user rates 12 displayed titles "very useful" that were retrieved by the narrower term "Buildings — Environmental engineering.") |
| fluids | Pinstripe System places the user at "Fluids" in an alphabetical list of subject headings. User selects this subject heading and browses several broader terms and subject subdivisions. Of the total 25 titles displayed, the user rates none "very useful." (In Blue-system search for this topic, the user gives a "very useful" rating to 1 of 67 titles displayed through the exact approach. The title is *The analysis and design of pneumatic systems.*) |

The descriptions of the two searches in Table 10.12 illustrated how users began with a query on a specific topic but the titles they rated "very useful" or the subject headings that delivered them to useful information were not always the same as their initial query.

#### 10.2.2.3 Conflicting Relevance Assessments

The researchers attributed the failure of five searches (two in the Blue System and three in the Pinstripe System) to conflicting relevance assessments. An example of a search in which we attributed failure to conflicting relevance assessments was a search of the Pinstripe System for "vibration." The user began searching in the Blue System and retrieved 54 titles in a title-keyword search. She displayed nine titles and gave four titles "very useful" ratings. She searched the Pinstripe System for the same topic, retrieved 85 titles in a keyword-in-record search, and displayed the first of 85 titles. She gave this title a "not useful" rating and ended the search. This title was one of the four titles the user rated "very useful" in her Blue-system search for the same topic.

#### 10.2.2.4 Database Failure

The failure of nine searches in the Blue System and the failure of nine searches in the Pinstripe System were attributed to database failure. Users performed eight of the nine Blue-system searches in the Blue System only. Users performed all nine Pinstripe-system

searches in the Pinstripe System only. The one search performed in two systems was a search for "eiffel." The Blue System responded with the alphabetical approach and placed the user at the subject heading "Eiffel Tower (Paris, France)." The user ended the search. The user's Pinstripe-system search for the same topic yielded one title in a keyword-in-record search for the title *Gustave Eiffel* to which the user gave a "very useful" rating. Since the query immediately preceding this search had been given as "eiffel language," we felt that the user was interested in the "Eiffel" computer language and would have been more satisfied with a title we found in Michigan's Mirlyn online catalog — *Eiffel: the language*. Since the UM-D database did not contain this title, we considered this search of the Blue System to be a database failure.

Examples of other queries that failed to retrieve useful titles in both Blue and Pinstripe Systems due to database failure were:

- smalltalk
- propeller
- maximum shear stress criteria
- solar powered automobiles
- poka-yoke system

#### 10.2.2.5 Large Retrievals

Large retrievals hampered six Pinstripe-system searches. Retrievals in all but one search were produced by keyword-in-record searches. Table 10.13 summarizes these searches.

**Table 10.13. Large Retrievals**

| Query | Search approach | Retrievals |
|---|---|---|
| fortran | keyword-in-record | 136 titles |
| science | keyword-in-record | 1,207 titles |
| data structures | keyword-in-record | 61 titles |
| materials | keyword-in-subdivided-heading | 120 subject headings |
| basic | keyword-in-record | 300 titles |
| architecture | keyword-in-record | 106 titles |

In all but one of the searches above, users walked away or they rated the first title that the system automatically displayed to them and then they walked away. In the search for "architecture," the user displayed 14 of 106 retrieved titles before terminating the search.

#### 10.2.2.6 Vocabulary of User Queries

Vocabulary problems adversely affected three searches in the experimental online catalog. Users conducted one of the three searches in the Blue System and two of the three searches in the Pinstripe System. In the left column of Table 10.14 are listed queries that failed due to their vocabulary. Under the right-hand column are suggested ways in which systems could process queries to retrieve titles and subject headings on users' topics of interest.

**Table 10.14. Vocabulary Problems**

| User query | Explanation |
|---|---|
| fundamental of data structures in c | Pinstripe System's keyword-in-subdivided-heading search fails to produce retrievals. If this query were reduced to "data structures in c," the experimental system would retrieve *Data structures: an advanced approach using C* in a title-keyword search. |
| how antilock brakes work | Pinstripe System's alphabetical approach places the user in an alphabetical list at the cross reference "How-to-do-it guides" for the subject heading "Do-it-yourself work." If this query were reduced to "antilock brakes," title-keyword and keyword-in-record searches would retrieve several titles. |
| reinforce concrete | Blue System exhausts search trees without retrieving titles or subject headings. If this query were truncated, the system would retrieve several titles bearing the subject heading "Reinforced concrete." |

#### 10.2.2.7 Subject Searching Approach Failures

Of the 24 search approach failures, only one failure occurred in the Blue System. This search began with the user-entered query "computers made simple." The Blue System exhausted the search trees without making a match. The Pinstripe System responded with the alphabetical approach. The user browsed forward in the alphabetical list to the subject heading "Computer," selected the heading, and gave "very useful" ratings to four titles. An example of a useful title was *High tech society: the story of information technology*. Table 10.15 summarizes search approach failures characteristic of searches in the Pinstripe System.

**Table 10.15. Subject Searching Approach Failures in the Pinstripe System**

| User query | Explanation |
|---|---|
| car suspension | Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (Blue System retrieves 2 titles in a title-keyword search for this query that the user rates "very useful.") |
| hovercraft | Pinstripe System responds to this query with the keyword-in-record search and is unsuccessful retrieving titles. (In the Blue System, "hovercraft" is a cross reference for "Ground-effect systems." Exact approach yields titles such as *The glideway system: a high-speed ground transportation system in the northeastern corridor ...* and *Jane's surface skimmer systems.*) |
| tcp/ip | Pinstripe System responds to this query with the alphabetical approach and places the user at the subject heading "TI Programmable 59 (Calculator)." (Blue System retrieves 3 titles in a title-keyword search for this query that the user rates "very useful.") |
| cad | Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (In the Blue System, "cad" is a cross reference for "Computer-aided design" which leads the user to useful titles.) |

In the first-listed search in Table 10.15, the Pinstripe System failed to retrieve titles through the keyword-in-subdivided-heading search and the Blue System was successful retrieving titles through a search selected by the search trees. Of the 23 Pinstripe-system searches that failed due to subject search approaches, this was by far the most common type of failure as it adversely affected 13 searches in the Pinstripe System.

Four searches failed because the Pinstripe System failed to retrieve titles through the keyword-in-record search. Blue-system searches for the same topic succeeded because the exact or alphabetical approaches in the Blue System featured cross references with terminology that matched user queries and delivered users to subject headings on their topics of interest. Pinstripe-system searches for these topics failed because the terminology used in cross references was not used in any subject-bearing fields of bibliographic records. Another four searches failed because the Pinstripe System responded with the alphabetical approach in which listed subject headings were not pertinent to users' topics of interest. Two Pinstripe-system searches failed because the subject heading(s) the system retrieved in a keyword-in-subdivided-heading search were not pertinent to users' topics of interest.

Subject searching approach failures in the Pinstripe System accounted for almost three-tenths (23 of 78 or 29.5%) of search failures. The Pinstripe System selected subject searching approaches randomly. Subject searching failures were rare in the Blue System, accounting for only one failure in 69 searches. The Blue System was guided in its selection of a subject searching approach by the search trees.

#### 10.2.2.8 Navigational Problems

Navigational problems occurred in Blue-system searches. The two searches plagued by such problems required users to navigate the many options connected with controlled vocabulary approaches to find subject headings of interest. In a search for "animation," the Blue System responded to the user query with the alphabetical approach and placed the user near subject headings such as "Animation, Computer" and "Animated cartoons by computer." The user browsed forward to "Casting, Steel," selected this subject heading, and displayed titles even though it was not related to his original query "animation."

#### 10.2.2.9 Spelling Problems

Spelling problems adversely affected the success of five searches in the experimental online catalog. Three searches were conducted in the Blue System and two searches were conducted in the Pinstripe System. In all five searches, users ignored the system's message that asked them to check their spelling and repeatedly clicked on the <Do search> option following the entry of their misspelled queries. These queries were:

- elecctronics
- assemlby language
- numeric rec
- software configration
- graphtheory

The systems specifically brought the potentially misspelled word to the user's attention. For example, figure 10.6 shows the system's response to the query "assemlby language" in which the system informed the user that the word "assemlby" might be misspelled.

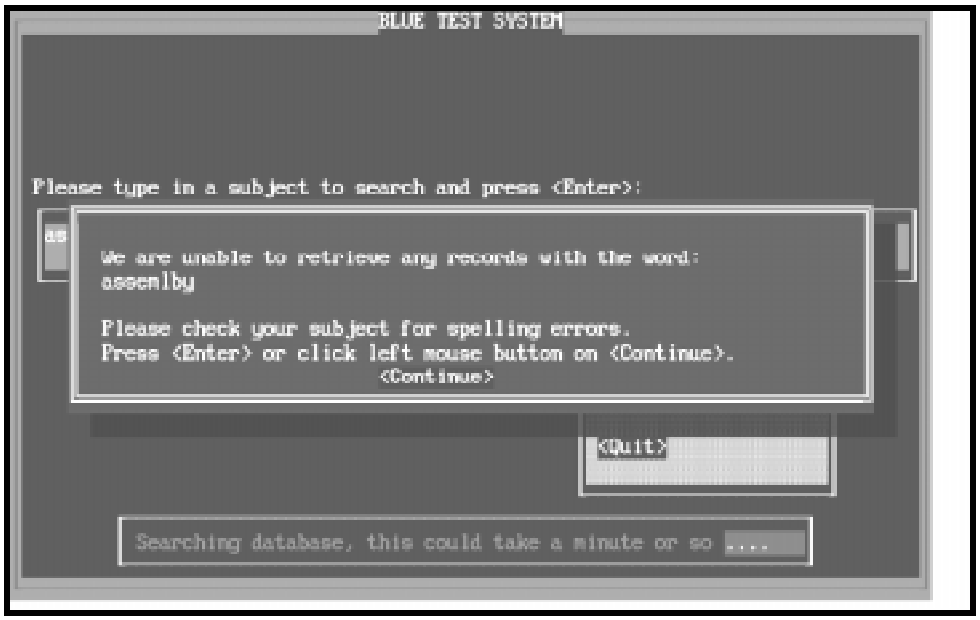


**Figure 10.6. System response to misspelled query words**

Computer-assisted spelling correction could have corrected these five queries. When users repeatedly ignore system messages to correct spelling, systems could automatically call up spelling correction algorithms and suggest corrections to possible spelling errors.

#### 10.2.2.10 Failure of One System Revisited

Searches that failed to retrieve useful retrievals in one system numbered 147. A failure analysis of these searches demonstrated the following problems that adversely affected the retrieval of potentially useful information from the experimental online catalogs: (1) user perseverance, (2) specificity that was not represented in user queries, (3) conflicting relevance assessments, (4) database failures, (5) large retrievals, (6) user query vocabulary, (7) search approach failures, (8) navigational problems, and (9) spelling problems.

Problems with user perseverance, specificity of user queries, and conflicting relevance assessments had little to do with the ability of systems to retrieve useful information. Sections 10.2.2.1 to 10.2.2.3 gave several examples of searches that would have retrieved potentially useful titles but for user actions that prevented retrieval. In this failure analysis, we considered such searches successful.

Database failures are bound to occur in searches of any database. And users are likely to search databases for information on topics that the database does not cover or covers only marginally. Database failures are a type of system failure. Other failures

connected with systems were due to large retrievals, user query vocabulary, subject searching approach failures, and navigational problems. Throughout the sections that described searches that failed for these reasons, we offered suggestions for making improvements to systems to reduce failure in future experimental and operational systems.

Table 10.16 reclassifies unsuccessful searches in the Blue and Pinstripe Searches. The reclassification is limited to the 147 unsuccessful searches conducted in both Blue and Pinstripe Systems and featured in section 10.2.2.

**Table 10.16. Reclassified Searches (Single system, UM-D)**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| Successful searches† | 53 | 76.8 | 36 | 46.2 |
| Database failures | 9 | 13.0 | 9 | 11.5 |
| System failures* | 7 | 10.2 | 33 | 42.3 |
| Total | 69 | 100.0 | 78 | 100.0 |

†Barring perseverance, query specificity, conflicting relevance assessments.

*Including problems with large retrievals, subject searching approaches, user query vocabulary, navigation, spelling.

Database failure accounted for less than fifteen percent of unsuccessful searches in the Blue and Pinstripe Systems. A much larger percentage (42.3%) of system failures was connected with the Pinstripe System than with the Blue System (10.2%). Many of these failures were search approach failures in the Pinstripe System. The failure analysis promoted some unsuccessful searches to successful searches. Over three-quarters of Blue-system searches and a little under one-half of Pinstripe-system searches were successful in terms of guiding users to titles on the topics their queries described. Many of the searches were promoted to successful status were due to perseverance in the Blue System.

Figure 10.7 is a graphical depiction of the statistics in Table 10.16. It shows obvious differences in the percentages of successful searches and system failures in Blue and Pinstripe Systems.

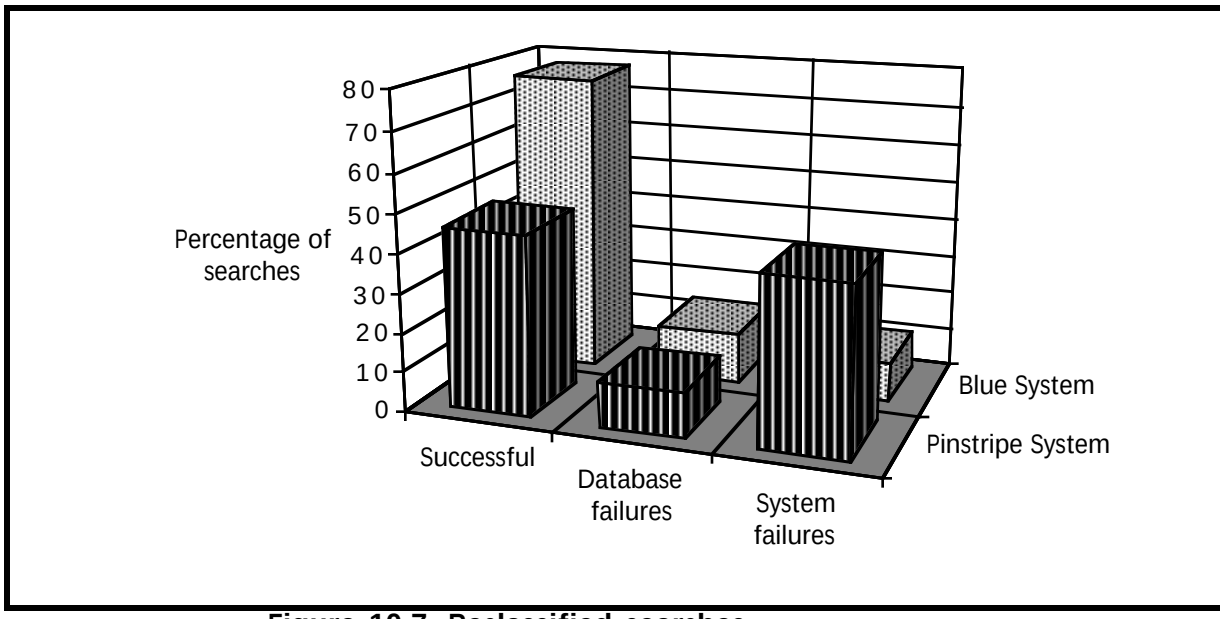


**Figure 10.7. Reclassified searches (Single system, UM-D)**

### 10.2.3 Revisiting Successful and Failed Searches

In the failure analysis of searches conducted by UM-D users, the researchers repeated user searches that failed to retrieve useful titles in an attempt to determine why they were unsuccessful in retrieving useful titles. In the process of repeating these searches, the failure analysis demonstrated that several searches might have been headed toward success but for various user actions that precluded the display of useful titles, e.g., perseverance, queries that were not indicative of user interests, conflicting relevance assessments. The failure analysis also demonstrated reasons for failure, e.g., design problems with a particular search approach or indexing technique, inability of a randomly-selected approach in the Pinstripe System to retrieve any titles in response to user queries. We can learn from the failure analysis in terms of designing subject searching approaches and reconfiguring the search trees. Chapter 12 makes recommendations for improving subject access in online catalogs based on failure analysis results.

Table 10.17 recasts the number and percentage of successful and unsuccessful searches based on the failure analysis of end-user searches at the University of Michigan-Dearborn.

**Table 10.17. Reclassified Successful and Unsuccessful Searches at UM-D**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Type of search** | **No.** | **%** | **No.** | **%** |
| **Successful searches:** | | | | |
| 1 useful titles retrieved | 156 | 44.2 | 150 | 42.1 |
| Retrieval of useful titles marred by perseverance | 73 | 20.7 | 42 | 11.8 |
| Retrieval of useful titles marred by query specificity | 16 | 4.5 | 9 | 2.5 |
| Retrieval of useful titles marred by conflicting relevance assessments | 7 | 2.0 | 6 | 1.7 |
| Subtotal | 252 | 71.4 | 207 | 58.1 |
| **Unsuccessful searches:** | | | | |
| Database failure | 68 | 19.2 | 70 | 19.6 |
| Search approach failure | 4 | 1.1 | 42 | 11.9 |
| Large retrievals | 1 | 0.3 | 19 | 5.3 |
| Vocabulary of user queries | 18 | 5.1 | 16 | 4.5 |
| Navigation | 7 | 2.0 | 0 | 0.0 |
| Spelling | 3 | 0.9 | 2 | 0.6 |
| Subtotal | 101 | 28.6 | 149 | 41.9 |
| Total | 353 | 100.0 | 356 | 100.0 |

A little under three-quarters of Blue-system searches were successful; about three-fifths of Pinstripe-system searches were successful. A large percentage (20.7%) of Blue-system searches would have been successful but users did not exhibit sufficient perseverance to stay with the search and see it to its conclusion. This percentage was somewhat smaller in the Pinstripe System (11.8%).

Although we have no user interview data, we can speculate on the reasons why many Blue-system users exhibited low levels of perseverance and terminated their searches prematurely. Our speculation is fueled by our experience searching the Blue and Pinstripe Systems and searching computer-based systems generally. We are quite sure that user perseverance in the Blue System was connected with amount of time the system required for users to browse the many options available in alphabetical and exact searches. These approaches allowed users to select listed subject headings, narrower terms, broader terms, or subdivided forms of selected subject headings. It took time for users to navigate between options and for the system to search for selected

subject headings. Recommendations for improving alphabetical and exact searches made in the lengthy section on navigational problems included (section 10.2.1.8):

- Reducing the response time for searching selected terms and creating exact-search option menus.
- Enlisting exact-search option menus to summarize large retrievals.
- Requiring the system to check whether keyword searches on user-entered queries that match cross references retrieve additional titles and enable/disable the "expand search" option based on this system check.

Of the several reasons for unsuccessful searches, database failure affected about the same percentage (a little under twenty percent) of searches conducted in the Blue and Pinstripe Systems. Vocabulary of user queries was also troublesome and led to about five percent of search failures in both Blue and Pinstripe Systems. Large retrievals were usually connected with searches in the Pinstripe System. Disregarding database failures, the largest percentage of unsuccessful searches was for search approach failures in the Pinstripe System (11.9%). Only 1.1% of searches in the Blue System failed because of subject searching approaches. Much of the time, the Pinstripe System's keyword-in-subdivided-heading search was unable to retrieve subject headings bearing words in user queries or subject headings that were pertinent to users' topics of interest. On occasion, its alphabetical approach placed users in an alphabetical neighborhood that was not fruitful for enabling users to find satisfactory subject headings for expressing their topics of interest.

The Pinstripe System's selection of a subject searching approach in response to a user-entered query was based on a random selection algorithm. In contrast, the Blue System's selection of a subject searching approach was governed by search trees. These data provide evidence to support the development of search trees in online catalogs. In a catalog without search trees, a large percentage of end-user searches are bound to fail because systems do not assist end users in subject searching approach selection.

## 10.3 Successful and Unsuccessful Searches at Earlham

Section 10.3 gives the results of the failure analysis of searches performed in the Blue and Pinstripe Systems at Earlham College. Problems were the same as problems for the failure of searches in both systems and in a single system at UM-D. Recommendations for system improvements that were discussed in section 10.2 for unsuccessful searches at UM-D were also applicable to unsuccessful searches at Earlham.

Table 10.18 lists the number and percentages of successful and unsuccessful Blue and Pinstripe Searches.

**Table 10.18. Successful and Unsuccessful Searches at Earlham**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| 1 useful titles retrieved | 46 | 44.2 | 37 | 37.8 |
| No useful titles retrieved in one of two systems | 22 | 21.2 | 25 | 25.5 |
| No useful titles retrieved in searches of both systems | 36 | 34.6 | 36 | 36.7 |
| Total | 104 | 100.0 | 98 | 100.0 |

One or more useful titles were retrieved in 44.2% of searches in the Blue System and in 37.8% of searches in the Pinstripe System. The Blue System had a lower percentage (21.2%) of searches in which no useful titles were retrieved than in the Pinstripe System (25.5%). Thus, the figures in Table 10.18 indicate that the Blue System performed a little better than the Pinstripe System in terms of retrieving useful titles in searches. These percentages were quite similar to the percentages of successful and unsuccessful searches in the experimental online catalog at UM-D (see Table 10.1).

## 10.3.1 Failure of Both Systems

At Earlham, users performed a total of 86 pairs of searches. That is, they performed one search for their topic of interest in the Blue System followed by a search for the same topic in the Pinstripe System. Of the 86 pairs, 36 search pairs failed to yield useful titles in searches of both experimental systems — about 42% of the search pairs. The project director, an expert searcher of information retrieval systems, repeated these three dozen searches in both Blue and Pinstripe Systems to determine why end users were unsuccessful in retrieving useful titles.

### 10.3.1.1 User Perseverance

User perseverance was a common reason for the failure of searches to retrieve useful titles. For example, a user entered the personal-name query into the Blue System "jefferson" (last name), "thomas" (first name), and "inventions" (topic). The Blue System was unable to make retrievals through the keyword-in-record search for the combination of name and topic elements so it ignored the topic element and showed the user the results of an alphabetical search. That is, it placed the user in an alphabetical list of subject headings for personal names and highlighted the best matching subject heading, "Jefferson, Thomas, 1754–1826." The user selected this subject heading and

displayed only the first of 62 retrieved titles. Had the user continued, the Blue System would have displayed titles on the user's topic of interest, e.g., *Thomas Jefferson: statesman of science, Thomas Jefferson: landscape architect, Jefferson and the arts,* and *Thomas Jefferson: the man, his world, his influence*.

Table 10.19 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to user perseverance.

**Table 10.19. User Perseverance**

| User query | Explanation |
|---|---|
| women | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, notes, topical and place subdivisions, and for expanding the search are enabled. The user ends the search. |
| reconstruction | In the course of performing an alphabetical search, Pinstripe System gives the user a choice of two "Reconstruction" subject headings. The user selects the heading assigned to titles on the Reconstruction following the Civil War but he does not display titles connected with the selected heading. |
| reconstruction | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, notes, topical, place, and form subdivisions, and for expanding the search are enabled. The user browses and selects several subdivisions connected to the main heading but he does not display titles connected with the selected heading. |
| hoover herbert economic | Failing to produce retrievals for combinations of name and topic elements, Blue System ignores the topic element, responds with the alphabetical approach which highlights the personal-name heading "Hoover, Herbert, 1874–1964." User selects this heading and displays one of 28 retrieved titles. Had the user continued displaying titles, he would have found titles such as *The shattered dream: Herbert Hoover and the great Depression* and *The poverty of abundance: Hoover, the nation, the depression.* |

If users had displayed several titles for the searches described in Table 10.19, they would have most likely succeeded in retrieving useful titles.

In the absence of interview data, we can speculate that the results of alphabetical or exact searches could have been too complex for users to understand. That is, they could have been lost or confused about which of the several options to pursue for continuing the search. This does not explain what was in the mind of the user who conducted the search for Herbert Hoover. The Blue System retrieved a manageable number of titles (28) and the user simply did not display additional titles.

A total of ten searches failed because of user perseverance. All but one of these searches were conducted in the Blue System. All nine searches conducted in the Blue System involved controlled vocabulary approaches. Perseverance was also a problem with failed searches at UM-D (see sections 10.2.1.1 and 10.2.2.1).

The failure analysis demonstrated that searches involving perseverance were really successful searches in terms of leading users to titles on their topics of interest. The problem was that users failed to display any or enough titles to find useful ones.

#### 10.3.1.2 Specificity of User Queries

The problem of specificity of user queries adversely affected the retrieval of useful titles in four Blue-system searches and in three Pinstripe-system searches. Users whose searches were classified in this category probably had a different (and maybe related) topic in mind but, for whatever reason, they entered queries into the experimental online catalog that were not indicative of their information needs.

Table 10.20 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to specificity of user queries.

Perhaps, the topics that users had in mind were not quite the same as the queries they entered into the experimental online catalog. These searches were really successful searches in terms of leading users to titles on the topics they expressed in the queries they entered into the catalogs. However, the searches were not successful in terms of retrieving titles that fulfilled the specifications of queries users had in their minds but did not make explicit in their queries.

**Table 10.20. Specificity of User Queries**

| User query | Explanation |
|---|---|
| boston tea party | Blue System responds with the results of an alphabetical search and highlights the subject heading "Boston Tea Party, 1773." The user selects the highlighted heading, displays the one title assigned this heading, i.e., *The Boston Tea Party*, and gives it a "not useful" rating. |
| richmond, indiana | Pinstripe System retrieves 16 titles through a keyword-in-record search. The user does not give "very useful" ratings to any of the 16 titles. |
| quaker | Blue System responds with the results of an alphabetical search that includes the subject headings "Quakers," "Quaker meeting houses," "Quaker church buildings." User displays titles assigned to these and other listed subject headings, expands the search by retrieving the results of a title-keyword search for this query. Examples of retrieved titles are *The Quaker colonies: a chronicle of the proprietors of the Delaware* and *The Dutch and Quaker colonies in America.* User does not give "very useful" ratings to the dozen titles displayed in this search. |

#### 10.3.1.3 Database Failure

Table 10.21 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to database failure.

In all three searches described in Table 10.21, the systems retrieved very few titles on the topics of user queries and displayed titles were probably too specific for user needs (e.g., the title on the hanging of a sheriff in nineteenth-century Montana in response to a query on 19th century Montana), or could have covered topics besides the one(s) users had in mind (e.g., the title on the witchcraft of native American peoples in response to the query on witchcraft). Post-search personal interviews would have provided explanations from users as to their reasons for giving negative relevance assessments to retrieved titles.

A total of 28 searches failed because of database failure. These 28 searches were pairs of searches, that is, searches conducted by fourteen different end users in the Blue and Pinstripe Systems.

**Table 10.21. Database Failure**

| User query | Explanation |
|---|---|
| folk songs | Blue System responds with four subject headings retrieved in a keyword-in-subdivided-heading search. The user selects the subject heading "Folk songs, American — New York," displays one of two titles assigned this heading, i.e., *Body, boots, and britches: folktales, ballads, and speeches from country New York* and gives it a "not useful" rating. |
| witchcraft | Pinstripe System responds with three subject headings in a keyword-in-subdivided-heading search. The user selects the subject heading "Witchcraft — New England" and gives all three displayed titles negative relevance assessments. Examples of displayed titles are *Witchcraft and sorcery of the American native peoples* and *Navaho witchcraft*. |
| montana 19th century | Pinstripe System displays the one title *Hanging the sheriff: a biography of Henry Plummer* retrieved in a keyword-in-record search. User gives it a "not useful" rating. |

A few searches in these pairs deserve special mention. Three searches were for combinations of names and topics:

- "king" (last name), "martin luther" (first name), "about the bycotts [sic] they had way back in mississippi" (topic)
- "king" (last name), "coretta scott" (first name), "learn about what happened in 1927" (topic)
- "king" (last name), "martin luther" (first name), "about in the 1925" (topic)

Since these three queries were alike in form, content, and were entered successively into the experimental online catalog, they were probably entered by the same individual. The Blue System responded to all three queries by ignoring the topic elements and placing users in an alphabetical list of personal-name headings where they could find the names "King, Coretta Scott, 1927–" and "King, Martin Luther, Jr., 1929–1968." The user selected these subject headings, displayed titles such as *Coretta: the story of Mrs. Martin Luther King, Jr., There is a balm in Gilead: the cultural roots of Martin Luther King, Jr.,* and *Martin Luther King, Jr.: the making of a mind.* The user did not give positive relevance assessments to any displayed titles despite their promising content. Most likely, this user would have only given positive relevance assessments to titles that explicitly referred to the specific topic the user made explicit in his query.

The Pinstripe System performed keyword-in-record searches for all three queries and was unsuccessful retrieving titles for them. One could consider this a system failure

instead of a database failure because the Pinstripe System failed to guide the user to potentially useful titles. However, if we attributed the failure of one of two searches in a pair to database failure, we almost always considered the second of the pair as a failure due to database failure. Other examples of queries that failed to retrieve useful titles in both Blue and Pinstripe Systems due to database failure are:

- dayton history
- women in reconstruction
- repeal of prohibition
- minorities in the west
- chicago strike

#### 10.3.1.4 Large Retrievals

We did not expect large retrievals to be a problem in searches at Earlham College because the database of the experimental online catalog was quite small. Yet, a few topics were represented by large numbers of bibliographic records, e.g., Afro-Americans, Civil War, history, Indians, Jews, presidents, and slavery.

Large retrievals hampered two Pinstripe-system searches. In both searches, the system responded to the user-entered query "civil war" with a keyword-in-record search and retrieved 446 titles. In both searches, users failed to display more than the first of 446 retrieved titles.

#### 10.3.1.5 Vocabulary of User Queries

Vocabulary problems adversely affected retrieval in 17 searches in the experimental online catalog. In the left column of Table 10.22 are listed four queries with vocabulary problems. Under the right-hand column are suggested ways in systems could process queries to retrieve titles and subject headings on users' topics of interest.

Of the seventeen queries that failed due to the vocabulary of user queries, eight queries had vocabulary problems that adversely affected both searches in Blue and Pinstripe systems. Vocabulary problems adversely affected one query in the Blue System; the Pinstripe-system search for this same query failed because of search approach problems. Thus, nine searches conducted in the Blue System and eight searches conducted in the Pinstripe System failed because of the vocabulary of user queries.

**Table 10.22. Vocabulary Problems**

| User query | Explanation |
|---|---|
| black civil rights in the reconstruction | In the Pinstripe System, the query "reconstruction and civil rights" retrieves three titles in a keyword-in-record search: *Death at Cross Plains: an Alabama Reconstruction tragedy, Reconstruction, the Negro, and the New South*, and *Politics, principles, and prejudice, 1865–1866 — dilemma of Reconstruction America.* |
| hopi navajo indian land dispute | In the Blue System, query for "hopi navajo" retrieves one title in a keyword-in-record search, *The second long walk: the Navajo-Hopi land dispute.* |
| navajo crafts | In the Blue System, query for "navajo" results in an alphabetical search. Selection of the longer subject heading "Navajo Indians" results in an exact search options for subject subdivisions. Under the subdivisions "Material culture" and "Textile industry and fabrics" are potentially useful titles such as *Navajo material culture, Walk in beauty: the Navajo and their blankets,* and *Navaho weaving: its technic and its history.* |
| origins of the ku klux klan | In the Blue System, query for "ku klux klan" prompts the system to respond with the exact approach. The general works option produces titles such as *White terror: the Ku Klux Klan conspiracy and Southern Reconstruction* and *The Ku Klux Klan: a century of infamy* |

In a large database, the vocabulary of some queries — "navajo crafts," "origins of the ku klux klan," "black civil rights during the reconstruction" — would not adversely affect the search because there would be a much larger database of unique words to combine in keyword searches. However, the vocabulary of the queries in this category of failed searches adversely affected retrieval in a small database such as the experimental online catalog database. If we could redesign the ASTUTE experimental online catalog, we would add system capabilities such as automatic stemming and/or best match to improve retrieval for the queries in this failure analysis category.

#### 10.3.1.6 Subject Searching Approach Failures

The experimental online catalog contained information on topics that interested several users; however, it failed to retrieve the desired information. For example, in section 10.3.1.1, we discussed a search for "jefferson" (last name), "thomas" (first name), and "inventions" (topic) that failed to produce useful retrievals in the Blue System because of end-user perseverance. The user's search in the Pinstripe System failed for a different reason. That is, the Pinstripe System conducted a keyword-in-record search for this query and failed to retrieve titles.

The failure of eight searches was attributed to subject searching approach failures. All eight searches were conducted in the Pinstripe System. This system failed to produce potentially useful retrievals through the subject searching approach that the system chose randomly. Blue-system searches for the same topics failed for reasons connected with user perseverance, query specificity, or vocabulary.

Table 10.23 lists user queries and gives explanations for subject searching approach failures in Pinstripe-system searches.

**Table 10.23. Subject Searching Approach Failures in the Pinstripe System**

| User query | Explanation |
|---|---|
| african-americans | The Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (Perseverance adversely affects the Blue-system search for this query.) |
| hoover herbert economic | The Pinstripe System responds to this query with the keyword-in-record search and is unsuccessful retrieving titles. (Perseverance adversely affects the Blue-system search for this query.) |
| slavery women | The Pinstripe System responds to this query with the keyword-in-subdivided-heading search and is unsuccessful retrieving subdivided headings. (Query specificity adversely affects the Blue-system search for this query.) |
| treatment of women slaves during civil war | The Pinstripe System responds to this query with the alphabetical approach. The user selects "Treaty of Ghent" but does not display titles assigned the selected heading. (Vocabulary adversely affects the Blue-system search for this query.) |
| women | The Pinstripe System responds to this query with the alphabetical approach. The user selects "Women spies," browses broader terms, selects "Spies," displays one title and deems it "not useful." (Perseverance adversely affects the Blue-system search for this query.) |

Most subject searching approach failures were attributed to searches that completely failed to retrieve titles.

A few searching approach failures were connected with user problems navigating alphabetical lists of subject headings. For example, the two users who entered the queries "treatment of women slaves during civil war" and "women" in Table 10.23 seemed to have difficulty navigating the alphabetical list because the subject headings they selected had nothing to do with their original queries.

#### 10.3.1.7 Failure of Both Systems Revisited

Pairs of searches that failed to retrieve useful retrievals in both systems numbered 36 — a little over two-fifths of the pairs of searches performed in the experimental online catalog at Earlham. A failure analysis of these searches demonstrated the following problems adversely affected the retrieval of potentially useful information from the experimental online catalogs: (1) user perseverance, (2) specificity that was not represented in user queries, (3) database failures, (4) large retrievals, (5) user query vocabulary, and (6) search approach failures.

Problems with user perseverance and the specificity of user queries had little to do with the ability of systems to retrieve useful information. Sections 10.3.1.1 and 10.3.1.2 gave several examples of searches that would have retrieved potentially useful titles but for user actions that prevented retrieval. In this failure analysis, we considered such searches successful.

Database failures were considered a type of system failure. Other failures connected with systems were due to large retrievals, user query vocabulary, and subject searching approach failures; improvements can be made to systems to help them cope with these problems.

Table 10.24 reclassifies unsuccessful searches in the Blue and Pinstripe Searches. The reclassification was limited to the 36 pairs of unsuccessful searches conducted in both Blue and Pinstripe Systems and featured in subsections of section 10.3.1.

**Table 10.24. Reclassified Searches (Both systems, Earlham)**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| Successful searches† | 13 | 36.1 | 4 | 11.1 |
| Database failures | 14 | 38.9 | 14 | 38.9 |
| System failures* | 9 | 25.0 | 18 | 50.0 |
| Total | 36 | 100.0 | 36 | 100.0 |

†Barring perseverance, query specificity.

*Including problems with large retrievals, subject searching approaches, user query vocabulary.

Almost forty percent of unsuccessful searches in the Blue and Pinstripe Systems were attributed to database failure. One-quarter of Blue-system searches and one-half of Pinstripe-system searches were due to system failures. The failure analysis promoted some unsuccessful searches to successful searches. A little over a third of Blue-system

searches and a little over a tenth of Pinstripe-system searches were successful in terms of guiding users to titles on the topics their queries described.

Figure 10.8 is a graphical depiction of the statistics in Table 10.24. It shows obvious differences in the percentages of successful searches and system failures in Blue and Pinstripe Systems.

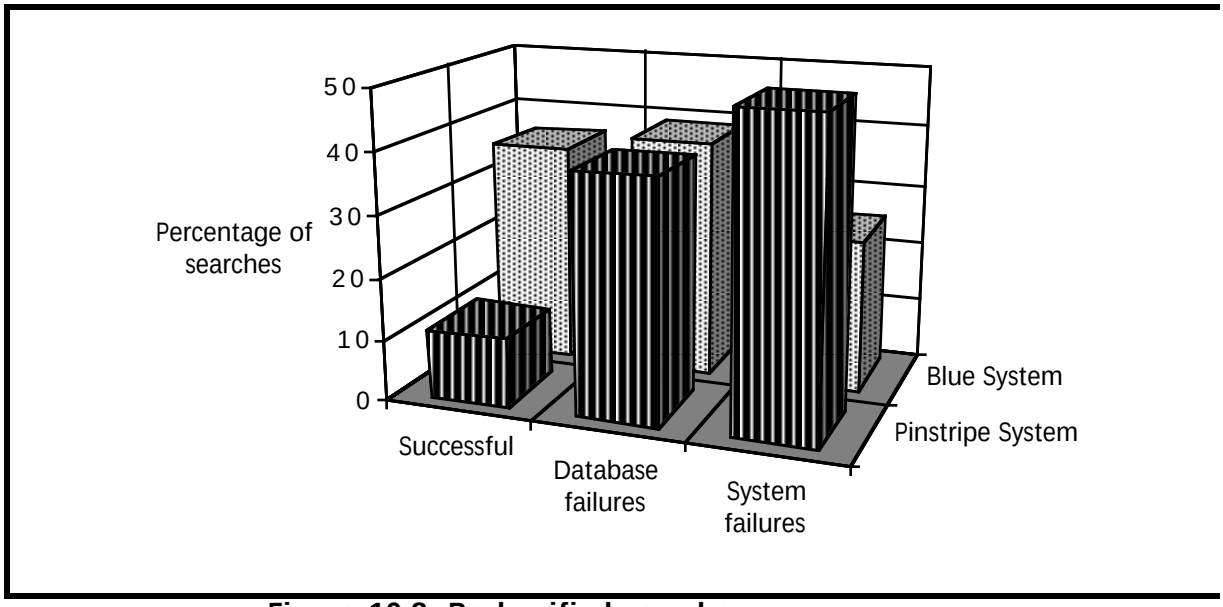


**Figure 10.8. Reclassified searches both systems, Earlham)**

## 10.3.2 Failure of One System

A total of 47 searches failed to yield at least one useful title in one of the two experimental systems (see Table 10.18). Users performed 22 searches in the Blue System and 25 searches in the Pinstripe System. Of the total of 47 unsuccessful searches, 35 searches were successful in retrieving at least one useful title in one of the two experimental online catalogs. The failure analysis involved recreating the search in both systems — the system that failed to yield useful titles and the system that was successful in producing useful titles — to determine the reason for failure in the former systems. Users performed the remaining 12 (of 47) searches in only one of the two experimental online catalogs. The failure analysis involved recreating the search in the one system searched by the user. It also involved performing a search for the same topic in the other system to determine the whether users would have been successful retrieving useful titles had they conducted their searches in two systems.

This failure analysis of searches resulted in a categorization of searches based on the problems that adversely affected the retrieval of potentially useful information from

the experimental systems. Problems were the same as problems for the failure of searches in both systems at Earlham except for the addition of categories for conflicting user relevance assessments and navigational problems.

#### 10.3.2.1 User Perseverance

User perseverance was again a common reason for the failure of searches to retrieve useful titles. For example, a user entered the query "roosevelt eleanor." The Blue System displayed an alphabetical list of subject headings for personal names and highlighted the closest matching name heading "Roosevelt, Eleanor, 1884–1962." The user terminated the search at this point without selecting the subject heading or displaying titles connected to it. Of the ten Blue-system searches marred by user perseverance problems, eight followed on the heels of a Pinstripe-system search for the same topic. Users could have displayed low levels of perseverance in the course of these ten searches because the results of the Blue-system searches were rather similar to the results of the Pinstripe-system searches.

Table 10.25 recounts user and system activity that resulted in attributing the failure of several searches conducted in the Blue and Pinstripe Systems to user perseverance. If users had continued searching, these searches would have most likely succeeded in retrieving useful titles.

Previously, we speculated that the results of controlled vocabulary searches could have been too complex for users to understand (sections 10.2.1.1 and 10.3.1.1). That is, they could have become lost or confused about which of the several options to pursue for continuing the search. This does not explain what was in the mind of the user who conducted the search for "united states imperialism." The Blue System retrieved a manageable number of titles (21), but the user displayed only four titles, and gave all four of them negative relevance assessments.

A total of seventeen searches failed because of user perseverance. Ten of these searches were conducted in the Blue System. The failure analysis demonstrated that these searches were really successful searches in terms of leading users to titles on their topics of interest.

**Table 10.25. User Perseverance**

| User query | Explanation |
|---|---|
| underground railroad | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, notes, place subdivisions, and for expanding the search are enabled. The user ends the search. |
| black women | Pinstripe System's keyword-in-record approach retrieves 42 titles including *Black women in American history: the twentieth century, Black women in American history: from colonial times through the nineteenth century,* and *I dream a world: portraits of Black women who changed America.* The user walks away without rating any titles. |
| minorities | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, notes, place subdivisions, and for expanding the search are enabled. The user terminates the search. |
| united states imperialism | Pinstripe System's keyword-in-record search retrieves 21 titles. The user displays the first four titles and terminates the search. Titles missed include *The rising American empire, American imperialism,* and *American's road to empire: the war with Spain and overseas expansion.* |

#### 10.3.2.2 Specificity of User Queries

The problem of specificity of user queries adversely affected the retrieval of useful titles in three Blue-system searches. The users whose searches were classified in this category probably had a different (and maybe related) topic in mind but, for whatever reason, they entered queries into the experimental online catalog that were not indicative of their information needs.

Table 10.26 recounts user and system activity that resulted in attributing the failure of three searches conducted in the Blue System to specificity of user queries.

**Table 10.26. Specificity of User Queries**

| User query | Explanation |
|---|---|
| jackson andrew jacksonian democracy | Blue System's keyword-in-record search retrieves three titles, *The rise and decline of Jacksonian democracy, The meaning of Jacksonian democracy,* and *Jacksonian democracy, 1829–1837*. The user gives the three titles "possibly useful" ratings. |
| civil war women | Blue System's keyword-in-subdivided-heading search displays the subject heading "United States — History — Civil War, 1861–1865 — Women." The user selects the heading and displays the three titles assigned this heading, *A woman's Civil War: a diary with reminiscences of the war from March 1862, Bonnet brigades,* and *Noble women of the North*. The user gives the three titles negative relevance assessments. |
| communism | Blue System responds with the results of an exact search in which options for displaying general works, broader terms, narrower terms, notes, place and form subdivisions, and for expanding the search are enabled. The user displays ten titles and gives none positive relevance assessments. |

#### 10.3.2.3 Conflicting Relevance Assessments

The failure of three searches in the Blue System was attributed to conflicting relevance assessments. All three searches followed a search of the Pinstripe System for the same topic. When users searched the Pinstripe System, they gave one or more titles "very useful" ratings. When they searched the Blue System, they gave the same title(s) "not useful" or "possibly useful" ratings. For example, a user who conducted a Blue-system search for "black families" rated the title *The Black family in slavery and freedom, 1750–1925* "very useful" the first time she displayed this title. She switched to the Pinstripe System, conducted a Pinstripe-system search for this same query and rated the same title "not useful" the second time she displayed it. Since we did not interview users following their searches of the experimental online catalog, we did not have the opportunity to ask users why they changed their relevance assessments from a positive to a negative rating. For the purpose of the failure analysis, we counted searches in which there was a conflicting relevance assessment as successful searches

#### 10.3.2.4 Database Failure

Three searches failed to retrieve useful titles because of database failure. The queries in these three searches were:

- jazz
- sit ins

- black church history

The former query was entered into the Blue System only. The two latter queries were entered into the Pinstripe System only. In all three cases, users retrieved titles in response to their queries but displayed titles were much more specific than the original user query. For example, the one title retrieved by the query "black church history" in a Pinstripe-system search was *Quest for equality: the life and writings of Mary Elizabeth Church Terrell, 1863–1954.* The Pinstripe System retrieved this title in a keyword-in-record search because of the occurrence of the word "church" in the title and the words "black" and "history" in the series title, viz. "Black women in United States history."

#### 10.3.2.5 Large Retrievals

Of searches that failed to retrieve useful titles in one of the two systems users searched, only one search retrieved too many titles. This was a Pinstripe-system search for "slavery" that retrieved 390 titles. The user responded to the display of the first of 390 titles by giving the title a "not useful" rating and terminating his search of the Pinstripe System.

#### 10.3.2.6 Vocabulary of User Queries

Searches placed in this unsuccessful search category involved user queries that were phrased in certain ways that prevented the experimental online catalogs from producing retrievals. Of searches that failed to retrieve useful titles in one of the two systems users searched, only one search was so phrased. This was a Blue-system search for "roaring twenties." The system exhausted the search trees without making a match. If the query were reduced to a single word — "twenties" — the system would retrieve titles such as *Into the twenties: the United States from Armistice to normalcy* and *Only yesterday: an informal history of the nineteen-twenties.* A one-word query such as "twenties" is not likely to retrieve many false drops in a small database limited to library materials on American history. Such a short query would be likely to retrieve many false drops in large library databases that feature encyclopedic coverage of recorded knowledge. To guide users to additional library material, systems should feature user-feedback mechanisms that prompt users for relevance assessments and use the subject headings assigned to useful titles to find such material.

#### 10.3.2.7 Subject Searching Approach Failures

The experimental online catalog contained information on topics that interested users. However, its failure to retrieve the desired information was attributed to the particular subject searching approach selected by the system. Search approach failures

were more common in the Pinstripe System than in the Blue System. The failure of fifteen searches was attributed to search approaches in the former system whereas the failure of two searches was attributed to search approaches in the latter system.

**Table 10.27. Subject Searching Approach Failures in the Pinstripe System**

| User query | Explanation |
|---|---|
| vietnam war | The Pinstripe System fails to retrieve titles through the keyword-in-subdivided-heading search. (Blue System retrieves 8 titles in a title-keyword search and the user rates 5 titles "very useful.") |
| calhoun (last name), john (first name), state rights (topic) | The Pinstripe System fails to retrieve titles through the keyword-in-record search. (Blue System ignores the topic element, places the user in an alphabetical list of subject headings where he chooses "Calhoun, John C. (John Caldwell), 1782–1850," displays three of twelve retrieved titles, and gives two of three titles "very useful" ratings.) |
| abolitionism | The Pinstripe System fails to retrieve titles through the keyword-in-subdivided-heading search. (Blue System retrieves eight titles in a title-keyword search and rates one title "very useful.") |

Table 10.27 lists three user queries and explanations for subject search approach failures in the Pinstripe System. Failures in the Pinstripe System were usually connected with the failure of the keyword-in-subdivided-heading search to retrieve subject headings bearing the words in user-entered queries.

**Table 10.28. Subject Searching Approach Failures in the Blue System**

| User query | Explanation |
|---|---|
| women in the westward movement | The Blue System informs the user that it is unable to retrieve titles. (In a Pinstripe-system search for the same topic, the Pinstripe System responds with the alphabetical approach and places the user at the subject heading "Women pioneers." The user finds useful titles bearing this heading.) |
| american indian religious freedom act | The Blue System exhausts the search trees without making a match. (Pinstripe System responds with the alphabetical approach and places the user in an alphabetical list of subject headings at the cross reference "American Indians." User selects subject headings beginning with the word "Indians" and rates retrieved titles "very useful.") |

Search approach failures were not as common in the Blue System. The two search approach failures that occurred in the Blue System are described in Table 10.28.

#### 10.3.2.8 Navigational Problems

The failure of two Blue-system searches was attributed to navigational problems (Table 10.29).

**Table 10.29. Navigational Problems**

| User query | Explanation |
|---|---|
| marxism | The Blue System informs the user that this term is a cross reference for "Communism." An exact search for "communism" results in enabled options for displaying general works, broader terms, narrower terms, notes, place and form subdivisions, and for expanding the search. The user displays titles connected with "Communism" and a narrower term "Marxian economics" and rates them "not useful." The Blue-system user seems close to finding useful titles but may be having difficulty navigating exact-approach options to find them. |
| pullman strike | Blue System responds to this query with an alphabetical list of subject headings bearing the cross reference "Pullman Strike, 1894." User selects the reference and the system responds by telling the user that "Chicago Strike, 1894" is the term used for this topic. System responds with the exact-search main menu with options for displaying titles and for expanding the search. User chooses latter option and the system continues searching using the original query words "pullman strike." The system fails to retrieve titles with these words because they occur in a cross reference, not in a bibliographic record. |

#### 10.3.2.9 Failure of One System Revisited

Searches that failed to retrieve useful retrievals in one system numbered 47 — about one-quarter of the searches performed in the experimental online catalog. A failure analysis of these searches demonstrated the following problems that adversely affected the retrieval of potentially useful information from the experimental online catalogs: (1) user perseverance, (2) specificity that was not represented in user queries, (3) conflicting relevance assessments, (4) database failures, (5) large retrievals, (6) user query vocabulary, (7) system failures, and (8) navigational problems.

Table 10.30 reclassifies unsuccessful searches in the Blue and Pinstripe Searches. The reclassification is limited to the 47 unsuccessful searches conducted in the Blue or Pinstripe Systems that were featured in section 10.3.2.

**Table 10.30. Reclassified Searches (Single system, Earlham)**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Useful titles retrieved** | **No.** | **%** | **No.** | **%** |
| Successful searches† | 16 | 72.7 | 7 | 28.0 |
| Database failures | 1 | 4.6 | 2 | 8.0 |
| System failures* | 5 | 22.7 | 16 | 64.0 |
| Total | 22 | 100.0 | 25 | 100.0 |

†Barring perseverance, query specificity, conflicting relevance assessments.

*Including problems with large retrievals, subject searching approaches, user query vocabulary.

The failure analysis promoted several unsuccessful searches to successful searches. A little under three-quarters of Blue-system searches and a little over one-quarter of Pinstripe-system searches were successful in terms of guiding users to titles on the topics their queries describe. Database failures accounted for few failures. System failures were more plentiful in the Pinstripe System than in the Blue System.

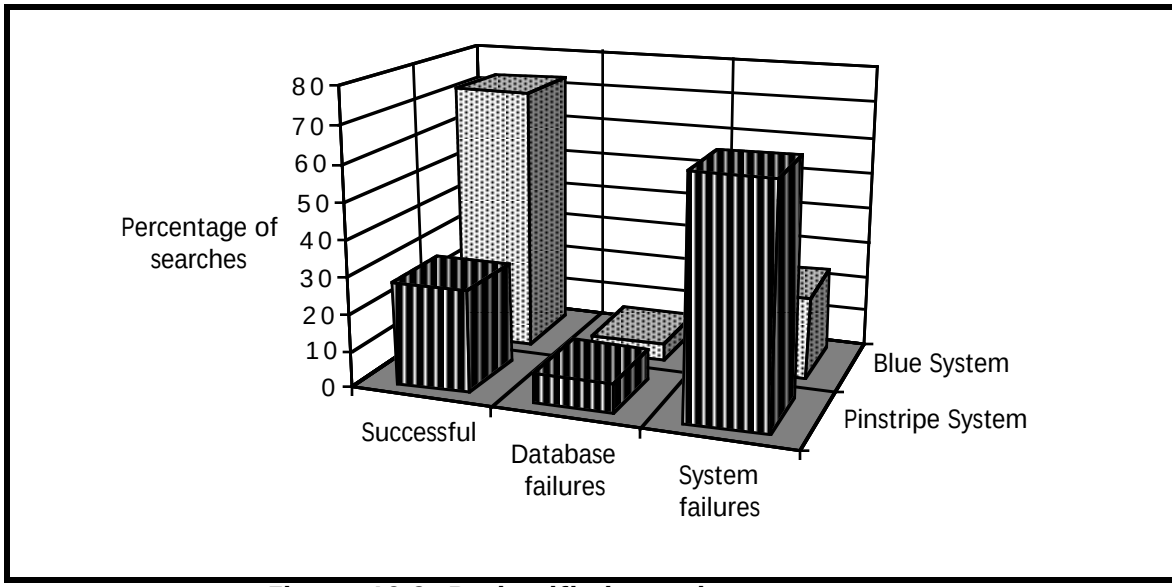


**Figure 10.9. Reclassified searches (single system, Earlham)**

Figure 10.9 is a graphical depiction of the statistics in Table 10.30. It shows obvious differences in the percentages of successful searches and system failures in Blue and Pinstripe Systems. Although sample sizes were quite low in this analysis, the percentage of successful searches in the Blue System was dramatically larger than the percentage of such searches in the Pinstripe System. Furthermore, the percentage of system failures

in the Pinstripe System was almost three times large as the percentage of system failures in the Blue System.

### 10.3.3 Revisiting Successful and Failed Searches

In the failure analysis of searches conducted by Earlham users, the researchers repeated user searches that failed to retrieve useful titles in an attempt to determine why they were unsuccessful in retrieving useful titles. In the process of repeating these searches, the failure analysis demonstrated that several searches might have been headed toward success but for various user actions that precluded the display of useful titles, e.g., perseverance, queries that were not indicative of user interests, conflicting relevance assessments. The failure analysis also demonstrated reasons for failure, e.g., design problems with a particular search approach or indexing technique, inability of a randomly-selected approach in the Pinstripe System to retrieve any titles in response to user queries. We can learn from the failure analysis in terms of designing subject searching approaches and reconfiguring the search trees. Chapter 12 makes recommendations for improving subject access in online catalogs based on failure analysis results.

Table 10.31 recasts the number and percentage of successful and unsuccessful searches based on the failure analysis of end-user searches at Earlham College.

About three-quarters of Blue-system searches were successful; about half of Pinstripe-system searches were successful. A large percentage (18.3%) of Blue-system searches would have been successful but users did not exhibit sufficient perseverance to stay with the search and see it to its conclusion. This percentage was somewhat smaller in the Pinstripe System (8.1%).

Speculating on the reasons why many Blue-system users exhibited low levels of perseverance, we are quite sure that user perseverance in the Blue System was connected with amount of time the system required for users to browse the many options available in alphabetical and exact searches. These approaches allowed users to select listed subject headings, narrower terms, broader terms, or subdivided forms of selected subject headings. It took time for users to navigate between options and for the system to search for selected subject headings. In Chapter 11, we take a close look at the alphabetical and exact approaches, identify troublesome features of these approaches, and make recommendations for improving their design in future online systems.

**Table 10.31. Reclassified Successful and Unsuccessful Searches at Earlham**

| | Blue System | | Pinstripe System | |
|---|---|---|---|---|
| **Type of search** | **No.** | **%** | **No.** | **%** |
| **Successful searches:** | | | | |
| 1 useful titles retrieved | 46 | 44.2 | 37 | 37.8 |
| Retrieval of useful titles marred by perseverance | 19 | 18.3 | 8 | 8.1 |
| Retrieval of useful titles marred by query specificity | 7 | 6.7 | 3 | 3.1 |
| Retrieval of useful titles marred by conflicting relevance assessments | 3 | 2.9 | 0 | 0.0 |
| Subtotal | 75 | 72.1 | 48 | 49.0 |
| **Unsuccessful searches:** | | | | |
| Database failure | 15 | 14.5 | 16 | 16.3 |
| Search approach failures | 2 | 1.9 | 23 | 23.4 |
| Large retrievals | 0 | 0.0 | 3 | 3.1 |
| Vocabulary of user queries | 10 | 9.6 | 8 | 8.2 |
| Navigational problems | 2 | 1.9 | 0 | 0.0 |
| Subtotal | 29 | 27.9 | 50 | 51.0 |
| Total | 104 | 100.0 | 98 | 100.0 |

Of the reasons for unsuccessful searches, database failure affected about the same percentages of searches conducted in the Blue and Pinstripe Systems. Vocabulary of user queries was also troublesome and led to a little under ten percent of search failures in both Blue and Pinstripe Systems. What was most striking about the percentages of search failures was the large percentage (23.4%) of failures for search approaches in the Pinstripe System. Only 1.9% of searches in the Blue System failed because of subject searching approaches. Most of the time, the Pinstripe System's keyword-in-subdivided-heading search was unable to retrieve titles bearing words in user queries. On occasion, its alphabetical approach placed users in an alphabetical neighborhood that was not fruitful for helping users find satisfactory subject headings for expressing their topics of interest.

The Pinstripe System's selection of a subject searching approach in response to a user-entered query was based on a random selection algorithm. In contrast, the Blue System's selection of a subject searching approach was governed by search trees. These data provide evidence to support the development of search trees in online catalogs. In a

catalog without search trees, a substantial percentage of end-user searches are bound to fail because systems do not assist end users in subject searching approach selection.

## 10.4 Chapter Summary

The approach used in this study to determine why the experimental ASTUTE online catalog was successful or unsuccessful in retrieving useful titles was a failure analysis. The project director, an expert searcher of information retrieval systems, repeated end-user searches that were successful and unsuccessful in retrieving useful titles to identify the reasons why they succeeded or failed. This section features the results of the failure analysis.

The failure analysis focused on searches in which users displayed no useful titles. This was a majority of the searches conducted in the Blue and Pinstripe Systems at UM-D and Earlham (see Tables 10.1 and 10.18) as users failed to display of useful titles in almost three-fifths of their experimental online catalog searches.

The failure analysis of these searches demonstrated the following problems that adversely affected the retrieval of potentially useful information from the experimental online catalogs: (1) user perseverance, (2) specificity that was not represented in user queries, (3) conflicting relevance assessments, (4) database failures, (5) large retrievals, (6) user query vocabulary, (7) search approach failures, (8) navigational problems, and (9) spelling problems. Sections 10.2 and 10.3 described the failure analysis of searches conducted by users at UM-D and Earlham, respectively. The discussion in these two sections was divided into two parts for searches that failed to produce useful titles in both experimental systems and searches that failed to produce useful titles in one of the two systems.

Explanations and examples of searches placed in categories for perseverance, query specificity, and conflicting relevance assessments provided evidence that these searches were really successful searches in terms of leading users to titles on their topics of interest. With regard to user perseverance, the systems gave users ample opportunities for browsing terminology or displaying titles on their topics of interest but users terminated their searches. In searches marred by query specificity, users gave negative relevance assessments to the titles that seemed relevant to their queries. Their actions led us to conclude that the systems were successful in terms of retrieving titles on the topics users expressed in their queries but they were not successful in terms of retrieving titles that fulfilled the specifications of queries users had in their minds but did not make explicit in their queries. In the failure analysis, we also considered searches

bearing conflicting relevance assessments as successful searches because users had given the titles positive assessments in the course of their searches.

The reclassification of searches placed into categories for perseverance, query specificity, and conflicting relevance assessments from unsuccessful to successful searches resulted in success rate of about 72% for Blue-system searches at UM-D and Earlham and a success rate of between 49% and 59% for Pinstripe-system searches at UM-D and Earlham (see Tables 10.17 and 10.31). Clearly, the Blue System performed better than the Pinstripe System in terms of retrieving and displaying titles on the topics users made explicit in their queries.

Success in the Blue System, however, was marred by a large percentage of searches in which users did not persevere in terms of browsing subject terminology and displaying retrieved titles (sections 10.2.1.1, 10.2.2.1, 10.3.1.1, and 10.3.2.1). Almost ten percentage points separated the results of Blue- and Pinstripe-system searches for searches marred by perseverance (see Tables 10.17 and 10.31). Many Blue-system searches affected by perseverance featured controlled vocabulary search approaches. Navigational problems were also linked to user perseverance and controlled vocabulary searches (sections 10.2.1.8, 10.2.2.8, and 10.3.2.8). Of the several suggestions made to improve controlled vocabulary searches, the important ones were:

- Reducing the response time for searching selected terms and creating exact-search option menus.
- Reserving exact-search option menus for summaries of large retrievals and seeking other methods for summaries of one or a few retrievals.

Success in the Pinstripe System was negatively affected by the system's random selection algorithm for subject searching approaches. Between ten and twenty percentage points separated the results of Blue- and Pinstripe-system searches for search approach failures (see Tables 10.17 and 10.31). The keyword-in-subdivided-heading search was usually to blame for search approach failures; however, the alphabetical approach had a share of the failures. Subject searching approach failures were rare in the Blue System, accounting for only 1.3% of Blue-system searches. Search approach failures accounted for the failure of 14.3% of Pinstripe-system searches. The Blue System was guided in its selection of a subject searching approach by the search trees. Thus, the data and analyses in this project provided evidence that search trees improved the search performance of subject searchers at the online catalog in terms of guiding them to subject headings and titles that described the topics they made explicit in their queries.

## References

Lancaster, F. Wilfred, and Amy J. Warner. 1993. *Information retrieval today.* Arlington, Va.: Information Resources Press.

Lynch, Clifford. 1990. "Large database and multiple-database problems in online catalogs. In *Annual Review of OCLC Research, July 1989–1990,* 51–5. Dublin, Ohio: OCLC.

Markey, Karen, and Anh N. Demeyer. 1986. *Dewey Decimal Classification Online Project*. Dublin, Ohio: OCLC. OCLC Research Report Series OCLC/OPR/RR-86/1.

Wiberley, Stephen E. and Robert Allen Daugherty. 1988. "Users' persistence in scanning lists of references." *College & Research Libraries* 49, 2 (March): 149–56.

Wiberley, Stephen E., Robert Allen Daugherty, and James A. Danowski. 1993. *User persistence in displaying LUIS postings: A report to the Council on Library Resources*. Chicago: University of Illinois at Chicago. 23 July.