6 Methods Used to Test the New Subject Access Design

6.1 Introduction

The ASTUTE experimental online catalog was actually composed of two online catalogs: (1) the Blue Test System, in which search trees governed the system's selection of a subject searching capability, and (2) the Pinstripe Test System, in which the system selected a subject searching capability randomly. These systems were purposely designed to be very much alike to focus the attention of library patrons and staff on the retrieval of useful information in response to their queries. The Blue and Pinstripe Systems had virtually the same interfaces, and they accessed the same bibliographic and authority databases. Except for the Blue System's enhancement with the search trees, the two systems and their capabilities were the same.

In online retrieval tests, the Blue and Pinstripe Systems encouraged library patrons and staff to perform online subject searches in both systems. In introductory screens, the systems informed users that they were participating in an experiment and allowed them to graciously exit the system if they did not want to take part. Prior to conducting subject searches, the systems asked participants to respond to three pre-search questions that collected demographic information. During their searches, the systems recorded user activity and user relevance assessments for displayed bibliographic records to a time-stamped transaction log. The systems asked participants to answer eleven questions in a post-search questionnaire that compared the performance of the Blue and Pinstripe Test Systems. Thus, the retrieval tests yielded quantitative comparative data about the systems' capabilities and effectiveness, and about users' system preference. After completing the post-search questionnaire, library staff were asked additional questions that prompted them to make comparisons between the two systems with respect to the systems' ease of use, subject searching capabilities, and performance. These qualitative data were combined with quantitative data from patron searches to increase our understanding of search tree performance and enhancement.

In this chapter, the experimental design of the Comparison Search Experiment is described. Data collection procedures, online questionnaires (administered to patrons and staff), and offline questionnaires (administered to staff) are highlighted.

6.2Previous Studies Comparing Systems

The comparative method used in this project is based on the Comparison Search Experiment devised by Siegel et al. (1983) in their evaluation of two prototype online catalogs at the National Library of Medicine (NLM). The goal of Siegel's Comparison Search Experiment was to reproduce realistic search conditions in which library patrons performed a search on a topic of their own choosing on both systems. The researchers alternated the first system patrons used to minimize any transfer effect, that is, bias caused by patrons transferring their knowledge of the search and of searching in the first system to the second system. The researchers observed patrons and decided when they should switch systems. After completing their search of the second system, the researchers asked patrons questions about their experiences using the two systems, satisfaction with search results, and system preference. Siegel et al. used a second approach — the Sample Search Experiment — to control for potentially confounding variables. Siegel et al. used the results of both experiments to select an online catalog for use by library patrons and reference staff at NLM.

Several researchers have based their evaluation of two or more online catalogs on Siegel's Comparison or Sample Search Experiment. The project director used Siegel's methods to evaluate two experimental online catalogs, one with enhanced subject searching functionality using the Dewey Decimal Classification (DDC) and one without the enhancement (Markey and Demeyer 1986, 109–18). She expanded on Siegel's approach by asking patrons to assess the relevance of retrieved bibliographic records and used these assessments to compare the performance of the two experimental online catalogs with respect to precision and estimated recall. Markey and Demeyer used the results of their comparative study to evaluate the retrieval effectiveness of the DDC in subject searches.

Jones (1988) enlisted the Comparison Search Experiment to compare the performance of the operational LIBERTAS online catalog and the experimental Okapi online catalog. In the two separate studies by Markey and Demeyer (1986) and Jones (1988), interviewers approached prospective online catalog users, asked them to participate in the study, asked them a few questions prior to their search, observed their online searches in two systems, and asked them much longer lists of questions after the search about their searching experiences, satisfaction with search results, and system preference. Jones collected user relevance assessments and used the data to establish user assessments of system performance and user attitudes to Okapi's recall-improvement devices. Jones also compared user preferences between specific interaction features connected with ease of use, bibliographic format, browsing, menu design, and learning and relearning how to use the system.

Since its inception in 1982, the Okapi experimental online catalog has been under continual enhancement and redesign. Enhancement and redesign has been based on the findings of many evaluations including Jones' 1988 evaluation that enlisted the Comparison Search Experiment.

In 1987, the Okapi designers evaluated three different Okapi representations that were alike in all ways except for the following functionality: (1) the experimental (EXP) system featured "strong" stemming, a phrase dictionary, and some automatic cross-referencing, (2) the control (CTL) system featured "weak" stemming and spelling standardization, and (3) the third system (OSTEM) performed no stemming or spelling normalization (Walker and Jones 1987). The EXP and CTL systems were available for patron searching on different floors of the library. Although interviewers were present when patrons searched the EXP or CTL systems, their interaction with patrons was minimal. They recorded start and finish times and asked patrons four brief questions about themselves and their search results. Okapi logged patron searches to transaction logs. The researchers performed a failure analysis of EXP and CTL searches using questionnaire responses and logged searches. They also searched OSTEM for interesting searches and compared OSTEM results with patron searches in EXP or CTL. The goal of the Okapi designers in this particular evaluation was to assess the performance of several recall-improvement devices, i.e., automatic word stemming, synonym and cross-reference tables, soundex keys for matching personal names, and an n-gram technique for approximate word matching.

Based on 1987 and 1988 evaluations, the Okapi designers set to work at new representations of this experimental system. In 1990, they evaluated three different Okapi representations: (1) the DUMB system that was the same as the EXP system in the 1987 evaluation, (2) the QE (i.e., query expansion) system that was the same as "dumb" except for a "look for books similar to those already chosen" feature, and (3) the FULL system that was the same as QE except for a shelf-browsing feature (Walker and De Vere 1990). The Okapi researchers devised a method that "lay somewhere between the Comparison and Sample search experiments used in the NLM experiment" (Walker and De Vere 1990, 46). They recruited respondents, asked them a few questions about themselves before their searches, gave them a choice of questions to search on two of the three systems, recorded searches to transaction logs, asked them more questions about their searching experiences, results, and system preference after the search, and obtained relevance assessments for displayed bibliographic records from judges such as librarians and subject specialists. The goal of the evaluation was to compare the three systems with regard to effectiveness, efficiency, and user acceptability.

Hildreth (1993) employed an approach similar to the 1990 Okapi evaluation to evaluate the usability and retrieval performance of a navigation approach to subject searching in online catalogs. He evaluated three different experimental online catalogs: (1) a control catalog bearing no navigation features, (2) a catalog that supported related-record navigation based on title words, and (3) a catalog that supported related-record navigation based on subject headings only. He recruited respondents, asked them a few questions about themselves before their search, gave them a search task to perform in the three systems, recorded searches to transaction logs, asked them more questions about their searching experiences, results, and system preference after the search, and compared their search results with those obtained by expert searchers.

6.3Online Questionnaire Administration in Catalog Use Studies

Comparative approaches to system evaluation have used human intermediaries to collect information about searchers and their searching experiences. Such approaches are expensive and time-consuming for several reasons. Human intermediaries are needed to recruit searchers for participation in the system evaluation. An interviewer must be present to monitor end-user searches, ask questions, and record answers. If searchers are not recruited and scheduled ahead of time, interviewers could spend a considerable amount of time recruiting searchers themselves and explaining the study to them. Travel expenses for interviewers can be considerable for extended data-collection periods at off-campus sites. For these reasons, the project director used a comparative approach to system evaluation in this project that did not use interviewers or human intermediaries at all. The experimental online catalog performed and/or recorded all activity connected with data collection — searcher recruitment, questionnaire administration, search logging, and logging of searcher relevance assessments.

The comparative approach used in this study was innovative and experimental. Although online questionnaire administration was used in an early online catalog use study, delegating complete control of retrieval tests to the experimental online catalog had not been done before.

The early study of online catalog use that enlisted online questionnaire administration was conducted by researchers at the University of California’s Division of Library

Automation (UC/DLA) who administered online questionnaires to users of the MELVYL online catalog in the Council on Library Resources-sponsored nationwide study of online catalogs (University of California 1982). Despite the success of UC/DLA's online questionnaire administration, this approach has been rarely used in online catalog use studies.

In a recent Department of Education-sponsored study, a research team at Rutgers University recorded searches and online questionnaire responses administered in online pre- and post-search questionnaires to transaction logs (Belkin et al. 1990). The purpose of the Rutgers study was to discover the goals, tasks, and behaviors of library and online catalog users with a view to using these data for the specification of online catalog design principles.

A research team at City University collected data on online catalog use using methods compatible with those enlisted by the Rutgers team (Hancock-Beaulieu, McKenzie, and Irving 1990). The City University team developed Olive, a front-end microcomputer program that recorded searches and pre-, in-, and post-search online questionnaire responses. Olive also featured search playback and printing capabilities. Although the City University team recorded searches and questionnaire responses made by users of two different online catalogs (i.e., CLSI and LIBERTAS), the team's aim was to build on previous research findings regarding predictors and categories of user behavior.

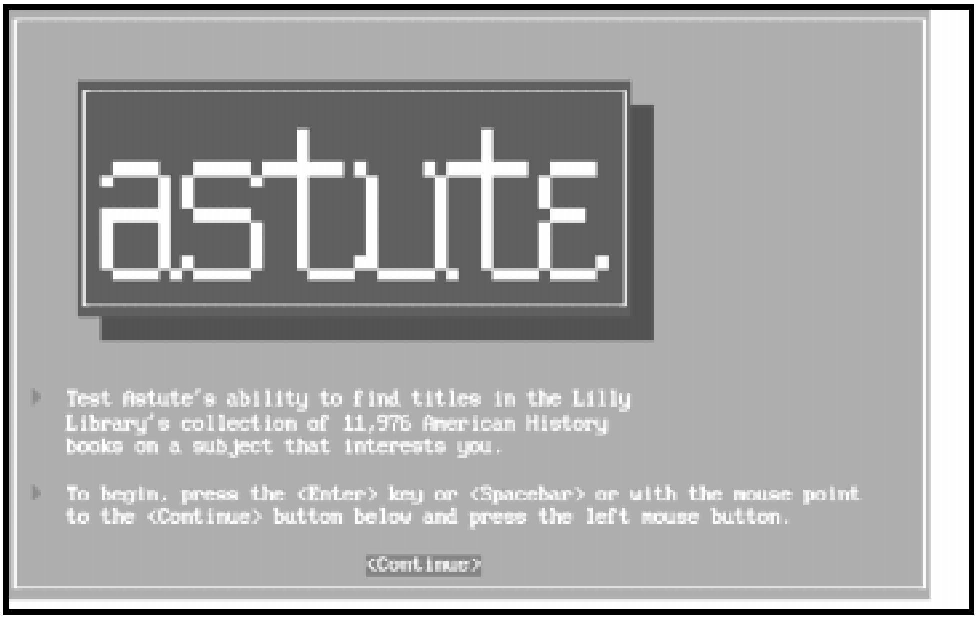
6.4Comparison Search Experiment (Library Patrons)

6.4.1 Recruiting Searchers

In this study, the researchers transported and assembled a Gateway microcomputer bearing ASTUTE to the two data collection sites — Mardigian Library at the University of Michigan-Dearborn and Lilly Library at Earlham College. The microcomputer was dedicated to use of the ASTUTE experimental online catalog. At UM-Dearborn, ASTUTE was located in a quiet study area of the library which was also near computer science, engineering, and technology stacks. Thus, ASTUTE searchers would not have to go very far to access the library material they retrieved in their searches of the experimental online catalog. At Earlham College, ASTUTE was located in the reference area of the library near the library's MARCIVE™ CD/ROM-based online catalog and other CD/ROM reference sources. Lilly Library reference staff were also nearby and directed patrons to ASTUTE when they felt patrons would find useful material in the system. At both libraries, signs were placed near ASTUTE to attract library patrons to use the system.

The ASTUTE experimental online catalog performed recruiting functions on its own. The presence of the microcomputer equipment in a public place in the library, posted written signs, and two alternating introductory screen savers attracted potential respondents. Reading subsequent introductory screens also piqued the interest or curiosity of some individuals to take part in the system evaluation. ASTUTE was installed in Mardigian and Lilly Libraries for several weeks at a time. If interested or curious library users did not have time to use the system the first time they saw it, they might have used it on subsequent library visits when they had more time.

ASTUTE featured five introductory screens. The first two introductory screens also functioned as screen savers to avoid the image of one of the screens burning into the monitor. Figure 6.1 shows the first introductory screen that invited users to take part in the evaluation of the system.

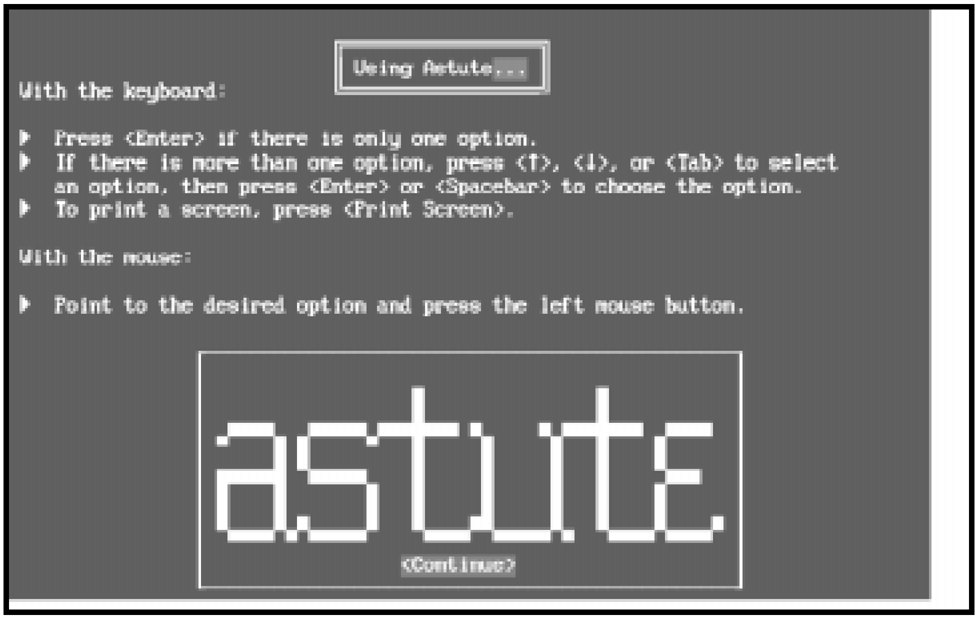


Test Astute's ability to find titles in the Lilly Library's collection of 11,976 American History books on a subject that interests you.

To begin, press the <Enter> key or <Spacebar> or with the mouse point to the <Continue> button below and press the left mouse button.

Figure 6.1 First introductory screen

Figure 6.2 shows the second introductory screen that told users how to operate the keyboard and mouse, make selections, and print screens.


Using Astute...

With the keyboard:

- ▶ Press <Enter> if there is only one option.
- ▶ If there is more than one option, press <↑>, <↓>, or <Tab> to select an option, then press <Enter> or <Spacebar> to choose the option.
- ▶ To print a screen, press <Print Screen>.

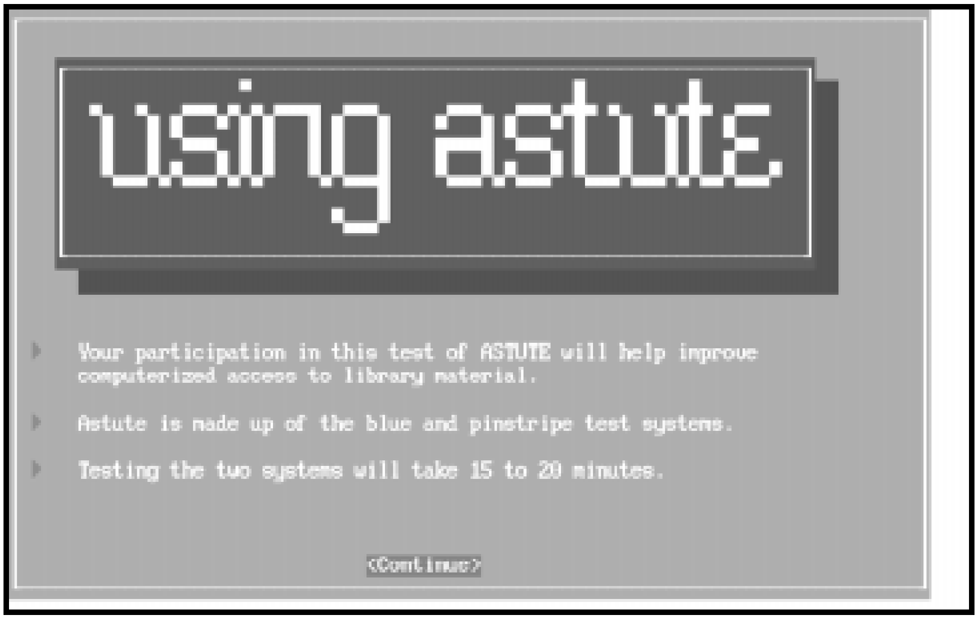
With the mouse:

- ▶ Point to the desired option and press the left mouse button.

Figure 6.2 Second introductory screen

If no one was using ASTUTE, the system alternated between displaying the first two introductory screens in figures 6.1. and 6.2.

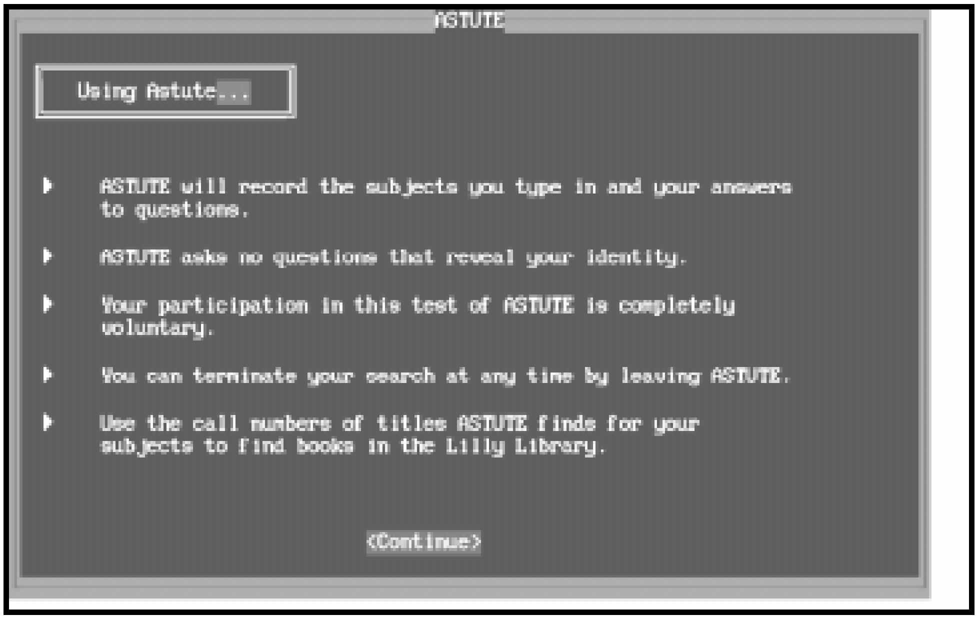
When users clicked on the <Continue> button in the first two introductory screens, ASTUTE responded with a series of three additional introductory screens that explained their participation in the test of the system. Prospective respondents were encouraged to read these three screens to find out about the system's purpose, subject coverage, and capabilities. These screens also invited users to conduct an online search on a topic of their own choosing in the system. ASTUTE told users it was logging their searches and responses to questions. Library users were entirely on their own to read screens, conduct searches, and answer questions. Figures 6.3–6.5 show the series of three additional introductory screens.



using astute

- Your participation in this test of ASTUTE will help improve computerized access to library material.
- Astute is made up of the blue and pinstripe test systems.
- Testing the two systems will take 15 to 20 minutes.

Figure 6.3. Third introductory screen

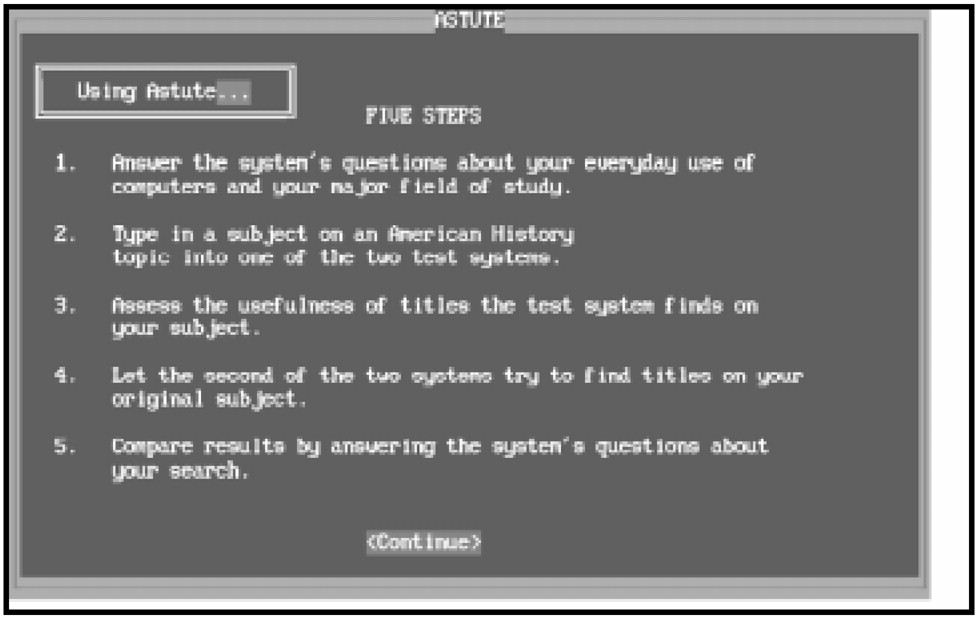


ASTUTE

Using Astute...

- ASTUTE will record the subjects you type in and your answers to questions.
- ASTUTE asks no questions that reveal your identity.
- Your participation in this test of ASTUTE is completely voluntary.
- You can terminate your search at any time by leaving ASTUTE.
- Use the call numbers of titles ASTUTE finds for your subjects to find books in the Lilly Library.

Figure 6.4. Fourth introductory screen



FIVE STEPS

1. Answer the systen's questions about your everyday use of computers and your major field of study.

2. Type in a subject on an American History topic into one of the two test systems.

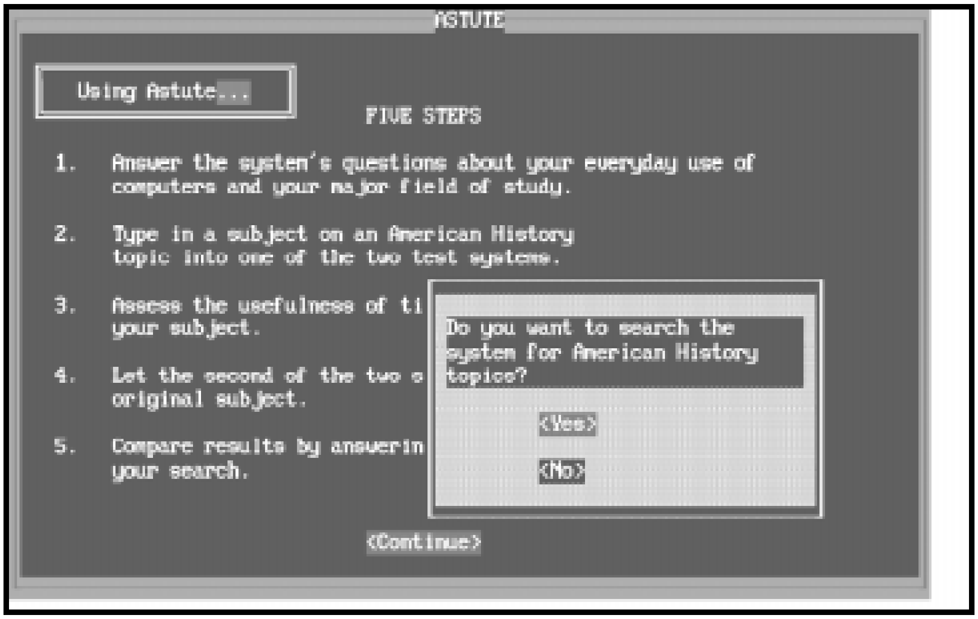
3. Assess the usefulness of titles the test system finds on your subject.

4. Let the second of the two systems try to find titles on your original subject.

5. Compare results by answering the systen's questions about your search.

Figure 6.5. Fifth introductory screen

When users clicked on the <Continue> button in the fifth introductory screen, ASTUTE responded with a pop-up window asking users whether they wanted to search the system for engineering or computer science topics (at UM-Dearborn) or American history topics (at Earlham). The answer to this question gave users who did not want to search ASTUTE a graceful way of exiting the system. That is, the system took users who clicked on the <No> button back to the introductory screen. Users could also terminate their use of the system at any time by just walking away. The system responded to users who clicked on the <Yes> button with the first of three pre-search questions. Figure 6.6 shows the pop-up screen that is the system's response to UM-Dearborn users who clicked on the <Continue> button in the fifth introductory screen.



ASTUTE

Using Astute...

FIVE STEPS

1. Answer the system's questions about your everyday use of computers and your major field of study.

2. Type in a subject on an American History topic into one of the two test systems.

3. Assess the usefulness of ti your subject.

4. Let the second of the two o original subject.

5. Compare results by answerin your search.

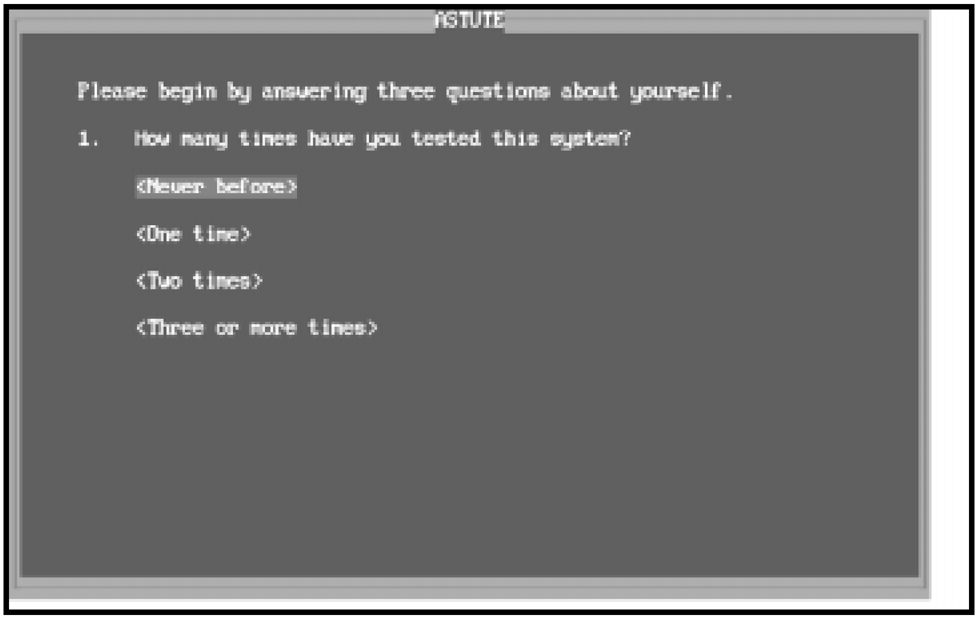
Do you want to search the system for American History topics?

Figure 6.6. Pop-up window with question on test participation

6.4.2 Administering the Pre-search Questionnaire

When users clicked on the <Yes> button in response to the pop-up window's question on searching for topics in a particular subject, ASTUTE began recording all subsequent user actions and system responses to a transaction log.

The pre-search questionnaire contained three questions. Figures 6.7–6.9 show these three questions. Users could use the mouse to highlight and click on a response category or they could use the arrow keys on the keyboard to highlight the desired response category and press on the <Enter> key to select it. The first pre-search question asked users how many times they had used ASTUTE (figure 6.7).



Please begin by answering three questions about yourself.

1. How many times have you tested this system?

<Never before>

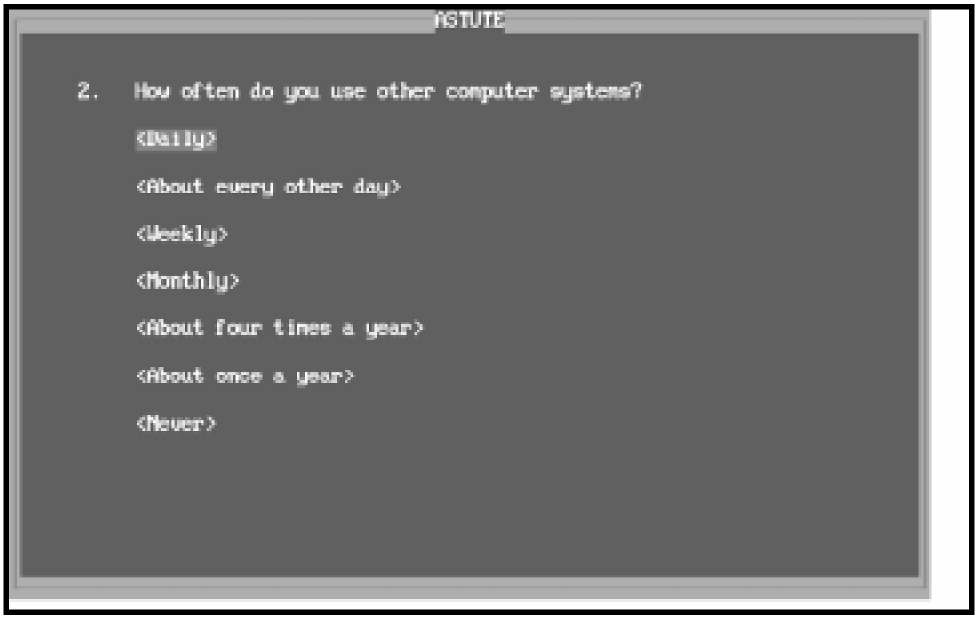
<One time>

<Two times>

<Three or more times>

Figure 6.7. First pre-search question on previous ASTUTE use

The second pre-search question asked users how often they used computer systems besides ASTUTE (figure 6.8).



ASTUTE

2. How often do you use other computer systems?

<Daily>

<About every other day>

<Weekly>

<Monthly>

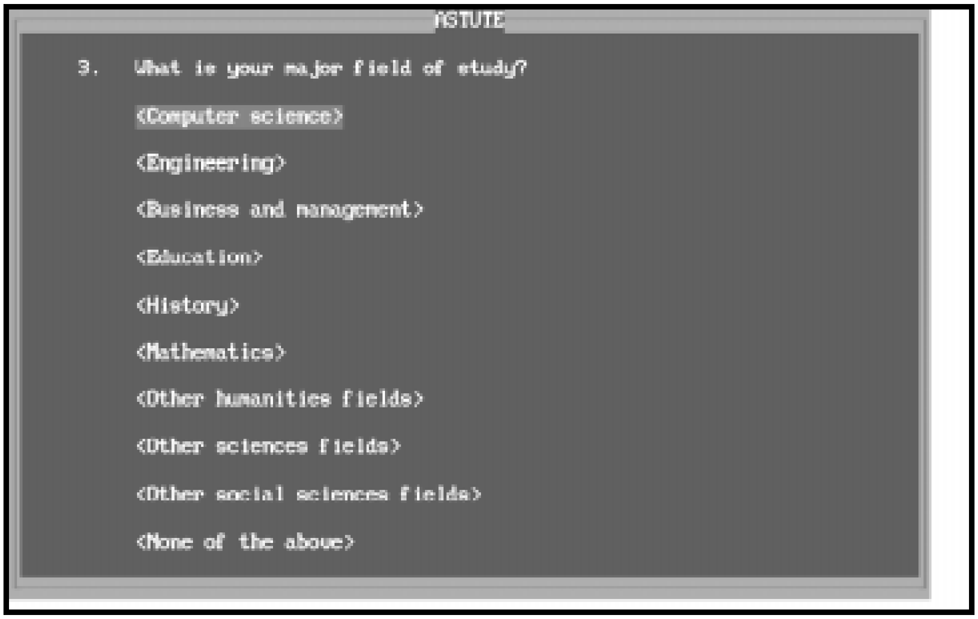
<About four times a year>

<About once a year>

<Never>

Figure 6.8. Second pre-search question on using other computer systems

The third pre-search question asked users to identify their major field of study (figure 6.9).



ASTUTE

3. What is your major field of study?

<Computer science>

<Engineering>

<Business and management>

<Education>

<History>

<Mathematics>

<Other humanities fields>

<Other sciences fields>

<Other social sciences fields>

<None of the above>

Figure 6.9. Third pre-search question on major field of study

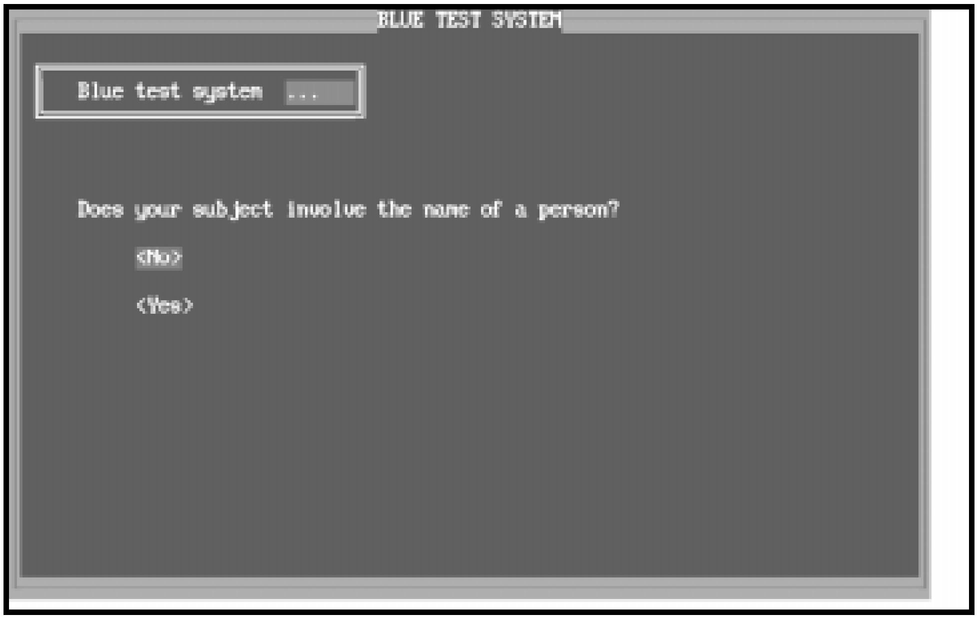
When ASTUTE began writing user actions and system responses to a transaction log, it activated a timeout function to determine whether users had terminated their search by walking away from the system. If ASTUTE recorded no user activity for four minutes, the system displayed a pop-up window asking the user to click on a <Continue> button to continue searching. If ASTUTE recorded no user activity within thirty seconds, the system closed the transaction log for the particular search administration, prepared to open the log for the next search administration, and displayed the first introductory screen (figure 6.1).

6.4.3 Searching the Experimental Online Catalog

The Comparison Search Experiment was designed to be as realistic as possible in that library patrons were searching the experimental online catalog with their own search topics and assessing the usefulness of retrieved items. Furthermore, transaction logging and online administration of pre- and post-search questionnaires made the data collection process as unobtrusive as possible.

When users clicked on <Yes> to the question in the pop-up window on the fifth introductory screen (figure 6.6), ASTUTE began logging user activity and system responses. It also assigned each participant to one of two odd/even experimental conditions. Odd-numbered participants searched the Blue Test System first, and even-numbered participants searched the Pinstripe Test System first.

Following the third pre-search question, the experimental online catalog asked users whether their query involved the name of a person. The Blue System used answers to this question to select between search trees for subjects generally and search trees for personal names as subjects. Figure 6.10 shows this question on personal names in subject queries.

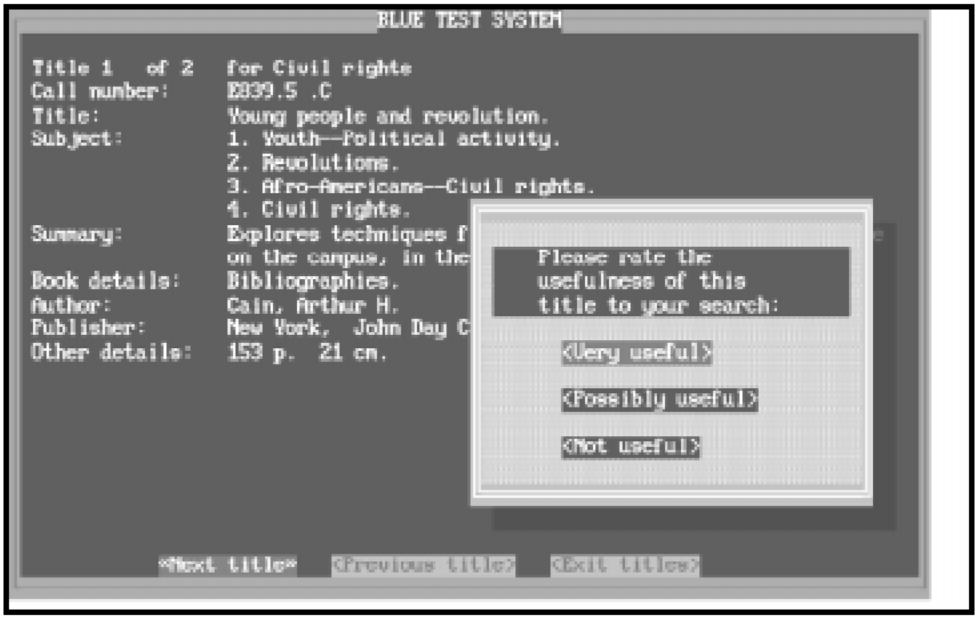


Does your subject involve the name of a person?

Figure 6.10. Question on personal names in subject queries

The Blue or Pinstripe System then asked users to enter their subject queries. (Chapter 5 details searching in the Blue and Pinstripe Systems.) As long as participants did not make a move to start a new search or enter a new query, they could search the Blue or Pinstripe System for as long as they wanted. When they made such a move, the Blue System (for odd-numbered participants) switched and automatically conducted a search for the original user-entered query in the Pinstripe System; the Pinstripe System (for even-numbered participants) switched and automatically conducted a search for the original user-entered query in the Blue System. As long as participants did not make a move to start a new search or enter a new query, they could search the second system for as long as they wanted. When they made such a move, the system initiated the post-search questionnaire. Both Blue and Pinstripe Systems recorded all user activity and system responses to transaction logs.

When users displayed bibliographic records, both Blue and Pinstripe Systems asked users to assess their usefulness. In figure 6.11, the Blue System displays a bibliographic record retrieved in a search for “civil rights” bearing a pop-up window that gives users three categories for rating usefulness: (1) 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. Both Blue and Pinstripe Test Systems recorded user assessments of retrieved records to transaction logs.



```
BLUE TEST SYSTEM

Title 1   of 2   for Civil rights
Call number:     E839.5 .C
Title:           Young people and revolution.
Subject:         1. Youth--Political activity.
                 2. Revolutions.
                 3. Afro-Americans--Civil rights.
                 4. Civil rights.
Summary:         Explores techniques f
                 on the campus, in the
Book details:    Bibliographies.
Author:          Cain, Arthur H.
Publisher:       New York,  John Day C
Other details:   153 p.  21 cm.
```

Please rate the usefulness of this title to your search:

Figure 6.11. Relevance assessment categories

6.4.4 Administering the Post-search Questionnaire

The post-search questionnaire contained eleven questions. Appendix C lists these questions. Questions 4–6 asked users to compare and assess the performance of the two systems in view of the useful titles they retrieved. Questions 7–12 compared subject searching capabilities and ease of use of the two systems and library computer catalogs generally. Questions 13–14 were general questions about the subjects users search in online catalogs. Users could use the mouse to highlight and click on a response category or they could use arrow keys on the keyboard to highlight desired response categories and press on the <Enter> key to select them. To answer question 12, users typed in numbers from 1 to 5; if they entered any character except a number from 1 to 5, the system displayed an error message and allowed users to re-enter a number. All questions were closed-ended except for question 14. If users responded “very interested” to a question about their interest in seeing ASTUTE’s capabilities extended to topics besides the ones in ASTUTE, the system asked users to type in the subjects that interested them. ASTUTE logged user responses to post-search questions to a transaction log.

6.5Comparison Search Experiment (Library Staff)

6.5.1 Background

In preparation for submitting the proposal to the Department of Education, the project director approached UM-Dearborn and Earlham College library staff about participating in the project. When the project director described search trees to UM-Dearborn staff, they bristled at the idea of an online catalog selecting a subject searching approach on its own in response to their subject queries. They suggested that the project director extend the Comparison Search Experiment to library reference and cataloging staff to obtain their reactions to the experimental system. Thus, the project director added a Search Experiment to the evaluation section of the Department of Education proposal that included library staff.

6.5.2 Recruiting Library Staff

The project director had informed UM-Dearborn and Earlham College library staff of their participation in the Comparison Search Experiment in written correspondence, proposal drafts, and face-to-face discussions before submitting the proposal to the Department of Education. In fact, the suggestion to include library staff in the search experiment came from UM-Dearborn library staff.

While the experimental online catalog was available for patron use at UM-Dearborn and Earlham, the project director sent staff letters reminding them of and describing the search experiment and asking for their participation. Library staff liaisons suggested the names of prospective participants, scheduled one-hour interviews with those who volunteered their participation, or the project director called them directly and scheduled them herself. The project director sent staff participants the cover letter and consent form given in Appendix D. The cover letter asks library staff to bring three to five subject queries in the area of computer science or engineering (at UM-Dearborn) or American history (at Earlham) to search. The queries could come from library patrons in the course of their reference work or could be subject queries in which they had a professional or personal interest. The cover letter also assured volunteers of confidentiality.

Interviews were scheduled on the day(s) immediately preceding and on the day that the ASTUTE project team disassembled the microcomputer equipment and transported the system back to Ann Arbor. This gave library staff several weeks to search ASTUTE on their own and saved the project team an extra trip to participating libraries.

6.5.3 Comparison Search Experiment Procedures

At the scheduled time, the project director and library staff member met where ASTUTE was installed and began the Comparison Search Experiment. Assuming the role of interviewer, the project director introduced herself to the staff member, gave a brief overview of the project, and described how the search experiment would proceed.

Library staff participants searched ASTUTE for one of the topics they brought. They responded to the three questions in the pre-search questionnaire, conducted their searches on the Blue and Pinstripe Systems, and responded to the eleven questions in the post-search questionnaire. The interviewer then asked participants nine open-ended questions about the impetus for their queries, difficulties searching the test systems, suggestions for system improvements, and system preference. She wrote their answers in spaces provided on the Post-search Interview Form with Library Staff (appendix E).

Participants then searched ASTUTE for a second time and responded to ASTUTE's pre- and post-search questionnaires. The interviewer asked participants the same nine open-ended questions about their searching experiences. Participants then searched ASTUTE for third and last time, responded to ASTUTE's pre- and post-search questionnaires, and the interviewer asked participants the same nine open-ended questions about their searching experiences. Following these nine questions, the interviewer asked four summary questions to elicit library staff comments on what they liked most and least about the experimental online catalog, their overall system preference, and any additional comments staff wanted to make.

Each library staff participant was randomly assigned to one of two odd/even experimental conditions. Odd-numbered participants first searched the Blue Test System, and even-numbered participants first searched the Pinstripe Test System. Occasionally, library staff entered queries on computer science/engineering topics (at UM-Dearborn) or American history topics (at Earlham) which were classified in different areas from the Library of Congress Classification areas selected for inclusion in ASTUTE databases. In such instances, the interviewer disregarded the query and allowed participants to conduct a search for another query.

Appendix E gives the questionnaire used by the interviewer in the Comparison Search Experiment with library staff. The questionnaire contains the interviewer's introductory statement, open-ended questions following each search in the Blue and Pinstripe Systems, and summary questions following searches.

6.6Recording Search Experiments to a Transaction Log

ASTUTE recorded individual administrations of the Comparison Search Experiment to a transaction log. The log consisted of up to four record levels for each patron or staff participant in the Comparison Search Experiment. Records at all four levels contained a unique identification number to enable the researchers to link the various records to one another. Level 1 records contained search experiment beginning and ending time stamps and answers to pre- and post-search questionnaires. Level 2 records documented user actions and system responses. Level 3 records contained relevance assessments for displayed bibliographic records. Level 4 records contained user responses to the only open-ended question in the questionnaire on user interest in seeing ASTUTE's capabilities extended to topics besides the ones in ASTUTE. Appendix F details the content and format of the four levels of transaction log records in the experimental online catalog.

6.7Timeout Function

The ASTUTE project team did not monitor ASTUTE use at the two participating libraries. Thus, users could come and go and use the system for as short or as long as they wanted. We added a timeout function to ASTUTE because there was nothing to stop users from walking away from the system at any time. FoxPro had a timeout function and we used this function to reset the program back to the initial screen after a period of time had lapsed without user activity. Originally, we set the timeout function to three minutes. During the pretest, the timeout function terminated several users' searches prematurely. We tried longer timeout periods and finally decided on a four-minute timeout period. If ASTUTE did not detect user activity for four minutes, it asked users the following question, "If you are still participating in the test, please press the <Enter> key or press the left mouse button to <Continue>." ASTUTE gave users thirty seconds to respond to this question. If users did not respond, the system recorded the time in level 1 transaction log records, terminated the search, and displayed the first introductory screen (see figure 6.1).

6.8Design of the Comparison Search Experiment

6.8.1 Test Administration

The project director chose the Comparison Search Experiment because it featured a partially controlled but authentic experiment. Respondents at the two participating libraries conducted online searches on a topic of their own choosing — sequentially —

on Blue and Pinstripe Test Systems, judged the relevance of retrieved records, and answered questions about their searching experiences in the two systems.

The Comparison Search Experiment that enlisted library patrons used the same design as Comparison Search Experiments conducted by library patrons in previous studies (Siegel et al. 1983; Markey and Demeyer 1986. 109 ff.). The Comparison Search Experiment involving library staff deviated from the experiment with library patrons as follows: (1) staff searched the Blue and Pinstripe Systems for three different topics, (2) staff took part in an extended post-search interview that featured open-ended answers to questions on their system use and experiences, and (3) staff conducted subject searches for topics that might have originated from sources other than their own personal experiences and interests.

Table 6.1 lists dependent and independent variables measured in the Comparison Search Experiments with library patrons and staff.

Table 6.1. Variables Measured in Comparison Search Experiment

Dependent Variables:
Number of retrieved bibliographic records
Number of displayed bibliographic records
Number of useful records displayed
Time spent searching Blue and Pinstripe Test Systems
System capabilities used
Ease of system and system capability use
Satisfaction with subject searches
System preference
Independent Variables:
Type of user (patron or staff)
System used (Blue or Pinstripe)
First system searched
Subject area (per library)
Subject searching approaches to be executed

Table 6.2 summarizes the procedures used in Comparison Search Experiments with library patrons and staff.

Table 6.2. Comparison Search Experiment Procedures Summary

Procedures	Patron experiment	Staff experiment
Source of search topic	Patron	Staff member
System searcher	Patron	Staff member
Source of relevance assessments	Patron	Staff member
Interviewer	Experimental system	Project director and experimental system
Estimated completion time	15–20 minutes	60 to 75 minutes

In the Comparison Search Experiment with patrons, two systems (Blue and Pinstripe) were studied at the two participating libraries. The system that was first searched by the patron was varied so that approximately half of the participants searched the Blue System first and the other half first searched the Pinstripe System. Since monitors were not present to encourage patrons to complete the entire Comparison Search Experiment, patrons could leave the experiment at any time. The researchers expected full and partial administrations of the Comparison Search Experiment. Table 6.3 summarizes the events completed in full and partial administrations of this experiment. Codes given in the bottom row of this table designate whether the collected data were considered full or partial usable administrations administrations.

Table 6.3. Full and Partial Administrations of the Comparison Search Experiment

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
Inconsistent post-search questionnaire responses				x	
Disposition of test administration	u, p	u, p	u, p	u, p	u, f

Key to disposition:

u = usable test
p = partial test
f = full test

There were four conditions under which usable, partial test administrations were possible. Full test administrations involved completion of pre- and post-search questionnaires, and searches in Blue and Pinstripe Systems. Post-search questionnaire responses in full test administrations had to be consistent with other data such as test system searches and relevance assessments.

Table 6.3 does not cite the characteristics of unusable test administrations. The project team expected several conditions under which unusable test administrations would occur, e.g., entry of queries on subjects other than computer science or engineering (at UM-Dearborn) or American history (at Earlham), occurrence of system errors. Several sections of chapter 7 detail characteristics of unusable test administrations.

Monitors were present to encourage library staff to complete the entire Comparison Search Experiment. Administrations of the Comparison Search Experiment included responses to pre- and post-search questionnaires, searches in Blue and Pinstripe Test Systems, and the staff member's answers to open-ended questions posed by the interviewer in the extended post-search interview. Except for the unfortunate occurrence of system errors that might have disrupted administrations of the Comparison Search Experiment with library staff, the project team expected to collect only full administrations of this experiment.

6.8.2 Quantitative Data Analysis

The experimental online catalog's transaction logging capability recorded subject searches, relevance assessments, and questionnaire responses. Subject searches provided quantitative data on the following:

- Length of user queries.
- Time spent searching.
- Frequency of system capability use.
- Number of titles retrieved.
- Number of titles displayed.

The combination of these data with user relevance assessments enabled the researchers to compare the performance of the Blue and Pinstripe Systems in terms of:

- Precision (defined as the proportion of retrieved documents that users rate "very useful").
- Estimated recall (defined as the fraction of titles users rate "very useful" in relation to all titles users rate "very useful" in the course of their searches for their topics of interest in both Blue and Pinstripe Systems). This

definition of estimated recall was adapted from a landmark study of information seeking and retrieving (Saracevic and Kantor 1988).

The researchers used t-test procedures to compare mean scores of precision and estimated recall in the Blue and Pinstripe Systems and to ascertain whether differences in mean scores were statistically significant (see chapter 9). We supplemented the results of statistical procedures with user responses to questions in the post-search questionnaire on ease of system and system capability use, user satisfaction with subject searches, and user system preference.

6.8.3 Qualitative Data Analysis

The failure analysis of Blue- and Pinstripe-system searches used a simple measure to characterize search success and failure. A successful search was one in which users retrieved at least one title that they rated "very useful." A failed search was one in which users failed to retrieve at least one title that they rated "very useful." The project director, an expert searcher of information retrieval systems, repeated successful and failed searches to determine the reasons why searches succeeded or failed.

The failure analysis featured an investigation of failed searches (see chapter 10). It also focused on failed and successful searches per subject searching approach type (see chapter 11). The latter identified problems with the implementation of a particular subject searching approach and needed improvements. It also demonstrated successful implementations of subject searching approaches that required little or no fine tuning.

Open-ended questions were limited to the extended post-search questionnaire administered by an interviewer to library staff. The interviewer asked staff nine questions about their subject searching experiences in the Blue and Pinstripe Test Systems after they had performed subject searches in both systems on a topic of their own choosing.

1. What was the impetus for the subject you searched?
2. What difficulties did you have searching for your subject in the Blue and Pinstripe Systems?
3. What improvements could be made to the Blue and/or Pinstripe Systems?
4. What differences did you notice between the ways in which the two systems responded to the subjects you typed in?
5. In what ways did the Blue and/or Pinstripe Systems give you new ideas for searching for subjects?

6. What differences did you notice between the titles the Blue and Pinstripe Systems found for your subject?
7. What did you like most and least about Blue and/or Pinstripe Test Systems?
8. What is your overall system preference and why?

Answers to these questions were not always on the tip of the staff member's tongue and had to be drawn out through the interviewer's probing. This probing helped to focus the staff member's attention on a specific Blue or Pinstripe System capability and to elicit their kudos about the system(s) as a whole or specific system capabilities and their suggestions for improving the Blue and Pinstripe Systems. (Probes are listed in open-ended questions in the extended post-search questionnaire in appendix E.) In two separate questions, staff were asked what they liked most and least about the Blue and Pinstripe Test Systems. They were also asked which system they preferred overall and why.

6.9Chapter Summary

This chapter featured a discussion of data collection procedures and instruments used in the evaluation of the performance of the experimental online catalog (see sections 6.4–6.5).

The researchers transported and assembled a Gateway microcomputer bearing ASTUTE to the two data collection sites — Mardigian Library at the University of Michigan-Dearborn and Lilly Library at Earlham College. The microcomputer was dedicated to use of the ASTUTE experimental online catalog. The ASTUTE experimental online catalog performed recruiting functions on its own. The presence of the microcomputer equipment in a public place in the library, posted written signs, and two alternating introductory screen savers attracted potential respondents.

This study enlisted the Comparison Search Experiment because it featured a partially controlled but authentic experiment (Siegel et al. 1983). Respondents at the two participating libraries conducted online searches on a topic of their own choosing — sequentially — on Blue and Pinstripe Systems, judged the relevance of retrieved records, and answered questions about their searching experiences in the two systems. When users displayed bibliographic records, both Blue and Pinstripe Systems asked users to assess their usefulness. Transaction logging and online administration of pre- and post-search questionnaires made the data collection process as unobtrusive as possible (see section 6.6 and Appendix F). Qualitative and quantitative analyses of collected data were described (see section 6.7).

References

Belkin, Nicholas J. et al. 1990. “Taking account of users tasks, goals and behavior for the design of online public access catalogs.” In *ASIS ’90; proceedings of the 53rd ASIS annual meeting; 4–8 November, 1990, Toronto*, edited by Diane Henderson, 69–79. Medford, N.J.: Learned Information.

Hancock-Beaulieu, Micheline, Lorna McKenzie, and Avril Irving. 1991. *Evaluative protocols for searching behaviour in online library catalogues*. London: British Library. British Library R&D Report 6031.

Hildreth, Charles R. 1993. *An evaluation of structured navigation for subject searching in online catalogues*. Ph.D. dissertation. Department of Information Science, The City University, London.

Jones, Richard M. 1988. *A comparative evaluation of two online public access catalogues*. London: British Library. British Library Research Paper 39.

Markey, Karen, and Anh N. Demeyer. 1986. *Dewey Decimal Classification online project: Evaluation of a library schedule and index integrated into the subject searching capabilities of an online catalog*. Dublin, Ohio: OCLC. OCLC Research Report OCLC/OPR/RR–86/1.

Saracevic, Tefko, and Paul Kantor. 1988. “A study of information seeking and retrieving.” *Journal of the American Society for Information Science* 39, 3: 197–216.

Siegel, Elliott R. et al. 1983. “Research strategy and methods used to conduct a comparative evaluation of two prototype online catalog systems.” In *National Online Meeting proceedings —1983*, 1983 April 12–14, New York, compiled by Martha E. Williams and Thomas H. Hogan, 503–11. Medford, N.J.: Learned Information.

University of California, Division of Library Automation and Library Research and Analysis Group. 1982. *Users look at online catalogs: results of a national survey of users and non-users of online public access catalogs: final report to the Council on Library Resources*. Berkeley, Calif.: University of California, November 16.

Walker, Stephen, and Richard M. Jones. 1987. *Improving subject retrieval in online catalogues; 1. Stemming, automatic spelling correction and cross-reference tables*. London: British Library. British Library Research Paper 24.

Walker, Stephen, and Rachel De Vere. 1990. *Improving subject retrieval in online catalogues; 2. Relevance feedback and query expansion*. London: British Library. British Library Research Paper 72.