The Effect of Interface Design on Item Selection in an Online Catalog

David H. Thomas

The effect that content and layout of bibliographic displays had on the ability of end-users to process catalog information was tested using a 2 x 2 factorial experimental design. Participants were asked to perform two related tasks during the course of the experiment. In the first task, they were asked to select a set of items that they would examine further for a hypothetical paper they must write, using a simulated online catalog to make their assessments of relevance. In the second task, they were asked to examine 20 bibliographic records, decide whether they would choose to examine these items further on the shelf, and identify the data elements that they used to formulate their relevance decision.

One group viewed bibliographic records on an interface similar to current online catalogs, one that used data labels and contained data elements commonly found. A second group viewed these records on an interface in which the labels had been removed, but the data elements were the same as those in the first. The third group viewed these records on a labeled display that included enhanced data elements on the brief record display. The final group viewed these records with the same brief record data elements as the third group, but with the labels removed, using ISBD and AACR2 punctuation standards.

For the first task, participants using enhanced brief screen interfaces viewed more brief screens and fewer full screens than their counterparts. Screen durations for the second 10 screens were found to have dropped from those of the first 10 screens. Statistical analyses comparing demographic variables to the screen frequencies uncovered many significant differences. Participants using the enhanced-content interfaces made fewer selections from index and full screens, and more selections from brief screens. For the second task, participants who used enhanced-content interfaces were able to make some sort of relevance judgment more frequently than those who used standard-content interfaces.

The widespread introduction of online public access catalogs into libraries over the last fifteen years has had a major impact on the way that end users utilize libraries and the resources that they contain. The development of online catalogs has transformed the primary locating tool in libraries—the card catalog—from a tool with limited means of exploitation to one with a much greater potential to help users find needed information objects. Research has suggested, however, that the online catalog brings with it a new set of problems for end users, as the added capabilities have also brought added complexities (e.g., Larson 1991a). Migrating bibliographic data from a print to an online environment has required reevaluation, redesign, and testing of many aspects of bibliographic information systems, from the methods of retrieval through the means of

David H. Thomas (sunfish62@yahoo.com) is an independent Database Design Consultant.

Manuscript received February 22, 2000; accepted for publication June 23, 2000.

presenting this information. While a large number of researchers have examined functional aspects of the system-user interaction with online catalogs (e.g., Bates 1989; Borgman 1986; and Harman 1992), few have investigated the effect that screen content and layout might have on the effectiveness of the system-user interaction. The screen content and layout of online catalog interfaces have been designed almost entirely based on expert opinion, with little use of empirical data on user preferences.

The purpose of this research was to contribute to the knowledge about online catalogs in order to help improve the effectiveness of those catalogs. The primary objective was to examine the communication process between the user and the catalog by gathering basic empirical data about user performance on different types of online catalogs. This was done by testing whether the content and layout of bibliographic displays in an online catalog influenced the effectiveness of the interface between online catalogs and end users.

The Information Seeking Process

In the information seeking process, an individual lacking knowledge to solve some need or problem *seeks* to remedy that lack by obtaining knowledge from some information resource (Buckland 1991). Such resources can range from other people (e.g., authorities) to museums or libraries. When the resource that the user consults is a library, the individual may employ the library's locating devices to identify potential informing instruments. The mechanisms in a library by which an individual can obtain needed information include consulting a reference librarian, reference sources, or the library's collections. In the latter two cases, the individual often must use the library's bibliographic catalog to locate the specific informing entities (e.g., almanacs, encyclopedias, journals, monographs, etc.) that are needed.

However, because the library serves the needs of many different users, its catalog contains data designed to address many potential problems or needs, and the process of locating useful informing entities for an individual's particular need can be difficult and frustrating. Individuals might seek information about particular attributes of a library's holdings, as when they look for items created on or after a certain date; they might require a particular bibliographic entity in the collection; or they might need information of a more general topical nature. Moreover, individuals must try to match their conceptualizations of the problems with the terminology used by the information system to represent those problems. These factors make a complicated process of bringing the information seeker together with the information sought.

This search process can be made easier or more difficult depending on many conditions, ranging from the ability of individuals to articulate their problems to the degree of helpfulness or obstructiveness of the library's information retrieval systems. However, the information system's failure might occur because the system fails to communicate clearly the content of the database. A failure of this type can occur because the user does not understand that the system has successfully retrieved bibliographic records to meet the user's need. A system that presents the user with incomplete or unclear data about the items it describes can mislead the user into thinking that the system does not contain items that meet the information need. For example, the data element that caused a particular record to be retrieved in response to an information seeker's request may be embedded in a part of that record that the seeker does not initially see. This might occur, for example, when an individual requests a specific work that is listed in a contents note and an added entry, neither of which the user sees on initial screens. In other words, how the system presents its findings to the user can influence the success of the communication process in an information retrieval exchange.

This problem underscores the fact that an information storage and retrieval system is, in fact, an information storage, retrieval, *and presentation* system. It is not enough for the system merely to store and retrieve information; to be useful, this information must be presented to the user in a manner that the individual can interpret meaningfully.

The adoption of cataloging standards, such as the *Anglo American Cataloguing Rules*, 2d. ed. (AACR2), represents a recognition of the user's need for standardized bibliographic descriptions. Without such standards, the user would have to interpret each library's catalog separately, which can hinder the communication process. Similarly, the International Standard Bibliographic Description (ISBD) represents an effort to establish guidelines for the arrangement of bibliographic data in a unit description, in part to make it possible for the user to interpret the contents of the description more easily.

With the advent of online catalogs, awareness of the user's need for standardized bibliographic description seemingly waned. This might have been due in part to the fact that the complexity of introducing computer technology in bibliographic applications distracted system developers, and it might have been due in part to the desire of developers to experiment with the power that the new technology offered. Moreover, the ability to manipulate the content of the online catalog more freely has given the library profession the ability to question the bibliographic display content guidelines previously established. The wide variation in current online catalog screen designs supports this view.

One negative result of this experimentation has been the loss of familiarity for the user, as each library's catalog can have different features and displays. The user is once again placed in the situation of having to learn how each library's catalog functions and how to interpret the output of

each of these systems. Different displays may contain the same kinds of data, while seemingly similar displays may have different information. Such confusion makes using online catalogs more complicated for the individual.

An examination of the library literature reveals that the content and layout of few—if any—online catalog displays have been rigorously tested to determine whether they meet users' needs, and that most catalogs have been designed without such input or testing. In human factors research into computer screen design, researchers have established broad guidelines for efficient and clear screen design—broad guidelines that library systems designers have used to inform online catalog designs. However, these researchers, too, have largely failed to base screen display design principles on empirically gathered data. This might be because the wide variation in applications makes such detailed study meaningless, inasmuch as the layout of data onscreen is so dependent on the purpose and content of the data. Library catalogs are bounded, however, by function and content. By and large, they serve the same purposes (identification and location of information objects), use the same data structure (the MARC format), and use the same data content (as prescribed by cataloging rules such as AACR2).

In addition, bibliographic data differ from typical computer data structures. Examination of manuals of screen design suggests that most applications use data of repetitive content and fixed length. In contrast, bibliographic data are highly variable in content and length, thus making the wholesale application of these general guidelines difficult. For example, labeling the data elements of even a simple MARC record can be challenging, because different types of data can reside in the same position of a display—as in the instance of the main entry, which can be a personal name, a corporate body, a conference name, or a uniform title. Furthermore, the existence of pre-established formatting standards for bibliographic data frequently render conventional design formulae inapplicable. ISBD, academic citation forms, and the catalog card—all of which existed prior to the introduction of online catalogs—are each familiar, alternative means of presentation of bibliographic data. These two factors—the nature of bibliographic data and alternative means of formatting bibliographic data—suggest that standard screen design criteria should be examined and tested.

Literature Review

The literature of human-computer interaction is large and diverse, and researchers in both the field of human factors and that of library and information science have studied many aspects of interface design. Researchers in the library and information science field have been concerned with the human-computer interaction specifically as it is manifested in the exchange between the online catalog and the library user. Indeed, the need for ongoing research into this interaction has been recognized for some time. For example, Cochrane and Markey (1983, 340) list "analyzing user requirements and behavior" as the first of four priorities in need of systematic study with online catalogs. In response to these perceived needs, a large body of data has been gathered regarding online catalog user behavior, but few authors have examined user preferences for online catalog displays. Human factors research is more varied, as the number and type of applications. It ranges from studies in which authors examine the effect that particular design attributes, such as color or highlighting, have on interface usability, to broad-based handbooks on general interface design. This literature includes numerous empirical studies that might inform online catalog design.

The efficiency of the information exchange between the user and the information system is a major concern in this study. The central question is whether participants perform better when using screens with particular data on them in particular layouts. Consequently, this review begins with a description of research about library catalog performance. Also examined is literature about human factors research and interface design, and cognitive aspects of users' utilization of an interface. Researchers interested in a more detailed review can see Thomas (1997).

Authors in each of these areas consider the human-computer interaction from different perspectives. Those who conduct system-oriented research examine this interaction in terms of information system features, and attempt to determine whether the system provides the user with appropriate functions and usability. This is relevant because screen design is integral to system design, and tests of system performance can include tests of screen layout. Some researchers in human factors attempt to view more specifically the effects of the presentation of system content and function on the human-computer interaction. Authors who study cognitive aspects of information system components seek to test the human-computer interaction in terms of how the brain processes information. Most of the researchers in each of these areas consider only the ability of the system to communicate to the user the structure and functions of the information system, however, and do not consider the effect of content presentation on the efficiency of the human-computer interaction.

System-Oriented Research

With the introduction of online catalogs to the library, the field of catalog use research has flourished. Computer-based systems have made it possible to gather data on users' searches that could not be gathered in the manual catalog environment. This capability has given librarians and researchers an

unprecedented opportunity to examine what users do with catalogs. Investigations of catalog use are relevant here because in this study the communication process between the user and the system is examined in an attempt to analyze sources of communications breakdowns. With regard to this research, however, most researchers have sought ways to improve the search engine or the database mechanics, rather than questioning the content of specific displays.

Two aspects of catalog use that have received a great deal of attention from researchers are subject searching and indexing, in large part because of their problematic natures (Larson 1991a; Markey 1985). Subject searching has been identified by users as the most difficult aspect of catalog use. Despite this, subject searching is one of the most frequently used methods of searching (Bates 1986). Authors of many such studies consider the mode of presentation of this information to be substantially deficient. Larson (1991a), in discussing the problems of zero retrievals and information overload, envisions a future interface with user-definable screen formats and functions. Similarly, Bates (1989) proposes an interface that uses a "Super-Thesaurus" to assist users in entering the catalog. Neither Bates nor Larson give details of how such innovative systems might be presented to the end user, although Larson does consider, in his discussion of methods for remedying subject search problems, the catalog interface to be one important factor in the overall success of the information transfer process.

Similarly, Borgman (1986) examines the difficulties that online retrieval systems pose to end users. Borgman identifies numerous problem areas that have been studied, grouping these into problems with mechanical aspects of retrieval systems and problems with conceptual aspects of the process of retrieval. She reviews the research concerned with the effect that individual differences play in this process, noting that most researchers have found that the search process differed depending on cognitive style, but that search results did not. Other researchers have found that overall experience with retrieval systems affected search results. Borgman points out, however, that little research has been conducted into the effect that interface design has on the success of searches. The only authors whom Borgman cites who are concerned particularly with online catalog interfaces, Martin (1974) and Hildreth (1982), list online catalog features, but do not evaluate how these features might be valued, interpreted, or used by the end user.

Harman (1992) places a different emphasis on this issue, and questions whether changes in the user interface can make a complex Boolean retrieval system truly easier to use. At their most basic level, the author argues, these systems require the end user to understand the theory behind Boolean logic, a requirement that she implies is too demanding. The author then describes four prototype search systems based on statistical retrieval techniques, each with a custom-tailored interface. Nonetheless, Harman concentrates on improving the functional aspects of the catalog—on improving search capabilities—rather than on improving displays. While these systems do offer a range of different display types, there appears to have been little testing of the screens on users, and thus little effort to determine whether they present the user with information in a more comprehensible manner.

Similarly, the majority of researchers who have gathered empirical data on user actions at online catalogs have attempted only to document the functional aspects of a system that are used, and not how users evaluated the data presented to them or whether they understood the meaning of these data. Both Borgman (1988) and Seymour (1991) provide reviews of this literature. Seymour arranges them by the method of data collection, with categories for surveys, experiments, transaction log analyses, interviews, and direct observation. In addition to the method of data gathering, Borgman also attempts to identify and categorize the variables studied. In this list, Borgman includes as a final variable for study the "effect of different interface methods (e.g., command, menu, form fill-in) on behavior," but then notes that this area "has not been studied empirically for information systems" (145). The need for research into mode of presentation is not raised.

Examples of use studies of online catalogs include: Larson (1991b); Millsap and Ferl (1993); Hancock-Beaulieu, Robertson, and Neilson (1991); and Siegfried, Bates, and Wilde (1993). The authors of these studies examine aspects of user online behavior, and typically offer suggestions on how these aspects might be changed or assisted in some way. Larson (1991b) used nearly complete transaction logs of a major academic library system for a six-year period to examine changes in search patterns. Using logs for the University of California's MELVYL online catalog, Larson found that subject searching over the period declined by about 2% per year as a percentage of searches performed. He suggested that this decline was the result of an increase in title word searching, as users moved away from the complexities and failures of subject searching. Larson recommended that online catalogs offer users assistance in mechanical aspects of searching, such as partial matching and stemming of terms, but he did not question whether the mode of presenting this information might be improved.

Millsap and Ferl (1993) examined the transaction logs of more than 1300 remote uses of MELVYL and found that a large number of searches were basic and retrieved one or more citations. However, a substantial minority of users had difficulty using the system, retrieving either no entries, or unmanageably large numbers of them. Again, the authors proposed adding online catalog features that would assist the user with mechanical aspects of searching.

Siegfried, Bates, and Wilde (1993) explored the search behavior of humanities scholars through transaction log and

protocol analysis. In their conclusion, they stated (288) that their "results suggest that the design of database search services . . . is still far from optimal for meeting the needs of humanities scholars." As before, the authors' main concerns were with finding means of improving the mechanical aspects of searching, and not the content of the user interface itself.

Thus, researchers in library and information science have focused largely on improving and documenting mechanical aspects of these systems. Most researchers would probably agree that online interfaces continue to need improvement; however, few have suggested that screen content or layout could play a role in this improvement.

Human Factors Research

Researchers into the human-computer interaction at online catalogs are acutely aware of the limitations of current-generation online catalogs, but their focus in finding solutions to these problems is on making the functions of the online catalog easier and clearer for end users. In the developmental stage of a new technology such as online catalogs, this focus is understandable and warranted. However, as online catalogs have become easier to use on a functional level, it has become increasingly important to test the ease of use of catalogs on an informational level—i.e., to test the means of conveying database content to the user. Library and information science professionals might look to the field of human factors research to find a body of research that can help inform many aspects of the design of online catalogs.

Human factors research is a broadly defined field that examines many aspects of the interaction that humans have with many different technologies. One facet of human factors research, however—that of computer interface design—has attempted to examine precisely some of the concerns that library and information science professionals need to consider. Human factors research into computer interface design will be examined here, with particular attention paid to the issue of screen layout and the effects that screen layout might have on system usability.

Several authors have written textbooks devoted to the design of user interfaces (e.g., Shneiderman 1992; Galitz 1989; Brown 1988; Dumas 1988; Crawford 1987; Norman 1990; Rubin 1994). Although the authors of these volumes support most of their design principles with empirical research results, this support weakens when each discusses layout and content choices. Instead, the authors tend to offer common-sense guidelines when discussing screen layout. The authors of these general texts rely largely upon common sense because there has been little empirical data gathered on the effect of a screen's layout on system usability.

Examples of empirical research in screen design include the work of Tullis (1981; 1983). In the first of these studies, Tullis (1981) found that reformatting screens to make them clearer significantly reduced decision-making time for users. From this, Tullis (1983) developed a means of measuring the complexity of screens based on groups defined by character proximity. Marcus (1982) discussed the role that graphic design principles played in the design of an interface for a geographic information management system. Kruk and Muter (1984) tested several aspects of reading text on video screens. The researchers found that reading text on video screens was slower than from printed matter, and that the vertical spacing of text on screen affected reading speed, but that screen contrast and distance from the screen had no effect on reading speed.

Trollip and Sales (1986) tested the effect that fill-justified text (i.e., text that has even margins on both sides) had on reading speed and comprehension. They found that reading speed was significantly slower for fill-justified text than for left-justified text, although comprehension did not differ. These studies, which were concerned with issues related to the impact that the layout of data has on interface effectiveness, might guide library systems designers in designing their displays, assisting with some of the many design decisions that must go into a typical online catalog display.

Tullis (1988) describes research conducted to identify and test the correlation between subjective and objective measures of screen complexity. In a preliminary experiment, Tullis isolated the correlation between these two measures by presenting participants with various screen displays and having them give a subjective rating of the screens. Participants were timed during the task, and these times were used to correlate the measures of screen complexity with time to complete the task. Tullis found that both time to complete the task and subjective rating were positively correlated to screen complexity. In a follow-up experiment, Tullis used the results of the first experiment to predict the outcome of a similarly structured experiment. Tullis found a high positive correlation between time and complexity, while the subjective rating/complexity correlation was not significant. The researcher then applied the evaluation criteria to earlier experiments of screen display, and found high correlation between the predicted values found by his criteria to the actual findings of those earlier studies, which lends support to the use of these criteria.

Tullis (1988) offers several design criteria that could be used to test alternative bibliographic displays. Tullis found that the number and size of the groups of elements onscreen had the greatest impact on search time, while the density of characters onscreen and the layout complexity (as defined by vertical alignment of screen elements) had the greatest impact on subjective rating. These criteria might well be applied to bibliographic displays, although the complex nature of bibliographic data and their potential for great length would need to be accommodated. Specifically, Tullis identifies methods for grouping data elements on-screen,

assuming that each element will occupy the same amount of space in every case—something that can't be assumed with bibliographic data. In online catalogs, data layout decisions must either be made once on a best guess basis, or for each record as it is prepared for display. Online catalogs to date have used the former method.

It is perhaps understandable that authors of human factors research—with the exception of Tullis—have given such scant attention to testing particular aspects of screen layout. Screen utility is closely related to the function of the application, the data that the system must convey, and the purpose to which the user will put the information. Given the variation in computer applications, such studies would likely be limited in their generalizability, and of only minor interest to nonusers of the particular application. Indeed, where there are areas of general concern, such as optimal menu design or the relative benefits of menu and command interfaces, there exists a body of research. For the purposes of this research, however, an examination of the human factors literature offers little guidance, because the authors of this literature mainly focus on conceptual and functional aspects of interfaces, whereas this study has been focused on those aspects of the interface that facilitate communication of information between system and user after the user has already navigated the conceptual barriers.

Other human factors researchers have conducted research into the mechanical aspects of the computer interface, examining the effects that various interface features have on user performance. However, the authors of these studies, like those in the library domain, have focused primarily on the mechanical or functional aspects of the interface, and not on the effect that layout might have on information transfer. Even so, such studies underscore the fact that display factors can affect user performance, even if the researchers do not generally address questions of screen layout.

In the more specialized area of library information systems, the literature is sparse and focused on content rather than presentation. Shaw (1991) reviewed the recent literature in this area, but did not discuss any studies that test the effect of layout on performance. It is interesting to note that several authors cited by Shaw noted the need for empirical evaluation of user interfaces. In a study of card catalog use, Palmer (1972) found that most users concentrated attention on author, title and call number data; subject headings were used by just under half of the users. From this, Palmer suggested that a five-item catalog using author, title, call number, subject headings, and date of publication would meet the needs of a substantial majority of the users' requests.

Seal (1983) reported that 90% of studied users were satisfied with a short entry catalog. Hufford (1991) found that the reference librarians he studied used only particular data elements in most situations, confirming the results of earlier studies. Matthews (1985) provided a set of guidelines for online catalog screen displays based on guidelines from human factors literature, again derived from common sense. Shires and Olszak (1992) reviewed basic interface design issues for online catalogs and focused almost exclusively on screen design for conveying command and search concepts more clearly. Yee (1991) surveyed research on user interfaces in order to identify research methods and findings applicable to the design of effective user interfaces to online catalogs. On the issue of single record display, Yee points out that many have asserted that labeled displays aid in user comprehension, but only one group of researchers has in fact studied this phenomenon.

In a handbook for library interface design, Crawford (1987) relies almost exclusively on his own judgments of interface design upon which to base design decisions. In this volume and in Crawford, Stovel, and Bales (1986), Crawford asserts the superiority of labeled displays without offering evidence to support this choice. He does examine this issue in some detail, noting, for example, the fact that choosing a citation format is complicated by the lack of standardized citation formats. In addition, he attempts to address some of the problems that bibliographic data commonly present in designing displays, especially the variability and complexity of the data, and its potential size. While the design principles that Crawford presents are sound, there is no attempt to test these principles empirically.

Crawford, Stovel, and Bales (1986, 2–3) identified five questions of importance to the design of bibliographic displays:

1. Does the display provide an appropriate amount of information?
2. Will patrons understand the information as it is displayed?
3. Is the display readable and attractive?
4. Will patrons be able to find information rapidly and to find all the information needed?
5. Will patrons be able to view the information on a single screen?

The first four questions deal with important display issues regarding the human-computer interface. The authors state that most work on interface design has been focused primarily on the first, and to a lesser degree the second, third, and fourth questions. The authors then examine the more practical fifth question. They do not attempt to address these first questions further; instead, they use aesthetic judgments to guide discussion of screen design. They then present a number of display options, and assert the relative value of these screens without offering any empirical data to support their claims. At no time do the researchers offer empirical evidence to support their choices of: fields to display, labels to use, or orders of information to present. Thus,

although the authors discuss the relative merits of differing display types, they do so from a perspective limited in its basis to personal opinion and not supported by the results of any wholly objective assessment.

In a different vein, Smiraglia (1992) provides a helpful outline of the basic goals of the catalog, which he labels identification, collocation, and evaluation. This outline might be used to help inform interface design. The first two functions, identification and collocation, depend upon presenting traditional bibliographic data elements. As Leazer (1993) notes, Smiraglia's identifying function closely parallels those of Cutter (1904) and the Paris Principles (1963)—that is, identifying particular items in a library collection. Two parts of Cutter's first function of the catalog—those of identifying whether a library has any given item by an author or with a particular title—clearly have as their primary focus the bibliographic item. The third part of this function—identifying whether a library has a book on a particular subject—does not focus as directly on the item. When users seek material on a given subject, they do not usually have particular items in mind. Because of this, this part of the function should perhaps be considered separately. For the first two parts, traditional bibliographic data elements—author, title, and publication information—might sufficiently serve in most cases, for specific item searches.

The collocation function seeks to bring together items with a particular attribute, e.g., author or subject. As with the identifying function, traditional bibliographic data elements serve this function in most cases. The most notable failure here is the lack of topical data elements included among most traditional bibliographic data elements. Moreover, this function is more complex, in that the nature of the use is less precise. Users who seek to know what a library has on a particular topic or by a particular author have a less precise knowledge of the items they seek than those who seek specific items. This places different demands on the information system, and requires that the system provide enough information to the user that an evaluation can be made.

The third function, the evaluative function, is problematic in that it is highly situational. It is the point at which the relevance of retrieved items is determined. Relevance is one primary means of measuring the effectiveness of an information retrieval system. However, the data necessary for deciding upon the worth of a given bibliographic item is dependent on the user's particular information need, a need that is likely to be poorly defined and subject to constant change. Indeed, the detailed study of the particular data elements that any user might need at any given time could form the nucleus of a large body of research, which is beyond the scope of this research.

Marchionini (1992) outlines the information retrieval process and features of systems to support that process. He uses three basic assumptions about information seeking, two of which have bearing on the current research: first, that end users aim to solve a problem, not use the information retrieval system; and, second, that end users seek to reduce the cognitive load of using an information retrieval system. This second observation is supported by Akeroyd (1990) as well as Tullis (1988). Marchionini, Akeroyd, and Tullis all seem to suggest that the online catalog interface should facilitate the search process in part by presenting bibliographic data to the user in the most effective manner.

Branching off the basic research in interface design is a broad area of research into prototype interfaces and system innovations. The authors of these studies typically take as their starting point an acknowledged failure of current interfaces, and devise some new means of providing the user with system features. As with system-oriented studies, however, many authors (e.g., Frei and Jauslin 1983; Davis and Shaw 1989; Noerr and Bivins Noerr 1985; and Belkin, Marchetti, and Cool 1993) focus on simplifying search functions, and not on clarifying screen content.

Authors in one area of specialization do concern themselves to a great extent with the best means of presenting information to the user. These researchers, who are all studying the problems associated with the visualization of information, have begun to examine whether modes of communication other than textual communication might facilitate information retrieval. Olsen et al. (1993) presented one such system, VIBE, where users perform an information retrieval task using a two-dimensional plane with "Points of Interest" that can be manipulated to find relevant materials. Beheshti (1992) outlined another such system, where MARC records are rendered as virtual books on shelves that are displayed to the user on screen. The researcher envisioned this system enhancing users' abilities to browse library collections. In recent years, this area has received substantial attention, and offers promise for alternative means of informing users of database content.

In summary, then, researchers in the human factors area, both within and outside the library field, have explored many features of interface design. The problem remains that, in addressing the issue of screen layout, a great deal of this literature is not based on empirical research, but rather on aesthetic judgments or common sense. There are notable exceptions. For example, Tullis's work does present such empirical evidence. Unfortunately, this research has not been replicated using bibliographic data.

Cognitive Research

Research into cognitive issues in information systems could lead to the development of predictive models of specific display need. In one area of this field, researchers attempt to create models of the cognitive makeup of users, so that sys-

tems might be designed to predict the particular needs of users in given situations. Individual user profiles can then be created by building on these base models, thus enabling the system to anticipate an individual's preferences, and to modify the model iteratively. One aspect that these systems could manipulate in response to these models is the screen design and layout. For example, an expert system designed to use cognitive models could provide greater or lesser screen complexity depending on a number of user variables.

Morehead and Rouse (1982) and Brajnik, Guida, and Tasso (1987) explored the use of such user models in information retrieval systems. Aykin and Aykin (1991) reviewed the literature of individual differences and their effect on human-computer interaction. The authors pointed out in their conclusion, however, that "there are only a few studies considering a limited number of user characteristics" (377). Although the authors of much of this research again focus on user models to predict system functional needs, research into screen design could potentially be affected by such considerations.

Several authors have examined the impact that cognitive differences between individuals can have in the information exchange process. Hensley (1991) offered suggestions on how reference librarians can use learning style theory to facilitate the reference process. Prorak, Gottschalk, and Pollastro (1994) tested the effect that teaching method has on teaching bibliographic instruction; the researchers found no significant preferences. Allen and Allen (1993) tested the relative cognitive abilities of librarians and students, and found significant differences. They found that librarians had higher logical and verbal comprehension capabilities, while students had higher perceptual speed. The authors considered that these differences might suggest that individuals use information retrieval systems differently as a result of cognitive differences. Again, these researchers did not examine the effects that these cognitive differences had on screen preferences.

Ford and Ford (1993) tested whether different cognitive strategies affect success in addressing an information need. The researchers asked students to learn about a particular indexing system using a knowledge base that included (unknown to the participants) two human experts. An examination of the search sessions showed that participants asked different types of questions of the knowledge base, and that the question types did affect problem-solving success. The researchers found that two approaches—one associated with female participants and focused on detailed information, the other with males and focused on broader analytic information—led to success, while a third, which focused on middle-level concepts, did not work. In their conclusion, they linked the two successful approaches with specific cognitive styles identified by earlier researchers: operation learners and comprehension learners. They then argued that establishing a system that accounted for cognitive style would be advisable, but that such a system should leave the choice of mode up to the user. However, asking all users to choose the search mode for a session seems cumbersome and requires users to be able to identify their own preferred style. The reader will recall, moreover, that Marchionini (1992) suggested that users do not care about how a system works, just that it works.

Allen (1994) used an experimental method to perform two tests of the relationship between a cognitive ability and the ability to perform a retrieval task. In the first test, Allen tested the relationship between logical reasoning, as demonstrated on a standardized test, and ability to identify relevant citations from two online catalog displays, one sorted in reverse chronological order, the other in relevance order. The researcher found a significant interaction between logical reasoning ability and the type of display, and particularly that users with lower logical reasoning ability benefited most from the relevance order display. The second experiment tested the relationship between perceptual speed and the ability to examine subject headings arranged either alphabetically or hierarchically. In this experiment, the researcher found no significant results. However, users of the hierarchical display examined substantially fewer lines to identify relevant headings, which might indicate that the method of presentation affected participants' perception of the system. Both experiments, then, offered evidence that the interface can affect usability of the system.

Kerr (1990) tested the impact that different screen enhancements have on assisting users in navigating through electronic information systems. The author found no significant difference in usability between systems with no screen enhancements and screens with enhancements of color, icons, textual headers, or a combination of the three. Coll, Coll, and Nandavar (1993) presented research designed to test whether good screen design depends on layout or conceptual considerations; the researchers found that conceptual structure plays a greater role. Participants were asked to choose items from menus that were arranged alphabetically, categorically or randomly, and on which the function key assignments changed from screen to screen. They found that task times for the two organized menus, while similar to one another, were significantly shorter than the unordered menu. This study reinforced the idea that how data are displayed can affect system performance.

Overall, the authors of the literature reviewed in this section focus primarily on testing aspects of the functional human-computer interface. Library researchers have extensively examined how users search and the errors that they have made, through transaction log analyses, user surveys, interviews, and other methods. A great deal of attention has been paid to the complexities of the search and retrieval process, with many studies comparing the performance of

different search methods. Similarly, researchers in human factors have tested the relative usability of command-driven and menu-driven interfaces, and the best means of organizing items in a menu. Some research into overall screen display has been conducted, and some measures of screen complexity have been devised, but these tests and guidelines have not been tested with bibliographic information.

Importance of the Study

Because end user access to remote databases of all types is rapidly increasing, library systems designers must assume greater ability to be responsive to the end user's needs. In a remote access situation, the user must rely entirely on the system designer's ability to present the system to the user sensibly; professional librarians cannot interpret displays. A library system that presents end users with data that are more directly related to the needs of the user's problem-solving process and are easier to interpret will require less explanation to the end user. As electronic resources become more central to library operations, libraries' overall effectiveness will be judged in large measure by the performance of their information systems.

To design systems that better serve the end user, it is essential that library and information science professionals understand more precisely what users need to see onscreen and how they want to see that information. While it is important to design more sensibly organized interfaces on a functional level, the question of display nonetheless remains critical—especially with data as complex as those found in a bibliographic database. An interface that offers information that is perceived as more pertinent will be less obtrusive, and will thus enable the user to focus on the primary task (Benbasat and Todd 1993).

It has been claimed (cf. e.g., Crawford 1986) that users of bibliographic systems are unaware of the content of particular screens, and that they cannot distinguish the different elements of a bibliographic record. While it might be true that typical end users do not consciously make these distinctions, they nevertheless might be aware of them on some level. It behooves librarians to make their systems as clearly organized as possible, both from functional and communications perspectives. This research was conducted in order to provide baseline data that could make it possible to determine how the content and layout of interfaces might be changed to make bibliographic displays easier for the end user to interpret.

Gaining an insight into the display needs of library catalog users should help librarians make informed decisions about changes to catalog content and structure. A number of library professionals have called for a radical reassessment of cataloging practices, so that both the MARC format and the cataloging process might be simplified (e.g., Gregor and Mandel 1991; Mandel 1985). Such reassessments have been hindered by the lack of data concerning such user preferences and needs. This study, while not aimed primarily at determining the importance of particular data elements in bibliographic databases, might nonetheless provide a baseline of preferences upon which others might build. Identifying individuals' preferences on display screens might provide a new perspective on which data elements they consider important to their problem-solving process.

It has been suggested that the need for evaluating user preferences in online catalog displays will soon be obviated in the client-server environment, where an individual will be able to configure the database view in any preferred format (e.g., Larson 1991a; Buckland, Norgard, and Plaunt 1993). While end user configuration might alleviate some problems of choice of sorting and display, it seems unlikely that it would help the casual user, who must use the online catalog "out of the box." Moreover, many users will not make the effort to customize their displays. For such users, default displays must exist to present online catalog results.

In the near term, it will continue to be necessary to provide all users with fixed display types, and common sense suggests that these displays should provide information to the end user in the most sensible context. Externally created limitations force users to adapt their needs to the system, and the system does little to mitigate the user's interpretive burden. Every such burden compounds the user's negative associations with the online catalog, while reducing this burden increases both the chances of a user's success and the user's positive associations with that system. An aim of the research reported here is to establish a basis for the design of more effective online catalog displays.

Method

A 2 x 2 factorial experimental design was used. The participants, first-year students in a four-year undergraduate program at a large university, were randomly assigned to one of four groups and then pretested for basic demographic information. They were asked to perform two related tasks during the course of the experiment. In the first task, they were asked to select a set of items that they would examine further for a hypothetical paper they must write. In the second task, they were asked to examine 20 bibliographic records, decide whether they would choose to examine these items further on the shelf, and identify the data elements that they used to formulate their relevance decision. In both tasks, the topic of the task was the same: big band music and the music of Duke Ellington. The instructions for forming relevance judgments in each task were also the same: participants were asked to use only the information provided on screen to make relevance decisions.

Each group was presented with a simulated online catalog that listed a set of bibliographic records sorted in reverse chronological order, from which they had to make their choices. The listing for each group differed in the data elements and screen layout used, but contained the same bibliographic information. That is, they all viewed the same bibliographic records, laid out in different on-screen configurations.

One group viewed bibliographic records on an interface similar to current online catalogs, one that used data labels and contained brief display data elements from MARC tags 1xx–3xx. A second group viewed these records on an interface in which the labels had been removed, but the brief display data elements were the same as those in the first. The third group used an interface that employed tags, but in which the brief display data elements were changed so that they included more subject-oriented fields (omitting main entry and physical description fields) while the full display was based on ISBD data elements. The final group viewed these records with the same brief record data elements as the third group, but with the labels removed, using ISBD and AACR2 punctuation standards. Using both tagged and untagged interfaces for each of the two content types helped control for differences in performance caused by particular layouts. Samples of these screens are provided in figures 1 through 9.

Retrieval speed and comprehension were tested in the first task by gathering transaction data for each participant. The system logged each participant's actions: the screens that were viewed and the time spent viewing them. Quantitative data were gathered on: the number of screens viewed; the number of citations examined at each level of specificity; the time to complete the task; and the time spent viewing particular screens.

Bibliographic records for the experiment were gathered by searching the Research Libraries Information Network (RLIN), the OCLC Online Union Catalog (OLUC), the Library of Congress catalog, and the University of Pittsburgh catalog. General searches were made for the keywords "Duke Ellington" and "Big Band Music"—both terms taken directly from the problem statement. Duplicate records were removed from the master list, under the assumption that a single library system would have one catalog entry for each of these items. 103 bibliographic records were used in the experiment. To populate the second task, a random set of 20 records was separated out, leaving the first task with a list of 83 records. For the second task, the 20 records were randomly ordered and randomly assigned to display in full or brief layout.

Screen Design Issues

The choice of screen elements for use in the different displays posed an important problem in the design of the experimental interfaces. Because the number of data elements included in the MARC formats numbers in the hundreds, an attempt to test the relative value of all different combinations of these elements would have been impossible. Previous research had shown that most users tend to employ only a limited portion of the bibliographic record (Palmer 1972; Seal 1983; Hufford 1991), and that briefer catalogs containing only a few of the data elements currently found in bibliographic databases would serve a substantial majority of end users' needs. Based on these earlier studies, brief display screens for this experiment were designed that included data elements from a limited set.

One method for choosing alternative data element sets for display to users might be to examine the data elements in a bibliographic record in relation to some of the accepted

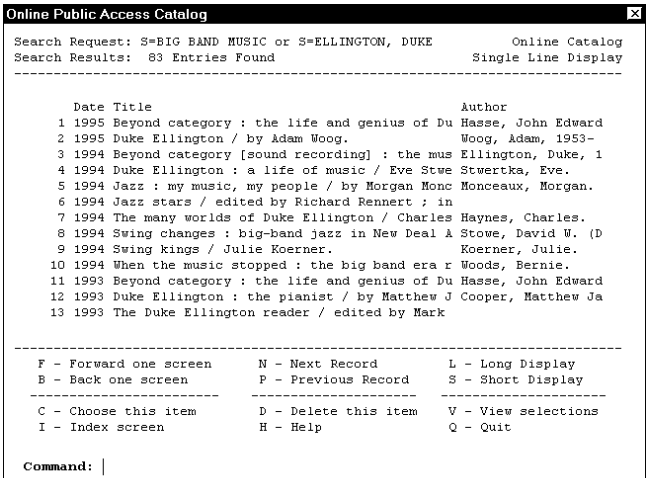

Figure 1. Single Line Listing, Used on All Interfaces

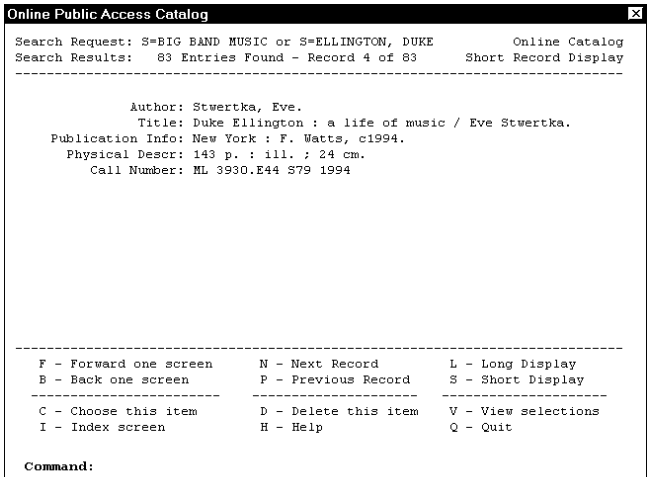

Figure 2. Standard Content, Labeled Brief Display

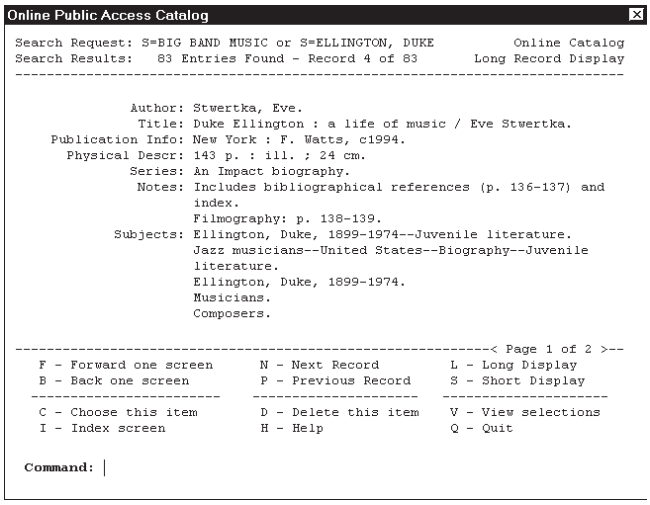

Figure 3. Standard Content, Labeled Full Display

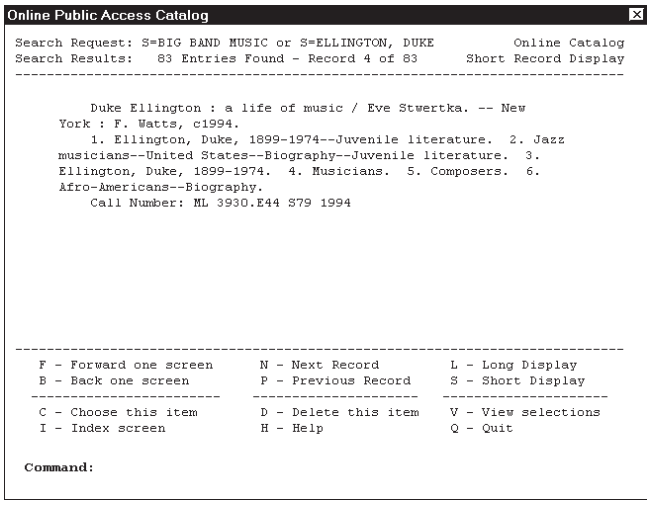

Figure 4. Enhanced Content, Unlabeled Brief Display

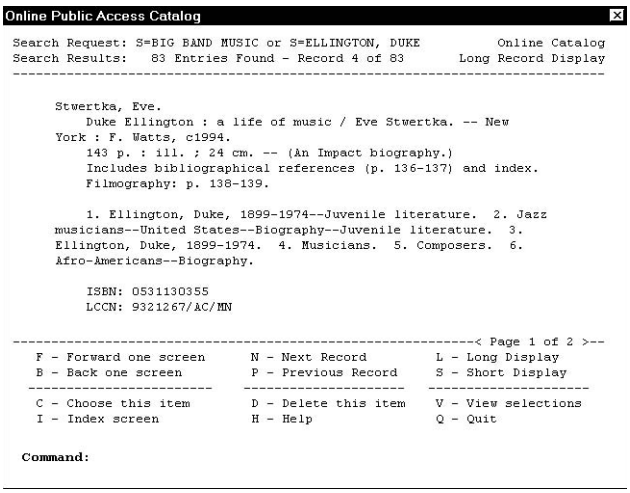

Figure 5. Enhanced Content, Unlabeled Full Display

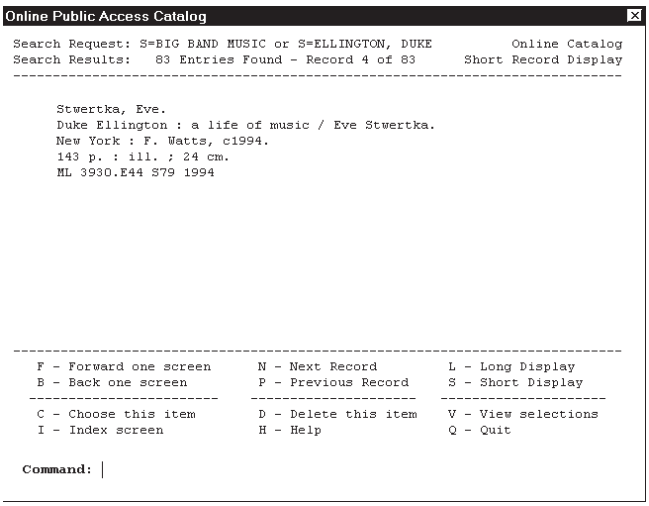

Figure 6. Standard Content, Unlabeled Brief Display

functions of the catalog—most notably Cutter's definitions (Cutter 1904) and those found in the Paris Principles (International Conference 1963). Based on such a functional analysis, a set of data elements can be chosen for inclusion in particular displays. These displays would then ostensibly support the stated functions of the catalog.

This functional approach was used, and provided a basic framework within which data elements were selected for inclusion on the bibliographic screens. In this research, attention was focused on the evaluative function of the online catalog by inducing the participants to use a topically oriented search strategy. By controlling the content of the user's information need—as in assigning an artificial task to a particular group—it might be possible to test whether the content and layout of a screen has any effect on participants' performance in that given situation.

The evaluative function was chosen because it would force the participants to rely more heavily on the information provided. Users who seek a particular known bibliographic item in a collection might not seek to view the same data elements as those users with a more generalized information need. Similarly, users who seek quick answers to solve an information need will potentially need to view different data elements (in this case, perhaps, the extent or availability of the bibliographic entity). For the purposes of this research, participants were given a topic of which they

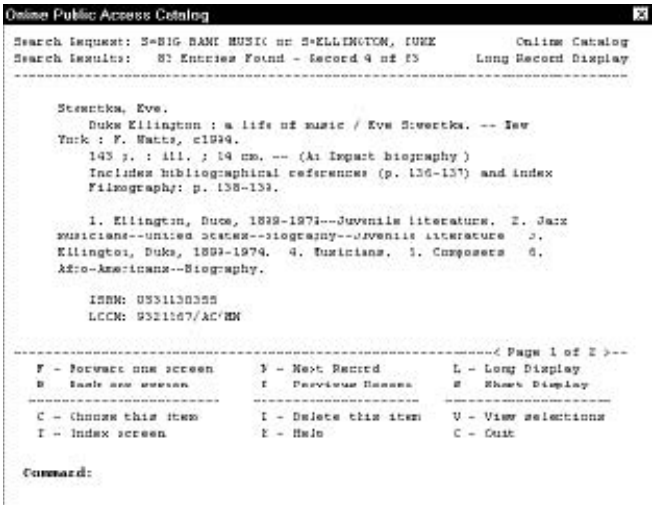

Figure 7. Standard Content, Unlabeled Full Display

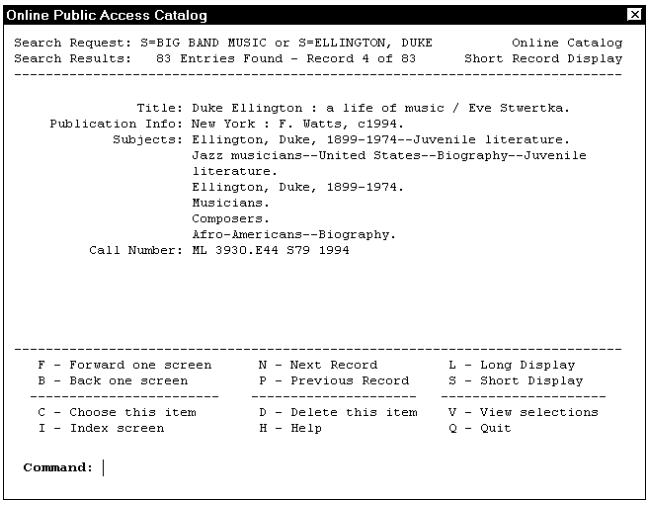

Figure 8. Enhanced Content, Labeled Brief Display

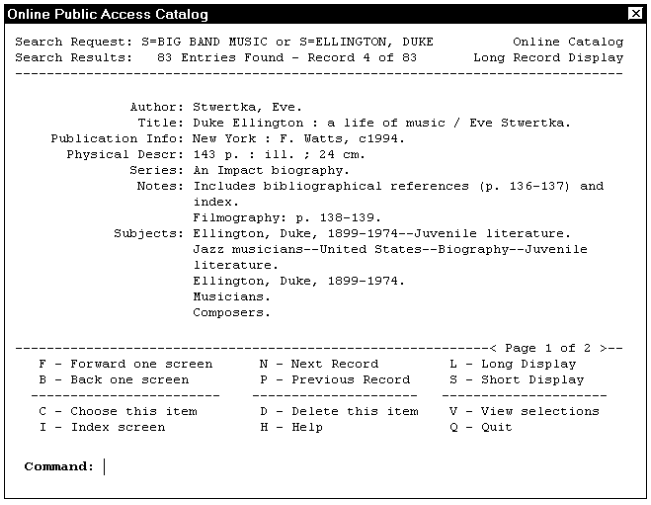

Figure 9. Enhanced Content, Labeled Full Display

presumably had little foreknowledge, and were asked to compile a list of most useful sources based on the information provided in the information system.

In known-item searches, users seek to verify or reinforce certain facts that are already known. In contrast, subject-oriented searches are more tentative and probing in nature. The user seeks to discover whether the library owns bibliographic entities—or more generally whether there exist any bibliographic entities—that might meet an information need. This more diffuse need places different requirements on the system. Users seeking known bibliographic items can rely on their ability to differentiate patterns of data and match on incomplete user knowledge, whereas those who seek to meet a more visceral need must evaluate the bibliographic data provided to them more closely. As a result, such users are more likely to rely heavily on the data presented by the system to determine the worth of those items for solving the information need. Such use places greater demands on the information system, which is the reason why this type of situation was established for this research.

In order to conduct this research, content designations for three screen types were designed: one that listed each item on a single line; a second that listed an individual item using a brief citation; and a third that listed the full contents of an individual item. These screens were designed based on the results of prior research (e.g., Palmer 1972; Seal 1983; Hufford 1991), as well as on library standards (e.g., ISBD and AACR2). For the single line display, brief title (MARC 245), author (MARC 1XX), and date information (MARC 008) were used; for the treatment group brief display, title (245), edition (250), publication information (260), subjects (6XX), and call number (090) were used; the treatment group full display included all MARC fields.

Administration of the Test

Participants volunteered for one of eight sessions over two days. Due to hardware limitations, 8 groups of 16 participants were scheduled for the experiment. Within each session, all four interfaces were tested, with each interface being assigned to an equal number of terminals, which were grouped together. Every participant thus had an equal chance of using any of the interfaces. They were grouped together to minimize the chance that neighbors would notice differences in the interfaces.

Participants were instructed to seat themselves at any terminal. Each interface used the same opening screen, which eliminated the possibility that participants might choose one interface over another based on some screen difference.

The use of several different administrations of the test to several groups of participants could have introduced bias in the administration of the experiment, as the investigator's behavior might have varied from session to session. To minimize this possibility, instructions for the experiment were scripted, and all groups thus received the same instructions. As a further means of guaranteeing impartiality, a third party was hired to administer the experiment. This person was responsible for interacting with and instructing the participants in the experimental setting. To minimize experimental bias, this person was not informed of the research questions. Additionally, the experiment administrator restricted her interactions with the participants. She stood apart from the participants, so she could not see any particular participant's progress, and only approached participants if they were having trouble and specifically requested assistance. When this occurred, the administrator answered only the particular question that the participant posed, and then moved away.

Participants were seated in the labs, and the administrator briefly explained the study, and said that all information gathered during the test was to be kept confidential. Participants then answered the questions of the pretest questionnaire. Participants were all given a tutorial session designed to help them familiarize themselves with the interface, during which time the administrator addressed any difficulties participants had using the system. Then the administrator presented to the participants the problem situation, reading from the problem statement. Participants selected the appropriate number of items for inclusion on their bibliography, and then answered the qualitative portion of the experiment. During the experiment, the administrator monitored the test site, to ensure that the experiment proceeded smoothly.

Results

The experiment was conducted in mid-October 1996. Eighty-two participants took part over a four-day period. The number of participants in each of the four interfaces was as follows: standard content, labeled display, 19 participants; enhanced content, unlabeled (ISBD) display, 19 participants; standard content, unlabeled display, 22 participants; and, enhanced content, labeled display, 22 participants. Thus, when examined by each of the two factors (absence or presence of labels, and standard or enhanced content brief screens), the total number of participants in each group was 41.

Data were compiled into a database, and output from this database was analyzed using a statistical software package and a spreadsheet program. Tests used for the analysis of the data are listed here, along with the identifying label used to state statistical results: student's t-test (t); Chi-squared (χ^2); One-way Analysis of Variance—ANOVA (f).

The statistical analysis covered three general areas: the demographic data gathered; the quantitative data gathered in the first task; and the qualitative data gathered from the second task. These results will be discussed later.

Overview of Demographic Data

Thirty-three males and 49 females participated. The participants ranged in age from 17 to 46, with a mean age of 20.21. As might be expected of a sample drawn from an undergraduate course, a large number of participants (67) were aged 17 to 20. Twenty-three participants (28%) had not yet determined their majors. The remaining data regarding majors were categorized broadly. Those that identified a major tended to indicate majors in the applied sciences (28, or 34.1%), social sciences (12, or 14.6%), or business (13, or 15.9%). The other categories were: humanities (3, or 3.7%); vocational (2, or 2.4%); and law (1, or 1.2%).

The participants claimed to use computers quite heavily: 44 (53.7%) said they used computers every day; 24 (29.3%) used them 2 to 4 times a week; 9 (11.0%) used them once a week; and 5 (6.1%) used them once a month. The mean computer use level was 3.30, which translates approximately to the 2 to 4 times a week category. Tests to see whether computer use differed between the factor groups found no significant differences(labels factor: χ^2=1.45758, p=.69210; content factor: χ^2=1.95758, p=.58126).

In contrast, library online catalog use was much lower: 6 (7.3%) stated they used library catalogs 2 to 4 times a week; 13 (15.9%) used them once a week; 34 (41.5%) used them once a month; and 29 (35.4%) had never used library online catalogs. Mean library online catalog use was .95, which translates to just below the once a month category. Once again, no significant differences were found between the groups for either of the factors.

It was found that that most participants had little knowledge of the problem topic. Knowledge of big band music broke down as follows: 2 (2.4%) said they knew a lot about it; 24 (29.3%) knew a bit about it; 32 (39.0%) had heard of it; and 24 (29.3%) knew nothing. The mean knowledge level of big band music was 1.05, which equates approximately to having heard about it. Knowledge of Duke Ellington broke down as: 2 (2.4%) were quite familiar with him; 16 (19.5%) knew a bit; 31 (37.8%) had heard of him; and 33 (40.2%) knew nothing. The mean level of knowledge about Duke Ellington was .84, which again equates approximately with having heard about him. No significant differences were found for knowledge of Duke Ellington on the primary factors.

There was no significant difference of knowledge of big band music when compared on the content factor; however, there was a significant difference when compared on the labels factor (χ^2=8.45833, p=.03743). Because this was the only significant difference found in the two topically oriented variables, a second comparison was made between the *average* of the two values in the topical questions and the experiment factors. When grouped this way, no significant differences were found.

In summary, the participants in the experiment were more likely to be female than male, and generally in the traditional age range for college students. They were most likely to be majoring in the sciences, although a good portion had not yet decided on a major. They used computers approximately 2 to 4 times a week, but rarely used library online catalogs. They had only a passing knowledge of the two aspects of the problem topic. The distribution of participants to the various experimental groups was not significant.

Task 1: Creating a Select List

Overview of the Data

For the first task in the experiment, a total of 5,970 log entries were tallied. A problem with the logging of the first screen in 24 of the sessions required that these entries be dropped from the data set, resulting in a valid set of 5,946 log entries. Help screens were also dropped, as they were used so infrequently (27 times overall), that their inclusion affected statistical analysis. The total number of log entries used for all further analysis thus was 5,919. Each log entry included: the command issued, the record number of the item being viewed, the type of screen being viewed (Index, Brief, Full, or Help), an error code, and the start and end times for the screen. These data elements make it possible to analyze the effects of the two factors on both the frequencies of screens viewed and on the duration of time spent on screens.

Within each data type category (i.e., frequencies and durations), analysis was performed in several different steps. First, all log entries for each complete session were analyzed in aggregate. Then, the data were separated by screen type to provide a more detailed method for testing the behavior of the participants while performing the task.

Making this distinction was necessary for a number of reasons. First, each of the three screen types used in the experiment had different content and complexity of data. Index screens are columnar, making scanning easier, because like elements are grouped visually on-screen. However, the inclusion of more than one bibliographic item complicates the user's decision on which items to view in greater detail. In contrast, both brief and full display screens present information for just one bibliographic item, thus eliminating this selection function; these screens differ in that the full screen is usually more difficult to process because there is more information on-screen to examine.

Another important reason for examining each type of screen separately was that the experiment's factors both relied on particular screen displays for the comparisons. The screen content factor—that is, the effect that enhanced screen content had on participant behavior—relied solely on the difference in brief display content for the different groups. Similarly, the absence or presence of labels occurred only on the single item brief and full display screens. Thus, it was decided to examine each type of screen separately, to test the effect these factors had on performance.

Further analysis was then carried out using data compiled for both the first 10 and the first 20 screens viewed. Because it was presumed that participants would become increasingly familiar with the interfaces they used, and that this learning process might manifest itself in behavior differences over the course of the session, it was concluded that an overall measure of session activity might not give an accurate view of the ease with which participants used these systems, and that separating the first screens each participant viewed might give a clearer sense of the learning process. It was judged, moreover, that comparing among the factors might also give an idea of whether one group of participants learned to use the system faster than others, and that further analysis based on a comparison of the data from the first 10 screens with those of the second 10 screens might enable us to pinpoint the rate at which the learning process occurs.

It must be noted that in compiling these initial screen statistics, erroneous durations for the first screen of 24 sessions were dropped from the sample; consequently, the statistics for these items includes data for screens beginning with the second screen viewed.

Several more analyses were carried out on the duration data alone. For this experiment, the question of the speed with which participants performed the tasks presented to them was an important one. Speed in a test like this might be measured in a number of different ways—for example, we can measure the overall time spent to perform the task or the mean time spent on specific screens. Another method for measuring speed is to use the number of characters processed per second to test participant performance. This measure can be derived by counting the total number of characters on each screen that the participant viewed and dividing that by the duration of time spent on the screen. This measure controls for the amount of information that participants viewed, and might help determine most accurately the speed with which they were able to use the catalog information presented to them.

Building on the idea that performance might change over time, analyses of screen durations for the first 5 screens

of each screen type were conducted. These were done based on the hypothesis that each screen type presents its own processing problems to participants; analyzing each screen separately might enable us to pinpoint to a greater degree where users have difficulties with catalogs.

One final category of data was gathered in this portion of the experiment: the participants' final selections for their hypothetical reading lists. Analyses of these data were conducted to test whether the experimental factors or demographic variables had any effect on the items that participants selected.

Analysis of Frequencies of Screens

Overall Frequencies. The mean number of screens viewed in each session was 72.51, with a standard deviation (SD) of 25.81 and a range from 30 to 177 screens viewed. Means for each of the groups in the two factors were compared using a t-test, with no significant differences being observed. Analyses of mean number of screens viewed for each type of screen were conducted. For index screens, the overall mean was 26.02 with a SD of 13.34 and a range from 1 to 58. Neither of the factor-based means varied from that noticeably. There was, however, a significant difference in the mean number of index screens viewed by gender, with males viewing on average 30.06 index screens and females viewing 23.31 (t=2.23, p=.030).

For brief screens, the overall mean number of screens viewed was 25.51 with a SD of 20.35 and a range from 0 brief screens viewed to 115. When examined by screen content, the means differed, with enhanced content participants viewing 29.46 brief screens compared to 21.56 screens viewed by standard content participants. This difference was not significant at the .05 level, but was significant at the .10 level—thus there is the likelihood that significant results would be found in further studies (t=-1.78, p=.08). When compared on the labels factor, the difference was not significant.

For full screens, the overall mean was 20.65 with a SD of 24.57 and a range from 0 to 146. When examined by screen content, the means differed. For full screens, the relationship of the numbers was reversed from that found with brief screens; enhanced content participants viewed on average 15.95 full screens, while standard content participants viewed 25.34 screens. Again, this difference was not significant at the .05 level, but was significant at the .10 level (t=1.75, p=.085). Differences between groups on the labels factor were not significant. The number of help screens viewed was so low as to preclude analysis.

Chi-squared tests were also performed, comparing frequencies of the different screens viewed to the two factors, as well as to the following demographic variables: gender, computer use, library use, big band knowledge, Duke Ellington knowledge, and major category. For the two main factors, the difference between groups on the label factor was not significant, while the difference on the content factor was significant (χ^2=138.40842, p=.00000); participants using the enhanced content interface viewed significantly more brief screens, and significantly fewer full screens.

Chi-squared tests on each of the demographic variables were found to be significant. Women viewed fewer index screens, fewer full screens, and more brief screens than did men (χ^2=62.28387, p=.00000). Infrequent users of computers viewed fewer index screens than did the heaviest users (χ^2=82.74381, p=.00000). Participants who said they had never used library catalogs, and those who used them weekly viewed fewer full screens than others, while monthly users viewed more index screens (χ^2=83.02628, p=.00000). Participants who knew nothing about big band music viewed fewer full screens (χ^2=100.41911, p=.00000). Similarly, those who knew nothing about Duke Ellington viewed fewer full screens (χ^2=32.52557, p=.00001). Participants who had not decided on a major and those in the social sciences both viewed fewer full screens, while those in applied sciences viewed more full screens (χ^2=182.65344, p=.00000).

First 10 Screens. Tests of the frequency of each screen type for the first 10 screens were conducted on both primary factors. No significant difference was found with the contents factor; a significant difference was found between groups on the labels factor. Participants using labeled displays viewed fewer full screens and more index screens than did those using unlabeled displays (χ^2=11.09244, p=.00390). T-tests for each screen type were conducted; a significant difference was found for the mean number of full screens viewed per session when compared on the labels factor (t=2.08, p=.040).

In analyzing the first 10 screens compared to the demographic data, a number of significant results were found. Chi-squared tests comparing screen type by level of computer use yielded significant results (χ^2=14.18658, p=.02762); participants with the least computer experience did not use full screens at all, while other groups used them in the neighborhood of 15% of the time. Novice users instead viewed more brief screens than others. Significant differences were also found when comparing library use to frequency of screen type (χ^2=20.16583, p=.00259). Those with the most library experience viewed more full screens than others.

Differences among groups with regard to knowledge of big band music were significant (χ^2=16.10567, p=.01320), while differences among groups based on Duke Ellington knowledge were not. Those who had heard of big band music viewed more brief and fewer full screens. Significant differences were found between gender groups (χ^2=15.16189, p=.00051); men viewed more index screens and fewer brief screens than did women. When examined

by major, those in the social sciences viewed more brief screens than other groups (χ^2=28.55680, p=.00458).

First 20 Screens. The first twenty screens of each session were analyzed separately, in order to isolate potential learning by the participants further. T-tests compare the mean number of each screen type viewed to each of the primary factors as well as to each of the demographic variables. No significant differences were found. Chi-squared tests of the frequency of each screen type yielded several significant results, similar to those found for the first 10 screens. For the labels factor, participants using labeled screens viewed more index screens and fewer full record screens than did those using the unlabeled screens (χ^2=7.94, p=.01889). For the content factor, participants using the enhanced content interfaces viewed more index and brief screens, and fewer full screens, than did those using the standard displays (χ^2=7.64, p=.02190).

Chi-squared tests on demographic variables were all significant. Men viewed fewer brief screens and more full screens than did women (χ^2=16.98349, p=.00021). Daily computer users viewed more index screens and fewer full screens than did others (χ^2=30.85779, p=.00003). Those who said they used library catalogs 2 to 4 times a week viewed more full screens than did those with less catalog use (χ^2=31.57513, p=.00002). Participants who knew nothing about big band music, or had only heard of it viewed fewer full screens than did those more familiar with the topic; in addition, those who had heard of it viewed more brief screens than did others (χ^2=30.45690, p=.00003).

Comparing Screens 1–10 with Screens 11–20. The frequencies of screens viewed in the first 10 screens were compared to the frequencies of the next 10 screens to determine whether there were any significant differences. Paired-sample t-tests were conducted for each type of screen; in each case, there were significant differences. Participants viewed fewer index screens in the second 10 screens than they did in the first 10 (t=5.03, p=.000), more brief screens (t=-2.31, p=.024), and more full screens (t=-4.19, p=.000).

Analysis of Screen Durations

Overall Duration. The first measure—overall time spent to perform the task—is a broad measure that might provide a rough overview of relative performance. The mean session duration was 595.11 seconds, with a SD of 181.14, and a range from 153 to 984 seconds. When examined by either of the main factors the overall means were virtually identical and not significantly different.

We can also examine the mean per screen duration for each session. Overall, the mean per screen duration was 8.54 seconds with a SD of 2.34 and a range from 3.48 to 18.92 seconds. Means between the groups on the labels factor differed, although this difference was not significant (t=-1.38, p=.172). Nor were the means significantly different when compared on the content factor (t=-.55, p=.583). No significant differences were found for overall mean per screen duration among any of the demographic groups.

Next, each session's data, separated by screen type, was examined. For index screens, the mean was 11.90 seconds with a SD of 4.85 and a range from 5.25 to 39.00 seconds. When examined by screen content, the means were 12.77 and 11.04 seconds for standard and enhanced screens respectively, a difference that was not significant (t=1.63, p=.108). Comparing by layout, the means were not found to be significantly different (t=-.03, p=.977). When examined by library use some variation is seen, especially between users with no experience (13.27 seconds) and once-a-month users (11.12 seconds), but these differences were not significant (f=1.31, p=.3134). Comparisons on other demographic variables were not significant. It is interesting to note that the index screens were identical, regardless of the interface viewed.

For brief screens, the mean was 5.56 seconds with a SD of 2.38 and a range from 0 to 13.76. No noticeable differences were found in the means for each factor; no significant differences were found in the means for demographic variables. For full screens, the mean duration was 7.70 seconds with a SD of 4.76 and a range from 0 to 22.00 seconds. No noticeable differences were found in the means for each primary factor. When examined by level of computer use, a significant difference was found (f=6.2535, p=.0007). Analyses on other demographic variables were negative.

First 10 Screens. The next group of data used for analysis was made up of data from the first 10 screens that each participant viewed. For the first 10 screens, the mean duration was 13.25 with a SD of 4.57 and a range from 3.30 to 30.90 seconds. When compared on the labels factor, the difference was not significant (t=-1.25, p=.214). The difference between groups on the content factor was much smaller and also not significant. When compared by gender, the differences were similar to those found for content.

The mean durations for the first 10 screens varied more when compared by computer use, and the differences were significant (f=4.4088, p=.0064). However, the means were as follows: once a month, 14.02; once a week, 12.72; 2 to 4 times a week, 15.81; every day, 11.88. Thus, the means did not seem to follow any pattern; perhaps this was a result of the limited numbers of participants in the lower experience groups (5 and 9). A Scheffé test run on these variables identified the difference between groups 3 and 4 as the source of significance. On other demographic variables, no significant differences were found.

When mean duration was examined by each screen type, there were no significant differences found when compared by content. Similarly, there were no significant differ-

ences in mean duration by each screen type when compared by labels.

First 20 Screens. The first 20 screens of each session were isolated and analyzed. For the aggregate first 20 screens, there were no significant differences found when comparing mean screen duration by content or by labels. Once again, computer use yielded significant results between groups 3 and 4 (t=3.09, p=.0318). Analysis on other demographic variables yielded no significant results. Comparing each screen type for overall duration and mean screen duration yielded no significant results on either of the primary factors. Analyses of each screen type by the demographic variables were not significant.

Comparing Screens 1–10 with Screens 11–20. When comparing data for screens 11–20 with screens 1–10, we find some interesting differences. The mean for screens 11–20 was 8.39 with a SD of 2.80 and a range from 3.20 to 19.30. This represents a substantial drop in mean duration, from 13.57 seconds. In addition, the standard deviation dropped substantially, from 4.57 to 2.80, which indicates that the variance also dropped substantially. When using a t-test for paired samples, the difference in mean times was found to be significant (t=11.79, p=.000). Moreover, the ranges and histograms in each grouping suggest that the learning process affected primarily the longest times in the range.

Mean durations for each screen type for screens 11–20 were examined. No significant differences were found when comparing index screens on the primary factors. Brief screen mean duration differed for participants viewing standard or enhanced screens; participants viewing enhanced screens spent longer on these screens (enhanced=5.90, standard=4.78). This difference was not significant at the .05 level, but was significant at the .10 level, suggesting that there is a likelihood that significant results would be found in a study designed specifically to test for this (t=-1.93, p=.058). Full screen mean duration was significantly different when compared by content (standard=10.44, enhanced=8.16) (t=2.02, p=.043). Participants using the normal screen interface took longer than those using the enhanced interfaces. There were no significant differences found when compared by the labels factor.

First 5 Screens of Each Type. The first 5 screens of each type were separated from the main body of data in order to isolate the performance differences that participants might have had in relation to the different screens. The mean duration of time spent on all of these screens was tested for significance against each of the factors and the demographic variables. There were no significant differences when the primary factors were examined. Significant differences were found between daily users of computers and those using them 2–4 times a week (f=4.9600, p=.0034). When examining each type of screen separately, no significantly different results were obtained on the primary factors. Significant differences were found when comparing full screen duration to the computer use variable (f=6.9444, p=.0003). Otherwise, no significant differences were found.

Duration by Screen Length. One final analysis was conducted using a measure based on the number of characters viewed per second. These data were calculated by measuring the number of characters on each screen viewed, and dividing by the number of seconds spent on the screen. Overall, the mean screen length was 531.055 characters, and the mean screen speed was 109.68 characters per second. The mean index screen speed was 167.97 characters per second; the mean brief screen speed was 67.97 characters per second; the mean full screen speed was 87.75 characters per second. Neither of the primary factor groups differed significantly when compared on the basis of these measures. Similarly, most demographic variables yielded no significant differences. When compared by gender, however, a significant difference was found; males processed more characters per second than did females (t=2.04, p=.045).

Analysis of Selection Data

Data regarding the items that each participant selected were compiled and examined. Of particular interest in this area was the type of screen from which the participants made their selections. Overall, the participants made 809 selections out of a possible 820; one participant selected only 8 items, while a second selected only 1 item. Participants selected 71 items directly from index displays, 390 from brief record displays, and 348 from full record displays. When chi-squared tests were conducted against the two factors, a significant difference was found between groups on the content factor (χ^2=8.93966, p=.01145). Participants using the enhanced interface made fewer selections from index screens and full screens, and more selections from brief screens.

No significant differences were found with the labels factor. Significant differences were found between groups on computer use (χ^2=26.31462, p=.00019) and library use (χ^2=55.07852, p=.00000). On both measures, novice users were more likely to select items directly from index screens than experienced users. On computer use, experienced users were also less likely than all other groups to use index screens to make selections. On library use, the most experienced participants were more likely to use full screens to make their selections than were all other groups.

Tests on gender, knowledge of big band music, and knowledge of Duke Ellington were not conclusive.

Task 2: Judging Relevance

Overview

The second task in the experiment required the participants to provide basic relevance ratings for 20 items on the same

topic as in the first task, along with an indication of the fields that they used to make their relevance judgment. In this task, participants were presented with either a full or brief record in the display mode for their session (i.e., labeled or not, enhanced content or not), and were asked to decide whether they would choose to examine this item further. Participants could choose:

1. that they would examine this item on the shelf;
2. that they would NOT examine this item on the shelf; or,
3. that they could not decide from the information provided.

If they chose either 1 or 2, they were required to indicate which data elements enabled them to make their decision, by checking boxes next to the data elements that they used to make their decision.

The data for this task thus consist of two broad categories: the relevance rating, and the fields used in decision making. The relevance rating was required, so each session had a total of 20 relevance ratings. For the fields used in the decision-making process, it was necessary (for statistical purposes) to determine the total frequency that each field was viewed by each participant. This was a result of the variability of MARC records. This frequency count differed for each interface, because the brief screens differed. For all the data gathered in this task, chi-squared tests were used.

Analysis of Ratings

Participants viewed a total of 1,640 bibliographic records in either brief or full display. They checked 999 items (60.91%) as being useful, 390 items (23.78%) as not useful, and 251 items (15.30%) as not providing enough information to make a decision. The differences in frequency between groups on the content factor were significant (χ^2=48.54232, p=.00000). These results occurred in large part because of the high number of Not Useful entries by the enhanced, labeled interface, as well as a combination of high and low counts among the groups in the Don't Know category. What is more important is that the high frequencies for this category were observed in the standard content interfaces, while the low frequencies were observed in the enhanced content interfaces.

Further analysis of the data revealed that the ratings were linked to the type of screen that was viewed. When comparing the screen type to the relevance rating, a significant difference was found (χ^2=60.21908, p=.00000). This difference was due to the effect of the high rate of Don't Know classifications for brief screens.

Secondary analysis was conducted on the data set, in which the Don't Know category was combined with the Not Useful category. Contrary to results received when these categories were separate, no significant results were obtained when these categories were conflated.

Analysis of Fields Checked

The final group of data to be analyzed in the second task consisted of the check rates for each of the fields that participants viewed. Participants checked 2,058 data elements as being useful to their decision, a rate of 1.48 items per selection. The rates at which participants checked particular fields as useful ranged from a high of 70.73% for the summary field (MARC 520) to 0% for several fields. The title field had the second highest check rate at 60.55%, followed by the subjects field at 50.8%. The next group of fields was checked at a much lower rate, and consisted of the series (13.72%) and general notes (12.56%) fields. Four additional fields were checked at rates near 5%, and 5 more at rates around 1.5%. Four fields were not present in the sample.

One thousand six hundred and nineteen (78.67%) of the checks were made by participants in one of two fields: the title field (MARC 245), and the subjects field (MARC 6xx). Overall, significant differences between factors were found for the following MARC fields: 300, 4xx, 6xx, and 090. Field 300 was significant (χ^2=30.35, p=.000) in large part because of an extremely high check rate for the standard labeled display; secondarily, a low check rate in the standard unlabeled interface also contributed. Field 4xx was significant (χ^2=17.30, p=.001) primarily because there were no checks at all in this field for the enhanced unlabeled interface. Field 6xx was significant (χ^2=10.86, p=.002) because of a high check rate in the enhanced labeled display and a low check rate in the standard unlabeled display. Field 090 was significant (χ^2=10.81, p=.002) because of variance found in all 4 interfaces. This result might not be reliable because of low overall frequencies.

When examined by demographic variables, many similar results were obtained. Using computer use as the grouping variable, significant results were found for the Not Useful and Don't Know categories, and fields 300 and 6xx. Using library catalog use as the grouping variable, significant results were found for categories Not Useful and Don't Know, and fields 245, 300, 6xx. In this instance, field 245 was significant in large part because of the low check rates by the most experienced library catalog users. When using knowledge of big band music, significant results were found for the Not Useful category, and fields 245, 300, and 6xx. Field 245 was significant here because of a high check rate by those with no knowledge of big band music. Significant differences were also found for fields 020 and 090, but the individual cell counts were too low for conclusions to be drawn.

Summary of Statistical Analysis

The statistical analysis was presented in two broad categories: analysis of data from the first task, and analysis of

data from the second. For the first task, participants using enhanced brief screen interfaces viewed more brief screens and fewer full screens than did their counterparts. Screen durations for the second 10 screens were found to have dropped from those of the first 10 screens. Statistical analyses comparing demographic variables to the screen frequencies uncovered many significant differences. Participants using the enhanced content interfaces made fewer selections from index and full screens, and more selections from brief screens. For the second task, participants who used enhanced content interfaces were able to make some sort of relevance judgment more frequently than were those who used standard content interfaces. In the next section, the importance of these results will be discussed, with specific regard to the research questions and hypotheses.

Discussion

Research Question 1: Screen Content

The first research question asked whether there was a correlation between screen content and the time it takes a participant to perform a task. Three hypotheses were derived from this broader question. The first was that the number of screens viewed would differ between treatment groups. The second was that the number of screens viewed at each level of specificity would differ between treatment groups. The third was that the screens used to make selections decisions would differ between treatment groups.

Hypothesis 1.1

In this experiment, screen content was varied between treatment groups so that some groups viewed more topically oriented data on the brief screen than did other groups. Frequency data were tabulated for all participants in several categories—by the main factors in the experiment, by the demographic variables, and broken down by type of screen—and statistical tests were run. There were no significant differences between the groups in the overall number of screens that participants viewed on either the content factor or the labels factor. Thus the null hypothesis, which states that there is no difference in the number of screens viewed, cannot be rejected. Therefore, the first hypothesis—that there will be a difference in the overall number of screens viewed—cannot be supported.

Hypothesis 1.2

When considering the second hypothesis for this question—i.e., whether there were differences in the number of individual screen types examined—significant differences were found. It was found that participants who used the interfaces with enhanced content brief screens viewed more brief screens and fewer full screens overall than did their counterparts. The enhanced content of the brief display enabled participants to avoid viewing full displays. Given that full displays contain more data, more data elements, and by all measures require more time to process, a reduction in the number of full display screens viewed most likely represented a reduction in the effort that users expended to achieve their goals. Thus, the null hypothesis in this case was rejected; the number of screens viewed at each level of specificity did differ based on whether participants used enhanced content brief displays or not.

It should be noted that these differences were not observed when the first 10 screens were examined in isolation. This finding suggests that the participants altered their behavior during the course of the test. That behavior had changed by the twentieth screen, however, as the frequency data for the first 20 screens already show a significant difference between the groups based on-screen content.

It is also interesting to note that participants using labeled displays viewed more index screens and fewer full screens for the first 10 and first 20 screens than did those using unlabeled displays. This contradicts results for the labels factor that were obtained from the overall data, where there were no significant differences. This suggests, in turn, that users of labeled interfaces initially preferred viewing index screens to single record displays. That the absence or presence of labels would affect a participant's screen selection—even for a short period of time—was not an anticipated outcome.

The behavior change noted above is also reflected when both the frequencies and the durations of the first 10 screens are compared to the next 10 screens. The participants viewed fewer index screens and more brief and full screens in the second 10 screens than they did in the first 10. This might reflect the fact that many participants appeared to use the first few screens to familiarize themselves with the data set and with the system. Or it might be that they found the brief and full screens more informative for their needs. One way or the other, a second change in activity is seen in the fact that mean screen duration dropped a significant amount in all types of screen, as well as for all screen types taken together. These drops suggest that the participants became familiar with the system, its content, and how to use both to solve the problem at a relatively rapid rate.

It is possible that these drops—both statistically significant and substantial—are the result of some other factor, such as that the participants became dissatisfied with the system and simply hurried through the task to be done. There is some evidence in the post-test comments that suggest that this might be the case, inasmuch as some participants commented that there was not enough information in the system for them to make the types of decision that were

requested of them. However, it would appear from the transaction logs that the participants figured out their preferred method of solving the problem within the first 20 or 30 screens, and then used that method with increasing efficiency as the problem progressed. The participants seemed to develop routines that they followed throughout the experiment—for example, one participant would view an item in the brief display, then the long display, and finally select that item. It is assumed that in adopting such a routine, users can work more quickly because they gain practice and experience in carrying out the steps in the task.

Hypothesis 1.3

Because the problem presented to the participants was subject-oriented in nature, it was thought that the use of enhanced content brief displays would have an effect on the selection behavior of the participants. The third hypothesis for research question 1 asked whether the screens from which participants made selections would differ based on the experimental factors. Specifically, it was thought that those who used an enhanced content interface would be more likely to make their decisions from brief record displays. Statistical analysis of the selections that participants made shows that this in fact occurred—users of the enhanced content interfaces selected 217 items from brief displays while their counterparts selected only 173. Confounding this is the fact that users of standard content interfaces selected a larger number of items from index screens. This might have been skewed by the actions of the one or two participants who made all their selections from index screens. Despite this anomaly, the null hypothesis was rejected, and the hypothesis confirmed; the screens used to make selection decisions did differ between treatment groups.

Additional support for this conclusion can be found in the data from the second task, where participants were asked to make basic relevance judgments for 20 items presented to them either in brief or full display. Users of the enhanced content interfaces checked that they could not decide the value of an item significantly fewer numbers of times than users of the standard interfaces. The decrease in the Don't Know category was largely transferred to the Not Useful category, which suggests that the increased information helped participants make negative selection decisions.

It should be noted that the secondary analysis—in which the Don't Know and Not Useful categories were combined—confounds these conclusions to a degree. These categories were combined under the assumption that a user's indecision would most likely translate into a decision not to examine an item further, and the results were not significant. However, this assumption might not be valid; in an active environment, the user might choose any of a number of different steps to follow up on this state of indecision—request more information, browse the shelf, etc. It is quite possible that these two categories cannot be conflated.

Demographic Effects

Further complicating the findings of this study are the significant results obtained on many of the demographic variables. Most noteworthy of these results are the consistent significant differences found when examining the results using computer use as a grouping variable. More experienced computer users generally viewed fewer detailed screens. This might be explained by the fact that these users were less likely to accept the system limitations—both in terms of content and function—than less experienced users. Computer users have become more demanding as the capabilities of computers have grown, and it seems reasonable to imagine that they would be demanding in this experiment. Several participants commented on the system limitations; for example, some complained that there was not enough information in the system to make a decision, while others complained at having to use the keyboard to issue commands.

In addition, gender had a significant effect on several of the measures. The most interesting of these was the difference that was found in the frequency that males and females viewed different screen types. Women viewed fewer index screens and more brief screens than did men. These results bring to mind the study conducted by Ford and Ford (1993), where the researchers found that experiment participants used three broad search strategies that were associated with one gender or the other. The results of this study serve to confirm those findings, because women in this experiment tended to view records at a greater level of detail than did men.

The third variable with noticeable effect on the results was the library use variable. Those with higher library catalog experience were more likely to view full display screens than those with less library catalog experience. This suggests that experienced library users are more likely to exploit the resources in a catalog more fully, because they are more likely to examine the fuller information. Finally, knowledge of the topic was linked to the likelihood that the participant would view fuller information.

It must be pointed out that tests to determine whether the main treatment groups differed on any of the demographic variables recorded were negative in all cases except in the instance of the knowledge of big band music. It was concluded, therefore, that these demographic variables did not have a significant or systematic effect on the tests of the primary factors. With regard to the knowledge of big band music variable, only two participants said they were familiar with big band music, which makes it unlikely that their actions would adversely affect the validity of the results.

Research Question 2: Screen Layout

Research question 2 asked whether the layout of bibliographic data on-screen would affect the speed with which a participant could extract information from it. Two hypotheses were derived from this broader question. The first was that the time needed to complete the task would differ between treatment groups. The second was that the time that participants spend on specific screen types would differ between treatment groups.

For this experiment, layout was defined as the absence or presence of on-screen labels delineating the data elements of the bibliographic record. The data gathered that address this question consist of the transaction logs and the screen duration information derived from those logs. While previous researchers have found that labeled displays are easier for users to read (Tullis 1983; 1988), no such assumption was made for this experiment; thus, the t-tests that were performed were two-tailed.

Hypothesis 2.1

To test whether screen layout had an effect on the time it took to perform the task, the data were examined from several different perspectives. First, overall session duration was compared against the labels factor. This test failed to reveal significant differences. Second, the mean durations for all screens in each session were compared, with no significant differences observed. Additional analyses of the effect of labels on-screen duration included: mean duration of the first 10 screens viewed and mean duration of the first 20 screens. No significant differences were found. Thus, the null hypothesis cannot be rejected, and the hypothesis that screen layout had an effect on participant performance cannot be supported.

Hypothesis 2.2

To test whether screen layout had an effect on the time that participants spent on particular screen types, a number of tests were conducted. First, the overall mean screen duration for each screen type was compared, with no significant differences found. Additional analyses of the effect of labels on-screen duration included: mean duration of the first 10 screens viewed, separated by screen type; mean duration of the first 20 screens by screen type; mean duration of the first 5 screens of each screen type; and, the number of characters processed per second. In all these tests, no significant differences were found. Thus, the null hypothesis here cannot be rejected; screen layout had no effect on-screen durations for particular screen types.

With regard to the characters per second measure, it should be noted that while overall screen density was used as a measure, it is quite possible that participants only processed part of the information on-screen. Thus, a participant might only have read the title and subject data elements on a given screen, rendering the total screen measure less meaningful.

Demographic Variables

While layout of the data on-screen had no measurable effect on the ability of participants to perform the task presented to them, other factors did. The level of computer use affected participant performance in the first 10 screens, the first 20 screens, and the first 5 screens of each type, although these results are clouded by the fact that the differences in each case failed to follow a single progression. This might have been due to the uneven distribution of individuals over the computer experience variable, as there were so few participants who had little experience with computers. Consequently, it is not clear how these data might be interpreted. The complex nature of the values for mean screen duration here suggest the possibility that some other variable interacted to create these results.

However, even this difference disappeared when the length of individual screens was introduced as a control. Using this more exacting measure of speed, the only significant difference was found on the gender variable—males processed more characters per second than did females. This result might have been due in part to the fact that males viewed significantly more index screens than did females. Index screens had a higher mean processing speed than either of the other two screens; higher levels of index screen viewing would thus result in higher overall processing speed.

Based on the data gathered in this experiment, there is no evidence that screen layout has an effect on user performance. Neither of the two hypotheses under this question were supported, regardless of the method of measuring this difference. This conclusion is remarkable, given the body of human factors research that has been conducted that contradicts this (c.f., esp., Tullis 1981).

One possible explanation for these results is that it is possible that the design of the experiment was not sophisticated enough to obtain wholly reliable results; for example, it is possible that some participants just selected the first records they viewed simply to be finished with the test. Examination of the transactions logs, however, suggests that most participants made at least a modest effort at carrying out the task. For example, while the frequency distribution of screens viewed for each record in the set shows a general decreasing trend as the record number increases, there still remains a wide dispersion of activity.

If large numbers of participants failed to give some effort to the experiment, it seems reasonable to expect that

extremely high view rates for the first and last records in the list would have been observed—those closest to the starting point of the task. This was not the case. Even if it were, the fact that some participants chose not to expend a great deal of energy trying to select items from the display does not necessarily invalidate the results—especially given the fact that users often do not seem to be extremely diligent when examining catalog search results (e.g., Millsap and Ferl 1993).

Another explanation for the lack of significant results in this area might be that beyond a certain point the complexity of layout has no effect. Bibliographic displays include a large amount of data arranged in a complex set of data elements, the combination of which can present a great deal of information to the user. In this experiment, the mean display length was 531 characters out of a possible 910—a 58% screen density level. Most researchers recommend screen densities more in the neighborhood of 30 to 40%, a much lower figure. Added to this is the fact that many of the data elements in a bibliographic display are neither clearly explained to users nor inherently known by them, minimizing the effect that the absence or presence of labels might have. In this experiment, every attempt was made to use clear and unambiguous terms for the labels; however, it is not clear to what extent the participants knew and understood the meaning of the information presented to them. These complexities could have overwhelmed the effect that labels might have had on participant performance.

Research Question 3: Relative Value of Data Elements to Relevance Judgments

Research question 3 asked whether there were particular data elements that participants used more often than others to make relevance decisions in topically oriented retrieval tasks. Two hypotheses were derived from this broader question. The first was that those data elements most associated with topical bibliographic data would be selected most frequently by participants of the assigned task. The second was that the fields deemed useful would differ between treatment groups.

Hypothesis 3.1

One purpose behind including this question was to support or refute past research that identified the fields considered helpful by library catalog users—notably Seal (1983), Hufford (1991), and Palmer (1972). Those researchers found that the substantial majority of users considered just a few fields to be necessary or useful for their needs. In this experiment, this hypothesis relied on the subjective feedback provided by the participants in the second task of the experiment. More than three quarters of the checks that participants made were placed in 2 fields: the title field (MARC 245) and subjects field (6xx).

This represents an even higher concentration of interest than found in those earlier studies, but might be explained by a couple of factors. First, the problem in this experiment was highly subject-oriented; participants were asked to select items on a given topic. This might have caused the participants to focus on those fields that had the most topical content. This can be further supported by the fact that the summary field (MARC 520), while not present in many records (82 chances overall), was selected by participants an extremely high percentage of the time that it was present (70.73%).

Second, the nature of the experiment limited the chances that participants had to select other fields. The randomly selected records and randomly selected display screens emphasized primary fields and downplayed secondary ones. The title field was on every screen presented to the participants, and the subjects field in most. In contrast, the contents field—among other potentially useful fields—was not viewed at all. This uneven representation probably affected the relevance indications of participants.

A third reason for this concentration might have to do with the fact that the experiment was a simulation. One of the fields that previous researchers said that library users found important was the call number field. Obviously, without the call number, locating the actual item on the shelf is greatly complicated. In this experiment, however, participants were not asked to retrieve the items they selected; thus, their perceived need for the call number may well have been mitigated. It is perhaps instructive to note that some participants selected call number despite the nature of the experiment. This might reflect an effort on the part of those participants to take the simulation seriously, or it might be a retrieval strategy that they use.

The check rates of four other fields differed from the results of earlier research. Both the author and publication fields in this experiment were selected as useful only a small percentage of the time—at rates only half that found for the series field. On the other hand, the series and general notes fields were fields that were not identified as important in earlier research, but that were checked by participants at a low but steady rate. It is unclear why these results were obtained.

The evidence regarding useful fields that was gathered in this experiment seems to suggest that novice library users consider only a few fields to be helpful to their selections decisions, and ignore the rest. In addition to the explicit check rates in the second task are the transaction logs in the first task, which indicated that novice library catalog users viewed more index screens and fewer detailed screens than more experienced library catalog users. One possible explanation for this behavior is that novice users, unfamiliar with

the data elements in a bibliographic record, focused on the simplest displays, where the data were presented in the most uniform and unambiguous manner. Detailed screens can contain any of a wide variety of data elements, many of which are identified by terminology that is unclear to the user (e.g., "LCCN," "ISBN," "Music Number," etc.). Avoiding these screens might reduce the user's uncertainty and unease in using the system.

Hypothesis 3.2

With regard to whether the fields deemed useful differed between treatment groups, the significant results obtained for several of the fields partially support the hypothesis. When these results are considered in relation to the core fields identified by earlier research, it was found that both the subjects and call number fields differed significantly between treatment groups. Looking first at the call number results, it was noted that the check rates in each interface diverged from the expected rates, and that these discrepancies crossed both treatment factors. It might be that the low overall check rate makes these data unreliable in this context.

Turning to the check rates in the subjects field, participants using the enhanced content labeled display selected subjects at an especially high rate, while users of the unlabeled standard content display selected this field fewer times than expected. Finally, with both remaining significant results, it is interesting to note that the low check rates in unlabeled displays were primary contributors to the result. In addition, a high number of checks in the labeled display contributed to the results for the physical description field.

Taken together, the significant results for subjects, physical description, and series fields follow a pattern: participants who did not use interfaces with identifying labels indicated that they found these data less useful than those viewing labeled displays. While these results do not form a compelling argument to support the use of labeled displays in all cases, they certainly suggest that in some cases labels affect user perceptions of the data they see. Thus, while the absence or presence of labels did not have a significant effect on the speed with which participants performed the first experiment task, the results in the second task still might affirm their importance to end users.

Summary of Discussion

This discussion of the statistical tests included an analysis and comparison to the research questions and hypotheses. On research question 1—whether screen content affected user behavior—hypothesis 1.1 was rejected, while hypotheses 1.2 and 1.3 were confirmed. With respect to research question 2—whether screen layout affected user performance—hypothesis 2.1 and 2.2 were both soundly rejected. For research question 3—whether some data elements are more commonly considered useful than others—hypothesis 3.1 was supported, while the results for the hypothesis 3.2 are more ambiguous. Participants in the treatment groups did select some data fields as useful at different rates; however, these could at best be considered tentative results.

Conclusions

This research began with a desire to contribute to the knowledge about online catalog use. The primary objective was to explore the effect that layout and content might have on a typical information retrieval task. An experiment was designed to isolate these aspects of online catalogs and the experiment was run with participants at the University of Pittsburgh. The data gathered in this experiment were analyzed and the results were presented.

In this research, several important discoveries have been made. First, it was found that alterations to the content of particular screens—in this case, enhancing the content of the brief display—had a significant effect on the behavior of the experiment participants. Participants who used brief display screens that contained more topically oriented data elements resorted to full display screens significantly fewer times than did those using standard, citation-oriented brief displays. This finding is important because it suggests that—for topically oriented tasks, at least—brief catalog displays might be redesigned to include more fields with subject-rich content (e.g., MARC 520 and 6xx fields). Such a redesign would presumably reduce both the number of screens that users would need to view and the complexity of those screens. This might, in turn, simplify the user's task of finding wanted items in the catalog.

A second important finding of this research was that the layout of information on-screen had little effect on the time it took experiment participants to complete the assigned task. Whether measured as overall time spent to perform the task or as average time spent on particular screens, no significant differences between treatment groups were found. When combined with the fact that mean screen duration dropped substantially over the course of the first 20 screens, it seems likely that the participants of the experiment adapted to the mode of information presentation rapidly, and became equally comfortable with that mode of presentation—regardless of the specifics of the mode of presentation. These results contradict earlier related research, and raise questions about whether bibliographic data somehow differ from the data used in that earlier research.

A third important finding was the support for earlier research regarding the fields that participants felt were use-

ful to their judgments of relevance. Authors of earlier studies had found that most users consider only a few fields important in the bibliographic record. The results of this study confirm these findings, and take those results further with regard to topically oriented catalog tasks. The participants in this study overwhelmingly considered three MARC fields to be useful to their judgments of relevance—a fewer number of fields than was found in earlier studies. This discovery is important because it might be used by online catalog designers to select fields for inclusion on given catalog screens. This would give users more of the data that they feel are useful, without cluttering screens with data they feel are not.

Method Considerations

During the course of this research, two broad methods of conducting the experiment were tried. The first of these employed HTML and the Hypertext Transfer Protocol as the mechanism for carrying out the experiment. Several observations can be made with regard to this system. First, although there were very few problems with participants using a mouse to navigate the system, one problem that was observed was that many had trouble with bibliographic records that were long enough to require scrolling. They did not know how to view the bottom of the record. It is unclear why these participants had these troubles, but that these troubles existed should serve as a point of consideration for others doing such tests.

A second observation is that data-gathering from participants in the HTML environment requires access to server-side processing. Without server-side control of the flow of the experiment, participants can ignore requests for information posed to them. In the pilot test, 40% of the requests for participant feedback were ignored. While this response rate was due in part to the design of the particular experiment, a good portion no doubt resulted from the inability to enforce compliance. Use of a dedicated HTTP server with access to server-side control would have enabled the pilot test to have more success in this area; with the advance in technologies, this approach would be more feasible now.

Finally, extreme care must be taken in any such test to ensure that the variable being tested is not confounded by other related variables. In designing the pilot test, the variables were tested simultaneously and interacted in such a way as to make reliable measure practically impossible. It was assumed that participants would proceed sequentially through the screens presented to them, using one screen to examine the bibliographic information, and the next to provide feedback regarding the data elements used for their relevance judgment. In fact, the participants became familiar with the interface and often appeared to anticipate several screens in advance. This enabled them to ignore certain screens and focus on others, thus rendering the duration of screen measure meaningless in that particular experimental context.

The second method of conducting the experiment used a custom-made Visual Basic application, and was generally more reliable in performance. There were some problems, however. Whereas the first task relied exclusively on keyboard input, the second relied on mouse actions, and this change in system modes caused some confusion among the participants—even after the instructions had been modified to emphasize this change.

Future Research

A great deal of further research still needs to be conducted on the effect that layout of data on-screen has on user performance. While the research outlined here showed that no differences were found in performance based on the absence or presence of data labels, many other tests could and should be conducted. For example, the effect that overall screen density has on usability should be explored with respect to bibliographic data. In fact, many studies that have been conducted with other types of information systems (e.g., Kruk and Muter 1984; Marcus 1982; Trollip and Sales 1986; Tullis 1981, 1983) should be replicated in the special context of bibliographic data.

A number of other areas for further research have been uncovered. One such area is the effect that user knowledge of bibliographic data has on user behavior. In the current study, participants were asked to identify the data elements that they used to make basic relevance judgments. Left unanswered is the question of how the individual participant's knowledge of bibliographic data content might have affected the results. It seems reasonable to assume that users who know more about the content and structure of an information system will be better able to exploit the information in that system. Indeed, Brunner et al. (1992) found some evidence to this effect. Similarly, the selection of data elements used in a basic relevance judgment might depend on the user's knowledge of bibliographic data structures. Research must be conducted in which the effects that this knowledge has on user behavior and on user relevance judgments are explored. Some research into the effect that user knowledge—both of topic and of system function—can have on user behavior has already been conducted (e.g., Hsieh-Yee 1993), but additional research focused directly on layout and content needs to be conducted.

Linked to this study of users' knowledge of bibliographic data structures would be research into the labels that are used in online catalogs. Bibliographic data contain a large number of data elements. The labels that have been selected for these data elements have not been tested with users to see how clearly they communicate content and meaning to users. Research into the effect that different labels have

on user comprehension would enable catalog designers to use terminology with the widest recognition among users.

Another area of research that needs further exploration is determining the data elements that are most important for particular types of retrieval situations. This research would have to start with a greater consideration of the types of tasks users typically come to the online catalog to solve. While it might be obvious that many users come to the catalog to find out what a library has on its shelves, other uses are evident, if less well defined (e.g., finding all the editions of a particular work, in order to identify the most authoritative edition). Anecdotal evidence regarding the uses to which catalogs are put has long been offered; what is needed are studies that offer substantive, empirical data regarding these uses. Such knowledge is necessary before the library and information science field can act to improve catalogs in a meaningful and purposive manner.

Once a basic topography of the types of retrieval situations and their relative frequencies has been achieved, research into the data needs for each of these situations should be conducted. Common sense would suggest that certain data elements would be more useful in some situations than in others, and the research described here confirms that assumption in one particular instance. Moreover, it might be that the usefulness of data elements follows the same skewed patterns found in other areas of library and information science research. Using the frequencies of occurrence as a means of setting priorities, library and information science professionals can systematically begin to evaluate the typical data needs in each retrieval situation. Such research would make it possible to design online catalogs with basic screens that more closely reflect the needs of the user in a particular retrieval situation.

Research in this area might also be broadened to consider such basic display issues as the order of display of data elements on-screen. While online catalog displays have taken as their model the catalog card—a familiar, comfortable display—it is certainly possible that a different ordering of data elements on-screen might be more quickly and easily processed. Research should be conducted to determine the effects that data order might have on user performance.

It might be possible to use the results of these studies as a basis for redesign of online catalogs not just from a display perspective, but also from a functional perspective. With a clearer understanding of the types of retrieval problems with which users approach online catalogs, indexes and search modes could be reorganized to match those problems. For example, instead of indexing a bibliographic database using MARC fields as the focus, the indexing might be designed so that related data elements are joined in one index. To some extent, library systems designers have begun to create such indexes, as when they create keyword indexes that combine title and subject fields. With a better understanding of the types of fields that users utilize in given retrieval situations, this process can be accelerated, and indexes can be designed that truly match the user's information need.

It is necessary for additional research to examine the relationship between gender and online behavior from a number of perspectives. The current study seems to corroborate the findings of Ford and Ford (1993), who found behavioral differences between the genders in online text use. Studies that specifically test the effect that gender has on the types of bibliographic information deemed useful in given situations must also be conducted.

Works Cited

Akeroyd, J. 1990. Information seeking in online catalogues. *Journal of Documentation* 46: 33–52.

Allen, B. 1994. Cognitive abilities and information system usability. *Information Processing & Management* 30: 177–91.

Allen, B., and G. Allen. 1993. Cognitive abilities of academic librarians and their patrons. *College & Research Libraries* 54: 67–73.

Aykin, N. M., and T. Aykin. 1991. Individual differences in human-computer interaction. *Computers and Industrial Engineering* 20: 373–79.

Bates, M. J. 1986. Subject access in online catalogs: A design model. *Journal of the American Society for Information Science* 37: 357–76.

———. 1989. Rethinking subject cataloging in the online environment. *Library Resources & Technical Services* 33: 400–12.

Beheshti, Jamshid. 1992. Browsing through public access catalogs. *Information Technology & Libraries* 11: 220–28.

Belkin, N. J., P. G. Marchetti, and C. Cool. 1993. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing & Management* 29: 325–44.

Belkin, N. J., and A. Vickery. 1985. Precursors to information seeking behavior: "Information need." In *Interaction in information systems*. London: British Library.

Benbasat, I., and P. Todd. 1993. An experimental investigation of interface design alternatives: Icon vs. text and direct manipulation vs. menus. *International Journal of Man-machine Studies* 38: 369–402.

Borgman, C. L. 1986. Why are online catalogs hard to use? Lessons learned from information retrieval studies. *Journal of the American Society for Information Science* 37: 387–400.

———. 1988. Human factors in the use of information systems: Research methods and results. In *Information research: Research methods in library and information science*, edited by Neva Tudor-Silovic and Ivan Mihel. London: Taylor-Graham.

Brajnik, G., G. Guida, and C. Tasso. 1987. User modelling in intelligent information retrieval. *Information Processing & Management* 23: 305–20.

Brown, C. M. 1988. *Human-computer interface design guidelines*. Norwood, N.J.: Ablex.

Brunner, H. et al. 1992. An assessment of written/interactive dialogue for information retrieval applications. *Human-Computer Interaction* 7: 197–249.

Buckland, M. K. 1991. *Information and information systems.* New York: Praeger.

Buckland, M. K., B. A. Norgard, and C. Plaunt. 1993. Filing, filtering, and the first few found. *Information Technology & Libraries* 12: 311–19.

Cochrane, P. A., and K. Markey. 1983. Catalog use studies—since the introduction of online interactive catalogs: Impact on design for subject access. *Library and Information Science Research* 5: 337–63.

Coll, J. H., R. Coll, and R. Nandavar. 1993. Attending to cognitive organization in the design of computer menus: A two-experiment study. *Journal of the American Society for Information Science* 44: 393–97.

Crawford, W. 1987. *Patron access: Issues for online catalogs*. Boston: G. K. Hall.

Crawford, W., L. Stovel, and K. Bales. 1986. *Bibliographic displays in the online catalog*. White Plains, N.Y.: Knowledge Industry.

Cutter, C. A. 1904. *Rules for a dictionary catalog.* 4th ed. Washington, D.C.: GPO.

Davis, C. H., and D. Shaw. 1989. Comparison of retrieval system interfaces using an objective measure of screen design effectiveness. *Library and Information Science Research* 11: 325–34.

Dervin, B., and M. Nilan. 1986. Information needs and uses. In *ARIST* 21, edited by Martha Williams. White Plains, N.Y.: Knowledge Industry, 2–33.

Dumas, J. S. 1988. *Designing user interfaces for software*. Englewood Cliffs, N.J.: Prentice Hall.

Fidel, R. 1984. Online searching styles: A case study-based model of searching behavior. *Journal of the American Society for Information Science* 35: 211–21.

Ford, N., and R. Ford. 1993. Towards a cognitive theory of information processing: An empirical study. *Information Processing & Management* 29: 569–85.

Foss, C. L. 1989. Tools for reading and browsing hypertext. *Information Processing & Management* 25: 407–18.

Frei, H. P., and J. F. Jauslin. 1983. Graphical presentation of information and services: A user-oriented interface. *Information Technology: Research and Development* 2: 23–42.

Galitz, W. O. 1989. *Handbook of screen format design*. 3d ed. Wellesley, Mass.: QED Information Sciences, Inc.

Gregor, D., and C. A. Mandel. 1991. Cataloging must change! *Library Journal* 116 (April 1): 42.

Hancock-Beaulieu, M., S. Robertson, and C. Neilson. 1991. Evaluation of online catalogues: Eliciting information from the user. *Information Processing & Management* 27: 523–32.

Harman, D. 1992. User-friendly systems instead of user-friendly front-ends. *Journal of the American Society for Information Science* 43: 164–74.

Hensley, R. 1991. Learning style theory and learning transfer principles during reference interview instruction. *Library Trends* 39: 203–9.

Hildreth, C. R. 1982. *Online public access catalogs: The user interface*. Dublin, Ohio: OCLC.

Hsieh-Yee, I. 1993. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science* 44: 161–74.

Hufford, J. R. 1991. Elements of the bibliographic record used by reference staff members at three ARL academic libraries. *College and Research Libraries* 52: 54–63.

International Conference on Cataloging Principles. 1963. *Report: International Conference on Cataloging Principles, Paris, 9th–18th October, 1961*. London: Organizing Committee of the International Conference on Cataloging Principles.

Kerr, S. T. 1990. Wayfinding in an electronic database: The relative importance of navigational cues vs. mental models. *Information Processing & Management* 26: 511–23.

Kiger, J. I. 1984. The depth/breadth trade-off in the design of menu-driven user interfaces. *International Journal of Man-Machine Studies* 20: 201–13.

Kruk, R. S., and P. Muter. 1984. Reading of continuous text on video screens. *Human Factors* 26: 339–45.

Larson, R. R. 1991a. Between Scylla and Charybdis: Subject searching in the online catalog. *Advances in Librarianship* 15: 175–236.

———. 1991b. The decline of subject searching: Long term trends and patterns of index use in an online catalog. *Journal of the American Society for Information Science* 42: 197–215.

Leazer, G. H. 1993. *A conceptual plan for the description and control of bibliographic works*. Ph. D. diss., Columbia University.

Mandel, C. A. 1985. Enriching the library catalog record for subject access. *Library Resources & Technical Services* 29: 5–15.

Marchionini, G. 1992. Interfaces for end user information seeking. *Journal of the American Society for Information Science* 43: 156–63.

Marcus, A. 1982. Typographic design for interfaces of information systems. In *Human factors in computer systems.* New York: Association for Computing Machinery.

Marcus, A., W. Cowan, and W. Smith. 1989. Color in user interface design: Functionality and aesthetics. *ACM SIGCHI Bulletin*: 25–27.

Markey, K. 1985. Subject-searching experiences and needs of online catalog users: Implications for library classification. *Library Resources & Technical Services* 29: 34–51.

Martin, T. H. 1974. *A feature analysis of interactive retrieval systems*. Stanford, Calif.: Stanford University Institute for Communication Research.

Matthews, J. R. 1985. Suggested guidelines for screen layouts and design of online catalogs. In *Online catalog screen displays: A series of discussions*. Washington, D.C.: Council on Library Resources.

Millsap, L., and T. E. Ferl. 1993. Search patterns of remote users: An analysis of OPAC transaction logs. *Information Technology & Libraries* 12: 321–43.

Morehead, D. R., and W. B. Rouse. 1982. Models of human behavior in information seeking tasks. *Information Processing & Management* 18: 193–205.

Noerr, P. L., and K. T. Bivins Noerr. 1985. Browse and navigate: An advance in database access methods. *Information Processing & Management* 21: 205–13.

Norman, D. A. 1990. *The design of everyday things*. New York: Doubleday.

Olsen, K. A., et al. 1993. Visualization of a document system: The VIBE system. *Information Processing & Management* 29: 69–81.

Palmer, R. P. 1972. *Computerizing the card catalog in the university library: A survey of user requirements*. Littleton, Colo.: Libraries Unlimited.

Prorak, D., T. Gottschalk, and M. Pollastro. 1994. Teaching method and psychological type in bibliographic instruction: Effect on learning and confidence. *RQ* 33: 484–95.

Rubin, J. 1994. *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: Wiley.

Seal, A. 1983. Experiments with full and short entry catalogues: A study of library needs. *Library Resources & Technical Services* 27: 144–55.

Seymour, S. 1991. Online public access catalog user studies: A review of research methodologies, March 1986–Nov. 1989. *Library and Information Science Research* 13: 89–102.

Shaw, D. 1991. The human-computer interface for information retrieval. In *ARIST* 26, edited by Martha Williams. Medford, N.J.: Learned Information Inc., 155–95.

Shires, N. L., and L. P. Olszak. 1992. What our screens should look like: An introduction to effective OPAC screens. *RQ* 31: 357–69.

Shneiderman, B. 1992. *Designing the user interface: Strategies for effective human-computer interaction*. 2d ed. Reading, Mass.: Addison-Wesley.

Siegfried, S., M. J. Bates, and D. N. Wilde. 1993. A profile of end user searching behavior by humanities scholars: The Getty online searching project, report number 2. *Journal of the American Society for Information Science* 44: 273–91.

Smiraglia, R. P. 1992. *Authority control and the extent of derivative relationships*. Ph.D. diss., University of Chicago.

Thomas, David H. 1997. *The effect of interface design on item selection in an online catalog*. Ph.D. diss., University of Pittsburgh.

Troll, D. A. 1995. What's hot and what's not. *College & Research Libraries News* 56: 236–39.

Trollip, S., and G. Sales. 1986. Readability of computer-generated fill-justified text. *Human Factors* 28: 159–63.

Tullis, T. S. 1981. An evaluation of alphanumeric, graphic, and color information displays. *Human Factors* 23: 541–50.

———. 1983. The formatting of alphanumeric displays: A review and analysis. *Human Factors* 25: 657–82.

———. 1988. A system for evaluating screen formats: Research and application. In *Advances in human-computer interaction*, vol. 2., edited by H. R. Hartson and D. Dix. Norwood, N.J.: Ablex: 214–86.

Yee, M. M. 1991. System design and cataloging meet the user: User interfaces to online public access catalogs. *Journal of the American Society for Information Science* 42: 78–98.