DOCUMENT RESUME

ED 042 479 52 LI 002 098

AUTHOR Lipetz, Ben-Ami
TITLE User Requirements in Identifying Desired Works in a Large Library. Final Report.
INSTITUTION Yale Univ., New Haven, Conn. Library.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-7-1140
PUB DATE Jun 70
GRANT OEG-1-7-071140-4427
NOTE 121p.

EDRS PRICE EDRS Price MF-$0.50 HC-$6.15
DESCRIPTORS Algorithms, Automation, *Cataloging, *Catalogs, Computer Programs, *Information Retrieval, Research Libraries, Search Strategies, *University Libraries, *Use Studies
IDENTIFIERS *Sterling Mmeorial Library

ABSTRACT

Utilization of the card catalog in the main library (Sterling Memorial Library) of Yale University was studied over a period of more than a year. Traffic flow in the catalog was observed, and was used as the basis for scheduling interviews with a representative sample of catalog users at the moment of catalog use. More than 2000 interviews were completed. Data were collected on user objectives and starting clues. Follow-up studies were done on the matches among user clues, catalog card data, and information available in the front matter of cataloged documents. Reasons for search failures were determined. In terms of immediate intent, 73 percent of searches are document ("known item") searches and 16 percent are subject searches; in terms of underlying interest, 56 percent are document searches and 33 percent are subject searches. Remaining searches are 6 percent author searches (to find out what is on hand from a known author or institution) and 5 percent bibliographic searches (to complete or verify a reference on the basis of catalog-card data). The importance of secondary search clues in achieving retrieval despite incomplete or inaccurate primary clues is discussed. (Author)

Final Report

Grant No. SAR/OEG-1-71071140-4427

USER REQUIREMENTS IN IDENTIFYING DESIRED WORKS IN A LARGE LIBRARY

Ben-Ami Lipetz

Yale University Library
New Haven, Connecticut

June 1970

The research reported herein was performed pursuant to a grant from the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Contents

Acknowledgments iv

Summary
- Objectives and Methods 1
- Results 2

Introduction 9

Methods
- Environment 12
- Traffic Measurements 14
- Interview Schedule 16
- Interview Method 17
- Catalog Card Follow-up 20
- Book "Front Matter" Follow-up 21
- Miscellaneous Measurements 21
- Data Reduction 22

Results
- Publications, Papers 25
- Traffic 28
- Catalog Users 40
- Catalog Searches 43
- Search Success and Potential for Improvement 50
- Search Clues, Catalog Access, Catalog Data 55

Discussion 69

References 72

Appendixes

A. "User clues in initiating searches in a large library catalog," pp. 137-139 in American Society for Information Science, Proceedings, Vol. 5, Annual Meeting, October 10-24, 1968, Columbus, Ohio.

B. "A quantitative study of catalog use," 8 pp. Paper presented at Clinic on Library Applications of Data Processing, University of Illinois Graduate School of Library Science, Urbana, Illinois, April 27-30, 1969.

C. "Performance of Kilgour's truncation algorithm when applied to bibliographic retrieval from a library catalog," pp. 125-7 in American Society for Information Science, Proceedings, Vol. 6, Annual Meeting, October 1-4, 1969, San Francisco.

D. "Performance of Ruecking's word-compression method when applied to machine retrieval from a library catalog," Journal of Library Automation, Vol. 2, No. 4, 266-271 (1969 December.)

E. Forms Used for Interviewing and Follow Up

F. Factors Compared in Covariance Analysis

Abstract

Acknowledgments

Many people contributed in many ways to the work reported in this publication. Particular and emphatic thanks are due to Peter Stangl (who is now Associate Librarian, Yale Medical Library) for his contributions as supervisor of the research team that implemented the research plan through the data collection and data reduction phases. Mr. Stangl was responsible for detailing and testing much of the methodology of this project, and he was responsible also for training and coordinating all of the individuals who conducted interviews and traffic measurements at the card catalog in Sterling Memorial Library, Yale University.

Research personnel who contributed substantially to the collection of data and/or the development of data-collection methods are: Marianne Brindisi, Lynn S. Greenhouse, Barbara Hamilton, Alleavious Hill, Carolyn S. Potts, Katherine M. Skor, and Kathryn F. Taylor. Invaluable assistance in the programming of data reduction was given by John Dow.

Professor Herbert Menzel, New York University, provided useful advice on the development of interviewing technique. The original suggestion of the potential value of undertaking a research project of this type came from Frederick G. Kilgour.

This project was originated and completed under the cognizance of three different project officers at the Office of Education: Eugene P. Kennedy, F. Kurt Cylke, and Lawrence S. Papier. Each of these gentlemen

provided substantial assistance in terms of genuine interest, sympathetic attention to administrative requirements, and constructive criticism.

Assistance in preparation and editing of the manuscript of this report was provided by Roberta Milano and Elizabeth Patterson.

Summary

Objectives and Methods

A study was conducted of the utilization of the card catalog in the main library (Sterling Memorial Library) of Yale University. The study was motivated by interest in both short-term and long-term improvement in catalog performance.

The pattern of traffic flow in the catalog area was determined by means of frequent traffic counts that were continued for more than a year. Observed traffic was used as the basis for designing an interview schedule that would encompass a thoroughly representative sample of catalog users. More than two thousand catalog users were interviewed during a full calendar year. Users were approached and interviewed at the moment of initiating a catalog search. Interview techniques were designed to bring out many details of search objectives and starting clues with a minimum of probing by the interviewer. Information on the user's academic status and experience was collected also. Refusal of interviews was less than 1 percent.

After completion of their catalog searches, users were again approached and queried regarding the results. Call numbers of identified documents were recorded. Catalog cards for these documents and front matter from these documents were copied and compared with the users' starting clues.

A number of miscellaneous studies were made, including an investigation

of the causes of failure in unsuccessful catalog searches.

This study produced very voluminous records. Much of the analysis of results accomplished to date was facilitated by the use of computer programs for covariance analysis. There is ample opportunity for further analysis of available data.

Results

The findings of this study are listed below. Observations regarding gross traffic patterns in the catalog were an essential part of the study and, logically, should precede observations on the content of catalog searches; however, they provide few real surprises and are therefore listed last.

1. At the instant of approach to the catalog, 73 percent of the users are attempting a search for a particular document (known item); 16 percent are attempting a subject search; 6 percent are attempting an author search (to find out what documents are on hand from a known author, publisher's series, or other source); and 5 percent are attempting a bibliographic search (to use the information provided by the catalog card for some document without any intention of locating or borrowing the document).

2. Many users attempt document searches only as a means of locating some subject information probably contained by the documents. In terms of underlying objectives, only 56 percent of the searches are for the document as an end in itself; 33 percent are subject searches (16 percent directly by subject and 17 percent indirectly by a known document).

3. No significant variations in search objectives were detected with respect to time of year or type of catalog user.

4. Objects of attempted document searches were 80 percent monographs and 20 percent periodical articles.

5. Twenty-six percent of the catalog users are already familiar with desired documents. Probability of previous contact with a document tends to increase with years of library use by the individual involved.

6. The success or failure of a catalog search is determined right at the catalog in all bibliographic searches and in 98 percent of the document searches. In 40 percent of the subject searches and 30 percent of the author searches, the catalog user must go elsewhere (e.g., to the stack to look through possibly pertinent books identified at the catalog) in order to determine the success or failure of the catalog search. In general, 91 percent of catalog searches are evaluated at the catalog.

7. Eighty-four percent of document searches succeeded in locating the desired item and its call number. The success rate for author and subject searches appears to be the same or nearly the same.

8. No evidence of frustration or diminishing catalog use was found among catalog users in their first year of experience with the Yale libraries. Success rates appear to be about the same for all types of users.

9. The principal approaches by which users attempt to enter the catalog for document searches are: author, 62 percent; title, 28.5 percent;

subject, 4.5 percent; and editor, 4 percent. Inexperienced users sometimes search by author or title of an analytic (0.5 percent) but soon abandon this unproductive approach.

10. Users in their first and second years of experience at Yale altogether account for 55 percent of catalog use. Each of these two year groups has a different, somewhat atypical, distribution of approaches to the catalog which is not evident from the third year onward.

11. Of the 16 document searches that failed out of every 100 attempted, 10 failed simply because the documents were not in the catalog (one-fifth of these were added to the catalog between the time of the original search and the time of the follow-up study a few months later); 5 were in the catalog and could have been located by the user with his starting clues; 1 could not be traced because of inadequate user clues. Thus, the potential for improvement of catalog service through expansion (especially timely expansion) of the collection and through better orientation of catalog users is far greater than the potential for improvement through expansion of catalog accessibility.

12. There is no general agreement among catalog users regarding a best approach to improvement of the catalog; rather, there is diffuse interest in all possible approaches to improvement.

13. Title information and author information predominate over other types of search clues with respect to both availability and accuracy.

14. The availability and accuracy of search clues tend generally to favor the title approach over the author approach. But the difference in usefulness between the two approaches does not appear to be large.

Circumstances of catalog selectivity under particular entry terms can heavily favor the author approach for some searches and the title approach for others.

15. Among the accurate or possibly accurate clues known by catalog users that are not accessible under present cataloging practices, title-like clues (subtitles, short titles, analytic titles, etc.) are more common than author-like clues (editors, compilers, etc.).

16. Catalog users were usually able to identify desired documents in their searches despite incomplete or misspelled starting clues. Neither of two computer retrieval algorithms tested on data from this study could approximate human performance in overcoming inadequacies of search clues.

17. The use of combinations of computerized approaches shows more promise for effective retrieval than any single approach. However, devices will be required to suppress false drops, and this will probably necessitate retention of a variety of data elements in the computer store, even for simple document searches. Publication date is a prime choice for inclusion in such a store. Subject clues can also be of great value in document searches.

18. Subject searches in a catalog frequently identify far more potentially useful documents than a user can profitably examine. Non-subject data on the catalog cards can help the user to select the most promising documents and to reject the least promising. Virtually every type of data element provided on a card can help to narrow the selection, but relative usefulness of the different elements varies with users and with individual searches. Filing by publication date within a subject

heading would probably be of more general value than the current practice of filing by main entry.

19. Transcription of descriptive data from the title pages and other front matter of documents without creative human input could conceivably build a catalog that is adequate for document searches, but it would lack important conveniences of conventional catalogs.

20. Transcription of subject information from titles, contents, indexes, etc., of documents without human intervention could not possibly lead to a catalog suitable for subject searching unless a device was somehow incorporated for detecting and retrieving synonyms. Files built up by such transcription might tend to be excessively large and to be diluted with trivia not easily distinguished from the major subjects covered by the cataloged documents.

21. As expected, catalog use follows a consistent year-to-year traffic pattern.

22. Intensity of catalog use is twice as heavy during the academic year as during the summer period, but there are no distinct seasonal variations within these two periods. Holidays and recesses during the academic year are characterized by spikes of lower-than-normal catalog use followed by higher-than-normal use.

23. Catalog use and book borrowing parallel each other in an almost constant ratio from week to week. Measurement of catalog use can serve as a predictor of borrowing, or vice versa.

24. Patterns of variation of catalog use by day of the week, hour of the day, and fraction of the hour have been detected and described.

Knowledge of these patterns can be applied in planning physical facilities and in scheduling reference assistance at the catalog.

25. There is a strong tendency for catalog use to occur immediately after a user's entrance to the library, as one might expect.

26. Users of the catalog appear to be divided into two different but roughly equal populations on the basis of their length of stay at the catalog. The distribution patterns of the two groups exhibit modes of about 2 minutes and about 6 or 7 minutes, respectively.

27. Catalog users respond readily to interviews regarding their intended catalog searches. Refusal rates in this type of study are almost negligible.

28. As a group, graduate students are the heaviest users of the catalog, followed closely by undergraduates. Total faculty use is light by comparison. One fifth of the total catalog use comes from persons not directly associated with the university.

29. Per capita use of the catalog is somewhat higher for upperclassmen than it is for graduate students. Per capita use by freshmen and by faculty are equal and are at a level about half that of upperclassmen. (This does not take into account faculty use of departmental libraries.)

30. Catalog use throughout the year is relatively constant for faculty. Use by graduate students declines in the summer to half the academic year level; use by undergraduates drops to about one sixth during the summer. Use by visitors doubles during the summer.

The interpretation of these results can vary greatly, depending on whether a librarian is more interested in expanding service or in conserving

money and labor. Some of the apparent implications for expansion and for retrenchment are discussed. Improved user orientation is a particularly attractive approach to improved catalog service.

Introduction

The work reported here is a study of the utilization of the card catalog of a very large library, specifically the principal catalog of the library system of Yale University.

The study was motivated by two basic concerns, one of them of a long-term, or exploratory, nature, the other of a short-term, or operationally supportive, nature. The long-term concern is the question of how to design a computerized catalog for a very large library that can be expected to give the best possible performance. The short-term concern is the question of whether, and, if so, how, existing card catalogs in very large libraries may be made more responsive to user requirements. It was recognized that a carefully designed study of actual utilization of a catalog of a large library could shed useful light in both areas of concern.

The connection between the research design and these basic concerns is very straightforward. One cannot create an ideal tool of any sort on a rational basis (whether that tool be a conventional catalog, a computerized catalog, or any other device for any other application) without knowing a good deal about the purpose or purposes for which the tool is to be used, and about the manner in which the users interact with the tool. In the literature on libraries, there is a dearth of reliable information on the utilization of catalogs. This study was undertaken in an attempt to fill the void, at least to a degree.

In the preceding paragraph, there was no intention of implying that there have been no published studies of catalog utilization. In fact there have been many. The point is that their results have not been reliable. Most past studies have been unreliable because they were much too small, involving very small samples of actual catalog use. Usually in these reports the method of selecting a sample of catalog users is either unstated or clearly of such a nature as to invite slanting of results. And, in almost all instances, the method of data collection is suspect, making use of interviews or questionnaires administered after (sometimes long after) the instances of catalog use under investigation, and therefore inviting gross errors due to faulty human memory.

The intent of the study reported here was to circumvent the several shortcomings which made earlier studies unreliable. The study was designed to sample a significant fraction of actual catalog use (approximately one percent) over a significant period of time (a full year). The selection of the sample of catalog users to be studied was made as representative as possible by basing it solely on observed volume of traffic in the catalog during different times of the day, days of the week, and seasons of the year. Information on needs and approaches was gathered from catalog users immediately preceding a catalog search, rather than only after the search. The clues with which users began their searches were later compared with the search results, with the approaches afforded by the catalog, and with alternative search approaches that could conceivably be provided by making specific changes in cataloging rules that would take advantage of automatic indexing capabilities of computers.

The research proposal to the Office of Education was approved in July

1967. Work began in late September 1967. Systematic measurement of traffic through the catalog was begun in late November 1967 and was continued, with one brief interruption, through early February 1969. The gathering of usable data from interviews began in March 1968 and concluded in April 1969.

Methods

Environment

The catalog whose utilization was investigated in this study was the public catalog in the Sterling Memorial Library of Yale University. The total library collection of Yale University includes more than 5 million volumes and is divided into some sixty or so units housed in various buildings throughout the campus. Sterling Memorial Library functions as the main library for Yale University, and it houses the largest portion of the total Yale collection--approximately 3.5 million volumes. The largest of the other library units at Yale (none of them housing as much as a half-million volumes) are: divinity, law, medicine, science, rare books, art and architecture, and music. The public catalog at Sterling Memorial Library provides access to all of the holdings in the total Yale collection. It contains full sets of catalog cards for books shelved at Sterling Memorial Library; and it contains only main-entry cards for books shelved elsewhere at Yale.

Physically, the public catalog of Sterling Memorial Library is located near the front entrance of the building, in a rectangular area approximately 60 feet by 40 feet, immediately to the side of the principal thoroughfare leading to the circulation desk, stack, main reading room, reference area, periodical reading room, special collections, offices, etc. Four aisles of the catalog open from the long side of the catalog area onto this thoroughfare. A fifth aisle opens from the short side onto a spur corridor which leads from this thoroughfare only to the main reading room.

The only significant portions of Sterling Memorial Library which may be reached from the main entrance without passing by the public catalog are the undergraduate reserve-book room, a browsing room, and some lavatory facilities. However, there is a second entrance to the building which is normally open on weekdays until 6 P.M. and which is used a good deal; from this alternate entrance, it is possible to reach all portions of the building except those just mentioned without passing by the catalog. Thus, it was not safe to assume that a visitor to the building who has a problem that might warrant a catalog search will actually use the catalog, as a matter of convenience, before going elsewhere in the building. Neither was it safe to assume that the catalog users who enter the catalog through a particular aisle are representative of the users who enter through the other four aisles that are available, since users coming from the front entrance might tend to favor one aisle, users coming from the main reading room might tend to favor a different aisle, etc., and each group could conceivably have significant differences in their requirements.

The catalog is housed in cabinets that are 14 drawers high. Subject entries and name and title entries are all interfiled in a single alphabetic sequence. Subject headings are based on the Library of Congress arrangement, differing in only minor respects. Contents of the cards are very similar, in data elements and arrangement, to contents of Library of Congress catalog cards; indeed, a large fraction of the cards in the catalog are modified prints of Library of Congress cards. The catalog is estimated to contain approximately 8 million cards. These were housed in about 6,000 card drawers at the beginning of the study; about half-way through the data-gathering period, the catalog was expanded into 7,000 drawers by adding a bank of cabinets in the center of the rectangular catalog area.

It should be noted that Sterling Memorial Library is an open-stack library. Yale students, faculty, staff, and many outside users holding authorization cards, all have the privilege of entering the stack to browse and to remove books for borrowing. A page service is available to all users during normal working hours. It requires the prior filling out of a loan form, including the call number of the desired item. This service is heavily patronized by many users who do not wish to enter the stack; it is the only option available to the non-Yale visitor who lacks stack privileges and wishes to consult a book housed in the stack.

Traffic Measurements

The determination of the pattern of people entering the catalog area was a key factor in the later design of an interviewing schedule which would yield a clearly representative sample of catalog users. The pattern of entry to the catalog was determined by having observers assigned to count the number of people entering the catalog area through different entryways during different times of day and days of the week. Observers were stationed where they could observe simultaneously either the front three aisles into the catalog or the rear two aisles into the catalog. For a period of five minutes duration, they would count the number of persons entering each of the aisles being observed. Timing periods were rigidly predetermined to cover different hours of the day, different days of the week, and even different tenths of each hour. Observation assignments were rigidly scheduled; the schedule repeated every seven weeks. Observations were continued over a total of 62 weeks so as to provide a 10 week overlap period for determination of any annual variation in traffic which might occur. (During this 62 week period, there was a 5-week interruption in observations, during the late summer, while shifting

of catalog drawers was going on; the abnormal shifting activity tended to interfere with traffic flow.)

The total amount of time during which traffic was counted was somewhat over 4 percent of the time that the library was open during the total time span involved. For practical reasons, the coverage was more intense during weekday working hours (6 percent) and lower during evening hours and weekends (about 2.5 percent). However, observed traffic was also lower (by about one-fourth) during evening hours and weekends. Tallies of traffic counts by hour, day, and entryway for the first 10 weeks of observation were used as the basis for designing the interviewing schedule. Traffic counts were continued during the interviewing period to check on the continuing validity of the pattern observed during those first 10 weeks and to provide a rational basis for weighting of interview results if the interview schedule should prove to be biased with respect to observed traffic.

Several other traffic measurements were made in addition to the counts of persons entering the catalog area: At precise preassigned times, observers would follow anyone entering the catalog to observe where he went (which catalog drawer), how long he stayed at the catalog, and how many call-number notations he wrote down. Intervals for conducting these observations were scheduled in exactly the same pattern as intervals for gross traffic counts, so as to cover all times of catalog availability.

Observers of catalog traffic were instructed to avoid counting those library staff members who regularly work in the catalog area (filers, verifiers, reference librarians). The intent of the measurements was to count, as far as possible, only the "consumers" of the catalog service,

rather than the suppliers and interpreters.

Interview Schedule

The schedule for conducting interviews with catalog users was based on observed traffic into the catalog during an initial 10-week observation period. Projection of observed traffic for this period suggested that annual traffic into the catalog would be of the order of 300,000. (Full-year traffic observations later showed this estimate to be low.) When one adjusted this count to omit individuals who were found to be entering the catalog merely to use it as a shortcut between the front entrance and the main reading room, the indicated annual total of real catalog users was closer to 250,000. It was decided that about 2,500 interviews or more (i.e., at least something approaching one percent) should be attempted.

The interview schedule that was adopted called for an interviewer to be at a particular entry to the catalog area at a specified time on a specified day of the week. The first individual other than library staff to enter the catalog through that entryway during the next six minutes and to begin to use the catalog would be the person to be interviewed. (If no one entered during that interval, no one was interviewed until the next assigned time and place.) Interview assignments were set up on a revolving schedule very much like the schedule described above for traffic measurements. However, adjustments were made to reflect the observed relative traffic volume through each of the five entryways to the catalog, and to reflect the observed relative traffic during different hours of the day. Adjustments for minutes of the hour and for day of the (regular) week were not judged to be necessary. As with traffic measurements, the schedule for interviewing during the evening and weekend periods was made

lighter than during the regular weekday periods; this was done with the knowledge that compensations could be made later by weighting the results of actual evening and weekend interviews somewhat more heavily than the results of weekday interviews in compiling final statistics.

Interview Method

Interview content and technique were designed to elicit quite specific information from catalog users, with a minimum amount of bias due to prompting or leading by the interviewer. The method adopted made use of an interview guide in the form of a multiple-part questionnaire (Appendix E) which interviewers were required to follow uniformly. Interviews would begin with very vague, nondirective questions ("Please tell me precisely what you were about to do at the catalog the moment I interrupted you."), in order to give the user full opportunity to state whatever he happened to regard as important or significant. Only as the interview progressed would the questions become more specific, so as to fill in details which the user had not already supplied but which were regarded *a priori* as important to the study.

The underlying pattern of the interview involved identifying rather quickly the basic type of search which the user was about to make in the catalog (e.g., a search for the purpose of borrowing a specific known document; a search for the identity of documents on a specific subject; a search for the identity of documents from a specific source, as a particular author or a particular organization; a search for descriptive bibliographic information regarding a known document without any intent of borrowing the actual document). Identification of this basic type of search would then determine which of several possible lines of questioning

to follow in the remainder of the interview.

When it appeared that no more useful information could be gathered regarding the immediate search being conducted by the user, the interview would be terminated with a series of questions on the user and his personal background (but not his name). Background questions related to the user's status at Yale, his field of specialization, the length of his residence in the Yale community, and the general level of his use of Sterling Memorial Library and other libraries at Yale.

Questions asked during the main portion of the interview were intended to bring out everything of possible retrieval value that the user knew about the material he desired at the time of starting his search. This would include, as appropriate, the type of document (whether an ordinary book, or a series, periodical, report, etc.), descriptive data (author, title, date, publisher, etc.), physical characteristics of a document (size, color), contents (index, illustrations, bibliography), subject terms, translation specification, edition specification, and so forth. The questions also established whether or not the user was already familiar with the material he wanted, how he had first learned of the existence of the material, the connection in which he wanted to make use of the material, and the particular clue which he intended to use to begin his search of the catalog. Particular pains were taken to record descriptive data elements exactly as they were known to the user, taking nothing for granted: If the data came only from his memory, he was asked to spell out the authors and the longer title words; if the data came from class notes or duplicated lists which he had brought to the library, these were photocopied by the interviewer.

At the conclusion of the interview, the user was left alone to complete his catalog search. However, he was observed discreetly from a distance. The amount of time spent at the catalog and the number of catalog drawers searched were noted on the interview record. As the user was leaving the catalog area, he was stopped again and asked whether his search had been successful. If the answer was affirmative, he was asked to let the interviewer copy any call numbers that he had found in the catalog that satisfied his search needs. Users who were not certain whether their searches had been successful but who were going elsewhere in the library to find out (usually these were people who had identified a potentially useful stack area by finding some representative class numbers in the catalog and who intended to browse the stack for known and/or unknown documents) were given a self-mailing follow-up form on which they could conveniently note any call numbers that were subsequently found to satisfy their needs.

Several months were spent in developing and testing the interview outline and technique before starting the full year's run of data collection for the project. Only very minor changes were made as the year progressed. Five individuals performed practically all of the interviews. A comparison of the results of interviews conducted by different interviewers was made about four months after the start of the interviewing year; no serious biasing of results could be associated with the interviewers compared. Therefore, the interviewing technique was judged to be quite objective, as had been hoped.

The interviewing schedule that was adopted provided for a maximum of some 2700 interviews during the full year studied. Because of

various random factors (e.g., no user at the catalog at the scheduled time and place, unexpected library closings, or illness of the interviewer), the number of interviews actually completed in the year was 2,134. It is interesting to note that fewer than 1 percent of the catalog users who were approached refused to grant an interview (and usually only because of hurry to get to a class); most of the users interviewed were extremely pleased to learn that some people were really interested in their library needs, and they answered questions without reluctance. During the interview year it was inevitable that some individuals would be interviewed more than once, purely by chance; such instances were identified by one of the routine interview questions and noted in the interview records.

A large multi-library survey published in 1958 (1) used the technique of accompanying catalog users through their searches. The study reported here uses the technique of interviewing catalog users before the start of a search and later ascertaining the results of the search. An earlier independent application of this method occurred in an unpublished thesis project at the University of Chicago Harper Library (2); in that project, 100 searches for particular documents ("known item" searches) were studied. More recently, a similar but much larger study of "known item" searches at both public and university libraries in Ann Arbor, Michigan, was reported (3).

Catalog Card Follow-Up

Almost all of the catalog searches which users regarded as successful were searches which resulted in the identification of documents or catalog cards bearing sepcific call numbers. By looking up each call number in the shelf-list card file, it was comparatively simple for the project staff to obtain a photocopy of the basic catalog card for each item associated with a successful

catalog search. These photocopies were attached to the respective interview records for later use in comparing search clues brought to the catalog by the user with data and access points available in the catalog.

Book "Front Matter" Follow-Up

Follow-up on call numbers identified in successful searches extended not merely to catalog cards but also to the actual books which the call numbers represented. After allowing a period of several weeks, at least, for the user to finish with the items he identified, these books would be borrowed from the library shelves (or recalled) and examined by the project staff. Certain non-central portions of these books were photocopied when present and not redundant, including (but not necessarily in all cases) cover, title page, verso of title page, table of contents, preface, brief introduction, and index. This photocopied material was also attached to the respective interview records, to be used later for comparing search clues brought to the catalog by the user with potentially matching data elements that are conveniently available to a cataloger (or to a hypothetical optical-scanning device that could conceivably be substituted for a human cataloger in the fanciful future).

Miscellaneous Measurements

A number of miscellaneous measurements were made which are related to the understanding of the needs of catalog users.

In the case of known-document searches which failed, a second catalog search attempt was made by the project staff, using the clues supplied by the user before his unsuccessful attempt. In a number of

cases, the item wanted was clearly identified in this second search, indicating either a user error or inadequate familiarity with the catalog arrangement.

Statistics were compiled on the distribution of catalog entries in the Sterling Memorial Library public catalog in terms of the first two letters of the file term. Similar statistics were compiled for the first two letters of call numbers of items identified in successful catalog searches. Another compilation was made for the first three letters of catalog headings under which items were located. Statistics were compiled on observed use of individual catalog drawers.

Two different published formulas for achieving retrieval from computerized files despite spelling errors in the data to be searched were tested for effectiveness in catalog searching by applying the formulas to user clues and catalog-card data from successful document searches studied during the interviewing phase. One of the formulas tested attempted to negate misspellings by truncating the words to be matched; the other formula attempted to achieve the same result by applying specific rules for condensing the words to be matched.

Data Reduction

The records obtained from the thousands of interviews and the thousands of traffic measurements described above were extremely voluminous. Computer methods were invoked to assist in their analysis. Codes and formats were developed for keypunching much of the collected data on IBM cards. Virtually all of the data obtained from traffic measurement were susceptible to this treatment; data from a given

measurement instance could be recorded on a single punched card. Interview data, however, were much less tractable. For example, it was not generally feasible to keypunch every detail of information about a desired book supplied by the user, or available from the catalog card. But it was possible to keypunch indications as to whether or not certain types and ranges of information were available from the user or the catalog card. Some aspects of accuracy and correspondence could also be indicated. The selection of information characteristics to recognize in the punched card format is largely a reflection only of the judgment and intuition of the research group. Many descriptive characteristics that might have been reduced to punched card input were not. However, basic records were preserved to permit further analysis of data by manual methods or further reduction of data to machine-usable form.

Data on punched cards were analyzed mainly by the use of table, or matrix, programs provided by the Yale Computer Center. With these programs, the computer will take data concerning any two specific variables in the body of data supplied to it and will print out a table, or matrix, showing the co-occurence of these variables in terms of any individual value or any specific range of values for each of the two variables. Totals are provided for each row of figures and for each column of figures. The tables or matrixes can be made to show either raw data for each row-column position, or percentages of total populations for each row or column involved. One can specify that one table of each kind be printed. Furthermore, a third variable can be brought into the program and the computer can be required to print out a series of subsidiary tables or matrixes,

each containing just that portion of the data regarding the first two variables that applies to a particular value, or range of values, of the third variable. It takes very little arithmetic to determine that the number of tables which could conceivably be produced from a body of data involving scores of variables, as in this study, is quite astronomical and quite beyond the budget of most research projects. If all possible three-variable correlations were actually printed out, it is doubtful that anyone would have the time and energy to study them. Therefore, only a limited number of the more promising tables was printed, mostly of the two-variable type.

The selection of combinations of variables to be represented in tables or matrixes was based upon the results of still another computer manipulation. The randomness or nonrandomness of correlation (covariance) of each pair of variables in the data base was determined by a statistical program which printed out a short table showing the measure of degree of correlation for each pair. Pairs of variables showing relatively strong degrees of correlation were easily identified from this table; these were generally the variables selected for detailed elucidation by means of the table program described above. More than 40 variables (Appendix F) were compared for covariance. Over a hundred potentially interesting cases of covariance were identified; as yet, not all of these have been examined in detail.

Results

Publications, Papers

This report is final in the sense that it terminates the support grant under which data were collected and processed. It does not nearly exhaust the possibilities for useful analysis of the very large and multifaceted data collection that has been assembled. Results presented below represent only a first skimming of the data--an attempt to derive quick answers, where possible, to some of the more obvious and important questions about catalog use and catalog needs. It is hoped that the opportunity and means will be found in the future to make further use of the excellent data collection in order to gain better quantitative understanding of the questions considered in this report, and in order to answer many additional questions that have yet to be considered.

Preceding this report, there were four publications (4-7) resulting from this study; they are included here as Appendixes A, B, C, and D. Two of these publications (4, 5) are primarily descriptions of the design of the study, presenting very few and very preliminary findings. The other two publications (6, 7) present samplings of data on the number and types of errors found in the clues with which catalog users begin their searches; these data are used to assess the values of two different automatic searching algorithms which have been proposed by other authors for achieving retrieval from computerized bibliographic files despite inaccuracies in the data to

Table 1

Traffic in Catalog, Weekly

Week	First Day	Library Hours Day	Library Hours Eve	Open Hours Wkend	Open Hours Total	Cat. Users (by Extrap.)	Users Per Hour
1	671127	42.5	33.75	18.0	94.25	9,867	104.59
2	671204	42.5	33.75	18.0	94.25	10,028	106.10
3	671211	42.5	33.75	3.25	79.5	6,923	87.08
4	671218	41.25			41.25	3,028	73.41
5	671225	33.0			33.0	3,491	105.79
6	680101	33.75	20.25	18.0	72.0	7,719	107.21
7	0108	42.5	33.75	18.0	94.25	8,722	92.45
8	0115	42.5	33.75	18.0	94.25	7,979	84.58
9	0122	42.5	33.75	18.0	94.25	8,112	85.99
10	0129	42.5	33.75	18.0	94.25	9,410	99.75
11	0205	42.5	33.75	18.0	94.25	9,908	105.02
12	0212	42.5	33.75	18.0	94.25	9,996	105.96
13	0219	42.5	33.75	18.0	94.25	9,312	98.71
14	0226	42.5	33.75	18.0	94.25	8,461	89.69
15	0304	42.5	33.75	18.0	94.25	9,636	102.14
16	0311	42.5	33.75	3.25	79.5	7,618	95.82
17	0318	41.25			41.25	4,563	110.61
18	0325	41.25			41.25	4,705	114.06
19	0401	42.5	33.75	18.0	94.25	9,249	98.04
20	0408	34.0	27.0	18.0	79.0	6,352	80.41
21	0415	42.5	33.75	18.0	94.25	10,049	106.52
22	0422	42.5	33.75	18.0	94.25	9,637	102.15
23	0429	42.5	33.75	18.0	94.25	9,056	95.99
24	0506	42.5	33.75	18.0	94.25	9,500	100.70
25	0513	42.5	33.75	18.0	94.25	8,159	86.49
26	0520	42.5	33.75	18.0	94.25	7,288	77.25
27	0527	42.5	33.75	18.0	94.25	5,728	60.72
28	0603	42.5	33.75	18.0	94.25	4,660	49.40
29	0601	42.5	33.75	18.0	94.25	4,051	42.94
30	0617	42.5	33.75	18.0	94.25	2,408	25.52
31	0624	42.5	33.75	18.0	94.25	4,660	49.40
32	0701	34.0	19.0	13.0	69.0	2,772	40.17
33	0708	42.5	33.75	18.0	94.25	4,375	46.38
34	0715	42.5	33.75	18.0	94.25	2,841	30.11
35	0722	42.5	33.75	18.0	94.25	4,184	44.35
36	680729	42.5	33.75	18.0	94.25	3,302	35.00
37	0805	42.5	33.75	18.0	94.25	3,600	38.16
38	0812	42.5	33.75	18.0	94.25	2,660	28.20
39	0819	42.5	33.75	18.0	94.25	2,853	30.24
40	0826	42.5	33.75	18.0	94.25	(3,599)	38.15

Week	First Day	Library Open Hours Day	Eve	Wkend	Total	Cat. Users (by Extrap.)	Users Per Hour
41	0902	34.	19.	13.0	66.0	(2,998)	42.39
42	0909	42.5	33.75	18.0	94.25	(3,599)	38.15
43	0916	42.5	27.75	18.0	88.25	(8,496)	95.14
44	0923	42.5	33.75	18.0	94.25	(9,074)	96.18
45	0930	42.5	33.75	18.0	94.25	9,260	98.16
46	1007	42.5	33.75	18.0	94.25	9,948	105.49
47	1014	42.5	33.75	18.0	94.25	8,336	88.36
48	1021	42.5	33.75	18.0	94.25	9,772	103.58
49	1028	42.5	33.75	18.0	94.25	9,678	102.59
50	1104	42.5	33.75	18.0	94.25	9,961	105.59
51	1111	42.5	33.75	18.0	94.25	10,179	107.90
52	1118	42.5	33.75	18.0	94.25	9,586	101.61
53	1125	33.5	13.5	9.75	56.75	5,916	104.25
54	1202	42.5	33.75	18.0	94.25	12,627	133.85
55	1209	42.5	33.75	18.0	94.25	11,012	116.63
56	1216	42.5	33.75	3.25	79.5	9,091	114.34
57	1223	33.0			33.0	2,772	132.00
58	1230	33.0			33.0	3,218	97.52
59	690106	42.5	33.75	18.0	94.25	12,215	129.38
60	0113	42.5	33.75	18.0	94.25	9,167	97.17
61	0120	42.5	33.75	18.0	94.25	8,450	89.57
62	0127	42.5	33.75	18.0	94.25	8,219	87.12
Totals		2561.5	1814.0	945.5	5321.0	444,035	
1-53	total	2198.0	1577.75	834.25	4610.0	367,264	79.67
	average				86.98	6,930	
10-62	total	2198.5	1591.25	834.25	4624.0	378,166	81.78
	average				87.25	7,135	
1-9	total	363.0	222.75	111.25	677.0	65,869	97.30
	average				75.22	7,319	
54-62	total	363.5	236.25	111.25	711.0	76,771	107.89
	average				79.0	8,530	

be matched by the computer. An informal presentation on methods and results of this study was scheduled to be presented at the Gordon Research Conference on Problems in Scientific and Technical Information, Colby Junior College, New London, New Hampshire, July 12-17, 1970.

Traffic

Seasonal variation of traffic in the catalog was studied by determining weekly figures for traffic into the catalog area. This was done by taking the average number of users (catalog entrants) per observation period actually counted during that week and then multiplying by the number of equivalent periods during which the library was open during that week. In making this calculation, observations for weekday business hours (to 5 P.M.) were considered separately from observations for weekday evenings (to 11:45 P.M.) and observations for weekends (Saturdays and Sundays). This was necessary because these three time periods were sampled in different proportions. The indicated total traffic into the catalog area in a week was the total of separate calculations for the three time periods.

Weekly library hours and calculated traffic are shown in Table 1. Calculated weekly traffic is plotted in Figure 1. Also plotted in Figure 1 is the weekly average of catalog traffic per hour that the library was actually open during that week. Weeks during which hours were restricted or during which the library was closed for a holiday are indicated. Since no traffic counts were made or extrapolated for the very-low-traffic periods before 9 A.M. and after 10 P.M. the figures plotted in Figure 1 are systematically distorted to slightly less than their real values. The five-week late-summer gap during which

Figure 1. Weekly Catalog Traffic: Use Rate and Users

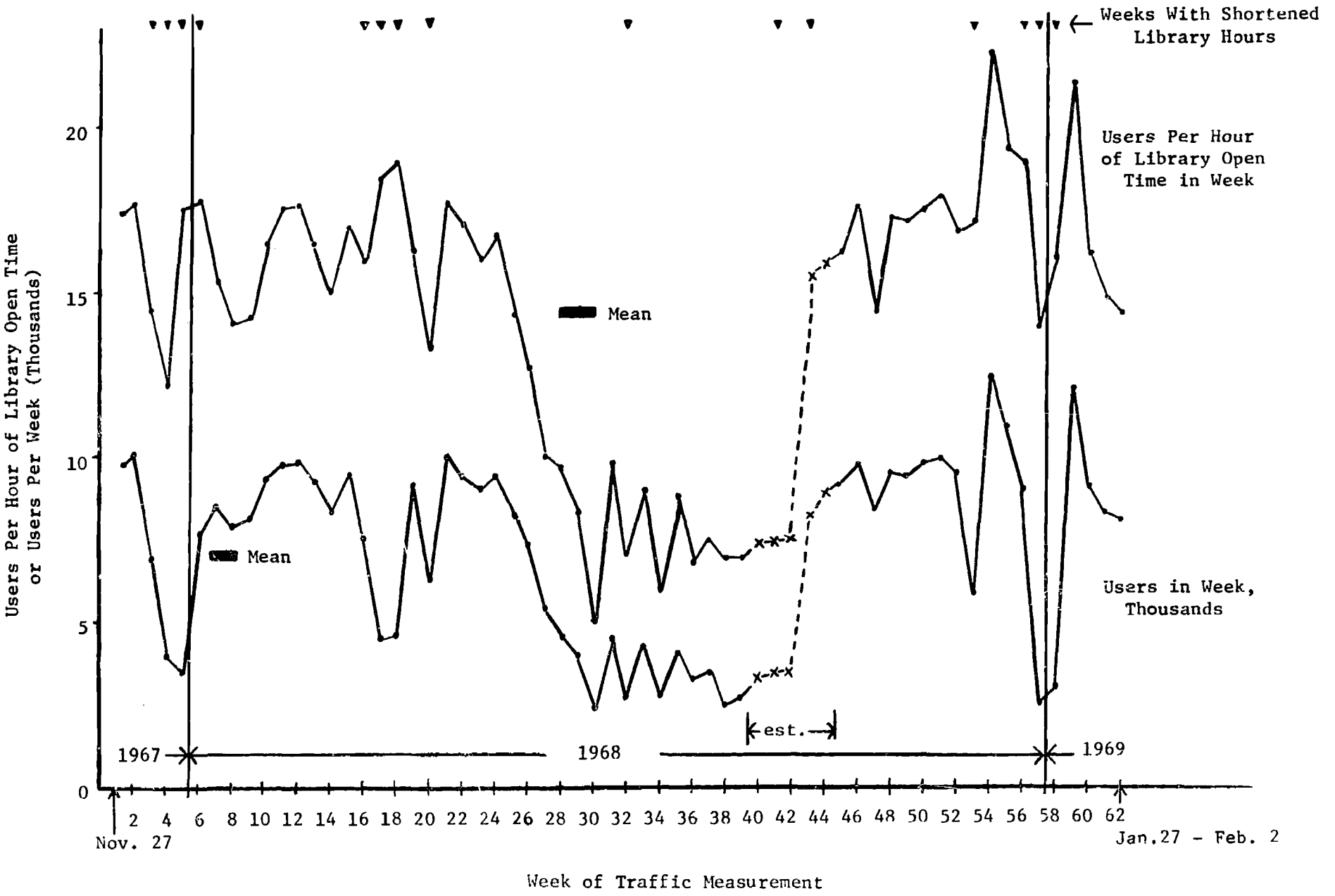
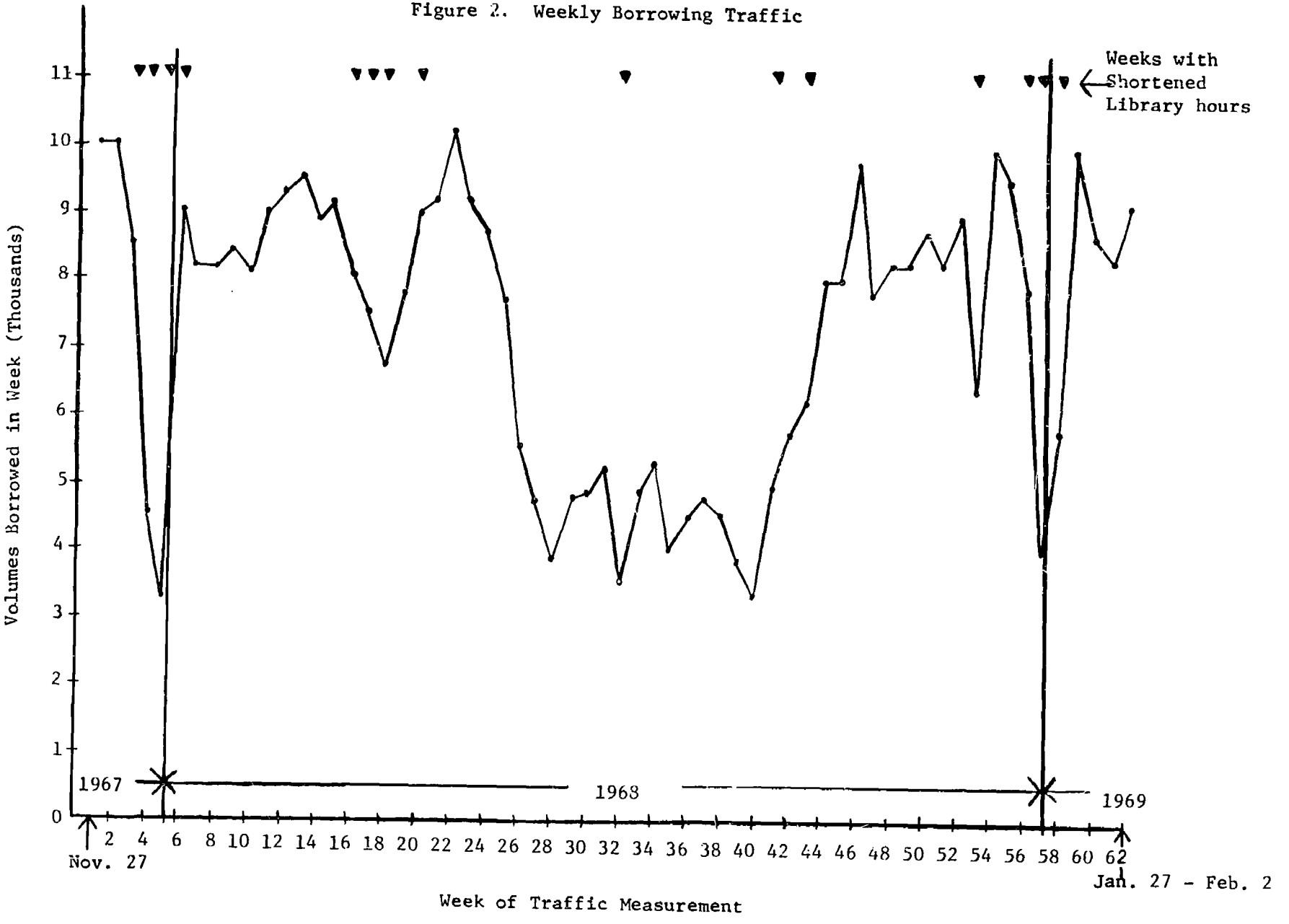


Figure 2. Weekly Borrowing Traffic



no traffic measurements were made has been filled in by extrapolating the curves from either side of the gap; the curve shape thus supplied to this gap is completely consistent with casual observations made by the research team during this period and with the past experience of Yale librarians who have witnessed the rapid back-to-school build-up of library use at the start of a new school year.

Figure 1 indicates that there are only two significant seasons for catalog use: the regular academic year, and the summer vacation period. During the summer period, activity is reduced to half the level for the regular academic year. (At Yale, various short summer courses are offered, but primarily for persons other than full-time Yale students. Virtually all undergraduates and many graduate students and faculty members are absent during the summer.) The activity pattern of the regular academic year is punctuated by irregular declines that are associated with holidays and recesses (as indicated by reduced library hours). This shows up clearly in the curve for total users per week. Because of shortened library hours, the curve for users per hour does not necessarily drop when the total user curve drops (see, e.g., the Easter recess period--weeks 17 and 18). The academic year pattern takes a full month to drop off into the summer pattern and somewhat less than a month to be resumed.

It is interesting to compare catalog traffic throughout the year with statistics for the borrowing of books at Sterling Memorial Library during the same time span. Circulation statistics supplied by the Circulation Department are plotted in Figure 2. These show all recorded book loans (whether for outside or in-building use; and whether borrowed

by students, faculty, library staff, or others). It can be seen that this curve follows the curve for catalog users in Figure 1 in most details. In fact, it appears that one can predict circulation rather accurately from a knowledge of catalog traffic and vice versa. This could be of some value to reference librarians and circulation librarians in scheduling their staffs. A discrepancy occurs at New Year, when book borrowing is somewhat heavier than catalog traffic would suggest; a similar discrepancy occurred in week 20 (Martin Luther King's funeral) when borrowing remained unchanged although catalog traffic declined sharply. At the beginning of the Fall semester, week 46, borrowing rises more sharply than the number of catalog users. There are only minor disparities during the summer period.

In general, the match of these curves is quite remarkable. However, there is no intention to imply here that book borrowing results solely and immediately from catalog use. Browsing is known to occur, and can result in formal borrowing. Catalog users can identify books and not borrow them until days or weeks later. Catalog users can use the catalog for purposes other than obtaining books. Yet, the interesting point is that all of these phenomena, if important, tend to even themselves out, leaving book borrowing and catalog traffic directly proportional to each other throughout virtually the entire year.

Both Figures 1 and 2 include annual overlap periods of about 10 weeks. In both cases, there is very great similarity in the pattern from one year to the next during this overlap period. Some variation in pattern might be attributable to the holidays of Christmas Day and New Years Day, which fell on Monday in 1967/8 and on Wednesday in 1968/9, possibly

causing different amounts of travel by potential library users in these respective years. Comparison of users in this overlap period suggests an annual growth rate of 10 percent in library use, but this particular time of year seems unreliable for predictions because of the differences in the holidays. Comparison of book borrowing figures in the overlap period also suggests an annual growth rate, but a smaller one of about 5 percent. It is interesting, however, that circulation figures for the successive July-to-June fiscal years 1967-1968 and 1968-1969 indicate no change in annual circulation; the total was 370,000 volumes borrowed in each year.

The total number of catalog users during a year, as indicated by traffic measurements (Table 1) appears to be 370,000 or 380,000, depending upon which overlap period one chooses in defining a year. The fact that this number is the same as the total of books borrowed is probably a coincidence; there is no reason to infer that each person entering the catalog seeks and borrows a single volume. It is well known that that is not the case. It should be noted that the measurements consistently ignored traffic during periods just after library opening in the morning and before closing late at night when traffic is very light. If these periods were included, it is estimated that the total number of persons entering the catalog area in a year would be about 400,000 or slightly less. Furthermore, it should be remembered that a substantial number of people who enter the catalog area merely walk through it as a short-cut between the front entrance and the main reading room. Observations indicated this to be about 80,000 "walkthroughs" in a year. Thus, the number of persons actually consulting the catalog in a year was of the order of 320,000.

The geometry of the catalog area was discussed above, and it was mentioned that there are five entryways into the catalog area. On the basis of gross traffic measurements, it was ascertained that the two entryways farthest apart (the one nearest the front entrance and the one nearest the main reading room) receive the most entrants. Traffic into these entryways is about equal. The three intermediate entryways have gross entry traffic ranging from one-half to two-thirds this level. However, the pattern changes if one ignores the people who walk through the catalog area without using the catalog. The "walkthroughs" enter mainly through the two portals nearest the main reading room. The rate of entry of actual catalog users is clearly highest through the portal nearest the front door. Entry of actual users through all other portals is roughly the same, and is about half the rate for the portal nearest the front door. This finding suggests that there is a strong tendency for catalog use to be undertaken immediately upon entering the library by those who use the catalog at all.

A clear pattern of catalog traffic variation with day of the week was observed, and is shown in Table 2. Figures are derived from measurements made between the hours of 9 A.M. and 10 P.M. It can be seen that the catalog use rate is heaviest during the early part of the week, especially on Tuesdays, and that it is lowest on Saturdays and Sundays, as one would expect. The grand mean use rate for hours in this range that the library was open is 95.9 entrants per hour. The span of deviations from this mean from the busiest day (Tuesday) to the slowest day (Saturday) is 26 percent of the mean. The percentages shown in the table refer to entrants per hour and not to total users in a day; the library was open only half as many hours on the average Saturday or

Table 2

Variation in Catalog Attendance by Day of Week

	Entrants per hour	Percent of Yearly Average
Monday	100.4	104.7
Tuesday	106.5	111.1
Wednesday	102.0	106.4
Thursday	95.0	99.1
Friday	92.6	96.6
Saturday	81.7	85.2
Sunday	82.5	86.0
Yearly Average	95.9	

Sunday as on the average weekday.

Catalog traffic varies greatly from hour to hour during the day. Average hourly traffic for an entire year is plotted in Figure 3. Variations show up more sharply if one plots only the average hourly traffic for weekdays (Figure 4). It can be seen that the rate of influx of catalog users builds to a morning peak at about 11:15, drops off through the early lunch period, then builds rapidly to a maximum just before 2 P.M., remains high until 4 P.M., drops rapidly to an extremly low level just after 6 P.M., rises again after 6:30, but not very much, and finally drops off again from 8 P.M. until closing time.

Data were collected which permitted articulation of use rate variations at 6-minute intervals. Such articulation produces few surprises. It shows, during weekdays, a clear build up of entry rate in the 12 minutes immediately after the hour in the morning and mid-afternoon (as users arrive from classes). From 10 A.M. to 5 P.M. there is a tendency for the entry rate to drop off shortly before the hour (except at 1 P.M.). At 5 P.M. sharp and 6 P.M. sharp, there are rapid declines in entry rate which are shown clearly by the data. At 7 P.M. sharp and 8 P.M. sharp there are rapid rises in entry rate. The range from highest to lowest entry rate during the five 6-minute intervals in a half-hour period is generally about 25 percent to 35 percent of the average entry rate for that period.

All of the foregoing discussion of catalog traffic referred to the rate at which people entered the catalog area. How long people tend to stay in the catalog area is also a valid question. Observation of

Figure 3. Yearly Total Catalog Traffic by Half Hour of the Day

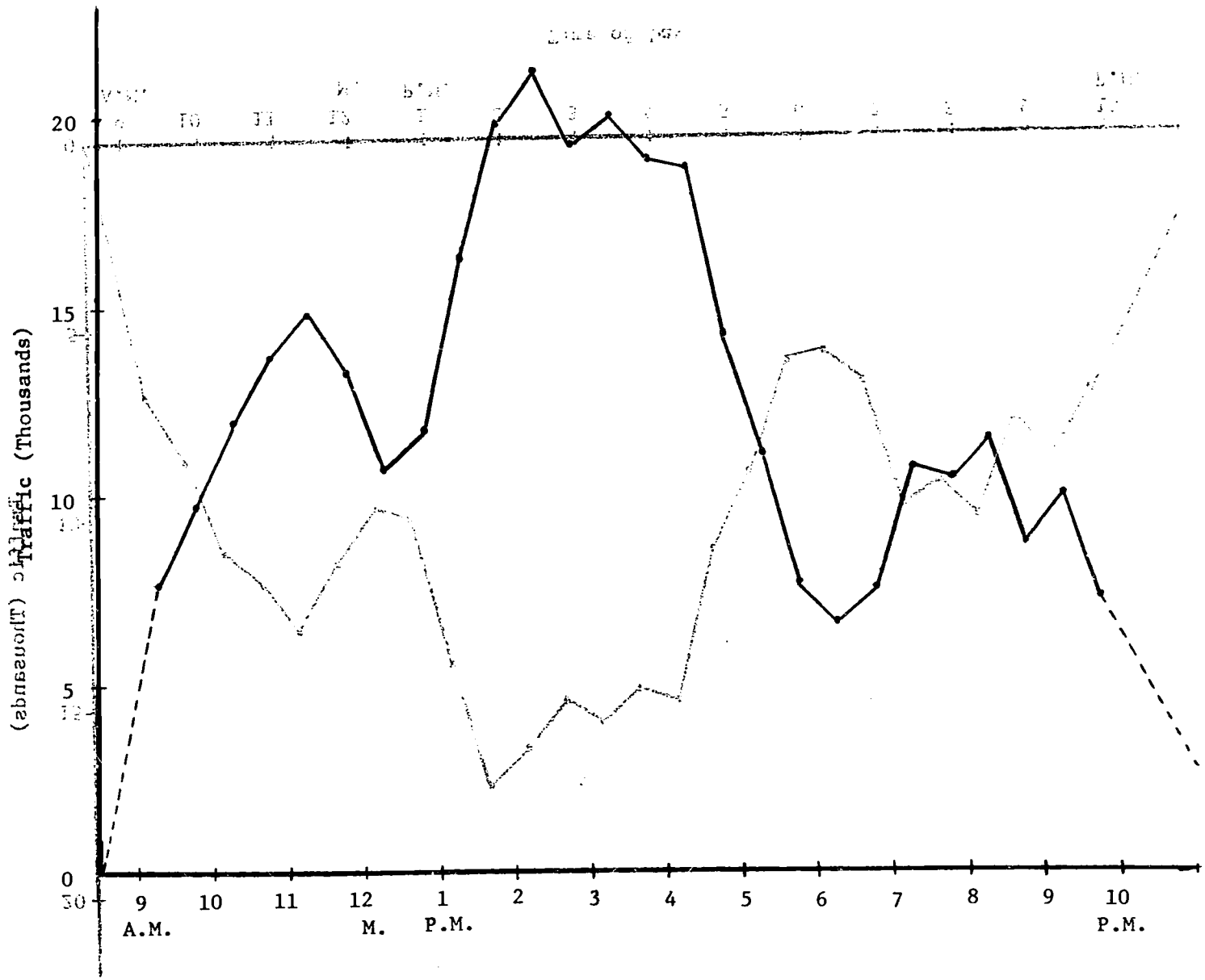
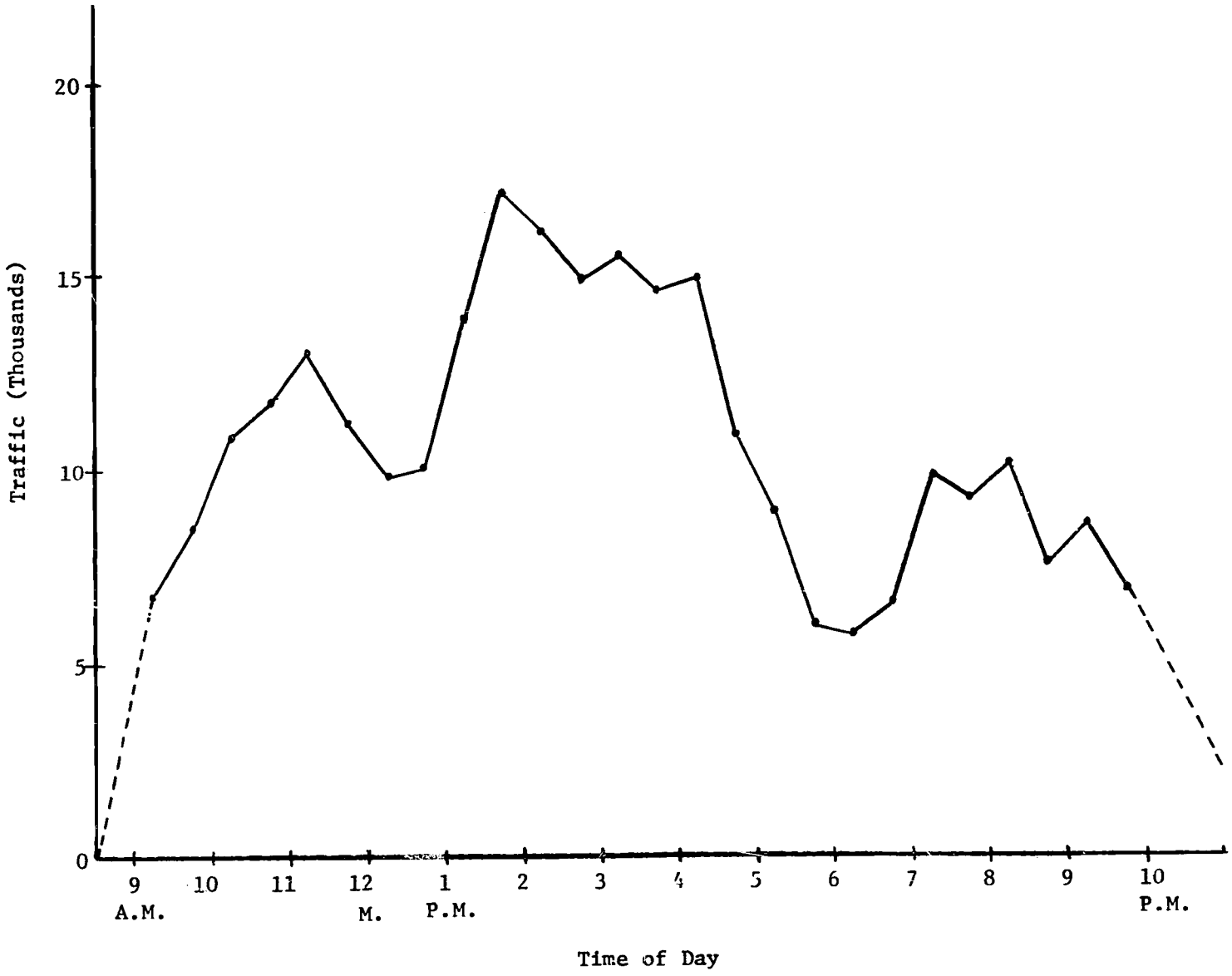


Figure 4. Yearly Weekday Catalog Traffic by Half Hour of the Day



non-users ("walkthroughs") permitted estimation of the correct number of real users of the catalog. Determination of the amount of time which real users spend at the catalog was also attempted through a regular sampling program in which the first users entering the catalog during specified 8-minute intervals were followed in order to note the amount of time spent in catalog use. At the end of the 8-minute period, a user (who could have entered the catalog at any time during the interval) might still be using the catalog, but the timing was broken off with only the notation that the catalog use was incomplete. Reconstruction of the profile of duration of catalog use from such data is a difficult matter, but not hopeless. The profile that emerges shows a peak for the most frequent catalog use period at 2 minutes. There are only about half as many 1-minute users. The number drops off from the 2-minute peak to about half of the peak value at the 4-minute interval; it decays slowly to a negligible value for intervals beyond a half-hour or so. Since there were no actual measurements of the longer intervals, it should be understood that this description is somewhat hypothetical. It is based on the assumption that there are two different normal populations of catalog users, with modes centering around the 2-minute and the 6- or 7-minute intervals. Such an assumption fits well with the data actually observed. The standard deviation for the second group is much broader than for the first, although the actual populations seem to be of roughly the same size.

Traffic statistics revealed no flaws in the design of the interview sampling method adopted in this study. On the contrary, the complexity of the traffic pattern strongly justifies the original decision to conduct interviews throughout an entire year, with representative coverage of different days of the week, hours of the day, periods within the hour,

and portals of entry to the catalog area. The interviewing schedule was deliberately non-representative in its coverage of the summer period as compared to the academic year, and in its coverage of evening and weekend hours as compared to weekday working hours. But simple weighting factors could be, and were, applied to the results of interviews that are completely representative of catalog use with respect to season, day of week, time of day, portion of the hour, and portal of entry to the catalog area. These results are discussed in the sections that follow.

Catalog Users

During the period March 25, 1968, through April 5, 1969, the interviewing schedule specified times and catalog portals for a maximum of 2699 possible interviews. The actual number of interviews conducted during this period was 2134, consitituting approximately two-thirds of a percent of the estimated 320,000 persons who actually made use of the catalog during the same period. Of the 565 scheduled times in which no interview was conducted, 384 are accounted for simply because no catalog users appeared at those times (six-minute intervals). Another 161 instances are due to illness or inadvertent absences of the interviewers or to misunderstandings regarding assignments. Only 20 resulted from the refusal of catalog users to grant interviews (less than 1 percent of the catalog users approached).

The extremly high degree of cooperation of catalog users was very gratifying; and it lends extra credibility to the findings of the study. Most interviewees were very pleased to be asked about their needs, and discussed them without much probing by the interviewer. Most of those who refused to be interviewed did so apologetically, explaining that they would be late for classes or appointments.

Table 3, based on adjusted data from interviews, shows the degrees of catalog use by the various groups within the academic community. Graduate students dominate over other distinguishable groups in terms of absolute use. However, in terms of use as related to size of the eligible group population, the Yale upperclassmen show somewhat greater use of the catalog than do graduate students. Yale faculty rank below undergraduates and graduate students in terms of absolute use of the catalog. In relative use they are probably comparable to freshmen (if one defines faculty to comprise a group of about 2000); but it should be remembered that unlike freshmen, faculty members can make heavy use of departmental libraries, can use assistants for library work, and can often go directly to desired subject areas in the Sterling Memorial Library stack without consulting the catalog. Other Yale employees rank below faculty, and wives and family of faculty rank last in use among people connected with the university. However, non-Yale students, non-Yale faculty, and other "outside" users account for a total of 19 percent of catalog use--a very respectable proportion.

The cross-section of catalog users varies with season. During the summer period, there is, understandably, relatively less use by Yale groups as compared to visitors, and less by undergraduates as compared to graduate students. Relative use by faculty and staff doubles in the summer period over the academic year. But since the rate of use of the catalog falls to half of the academic level during the summer, this merely means that total use per week by faculty and staff remains just about constant throughout the calendar year. Use by graduate students, although constituting a steady relative proportion throughout the year actually falls in the summer to half of the academic year level in terms

Table 3

Academic Status of Catalog Users by Season

Status	Aproximate Eligible Population	Percent of Catalog Users		
		Full Year	Acad. Year	Summer
Yale freshmen	1600	7.0	8.3	12.3
Other Yale undergrads	2900	24.9	27.7	
Grad., postgrad. students	4700	35.5	35.6	34.9
Yale faculty	5000	8.8	7.3	16.0
Yale staff		3.3	2.9	5.0
Non Yale students	?	9.5	9.9	7.6
Non Yale faculty	?	3.2	2.8	5.3
Yale fac. family	5000	1.5	1.1	3.3
Visitors, retirees, others	?	6.3	4.3	15.7
		100.0	99.9	100.1

of instances of use per week. Actual instances of use per week by visitors doubles in the summer period. Use per week by non-Yale faculty remains about the same; and use per week by non-Yale students declines, but not as sharply as use per week by Yale undergraduates. Use per week by faculty family increases by half in the summer.

Catalog Searches

Four basic types of catalog search objectives were identified in the study; they have been designated as document search, subject search, author search, and bibliographic search. In a document search, (often called a "known item" search) the catalog user is aware of the existence of some particular book or publication that he wants to locate. In a subject search, the catalog user is interested in both identifying and locating one or more documents pertaining to some known topic. In an author search, the catalog user is aware of some author, publisher's series, or other source of literature and is interested in identifying and possibly selecting specific documents from that source. In a bibliographic search, the catalog user is interested in using the catalog itself to supply or verify bibliographic information regarding a known document; he is not interested in locating and using the document.

The distribution of searches among these four basic types is given in Table 4. The distribution was determined in two different ways, yielding two different results. The first column is based on the immediate objective of the catalog user at the moment of his approach to the catalog. Most of the questions asked during an interview pertained to this immediate objective.

Table 4

Distribution of Search Objectives

Search Type	Immediate, %	Underlying, %
Document	73	56
Subject	16	33
Author	6	6
Bibliographic	5	5
	100	100

However, it was hypothesized that some users may try a search approach for which a library catalog is particularly suitable as an indirect means of performing a different type of search for which the catalog is not as well adapted. Specifically, it was hypothesized that some of the people performing document searches were really interested in subject information and were first seeking such information, for convenience, in known documents that they considered likely to contain the desired information but that they did not regard as the exclusive objectives of their searches. This hypothesis was tested by means of a simple question asked at the very end of an interview with a catalog user whose immediate objective was a document search. The user was asked what he would do if his intended document search should be unsuccessful--whether his search would end there, or whether he believed he might find what he wanted in some other publication. (For obvious reasons it was not necessary to ask this question of document searchers who were looking for works of fiction or for items on lists of assigned reading.)

The responses to this question revealed a rather dramatic difference between immediate objectives and underlying objectives in catalog searches. The distribution of underlying objectives is given in the second column of Table 4. It indicates that about a third of the catalog users are basically interested in subject or topical information, but that half of these users attempt to use a document search to make do for a subject search. In terms of underlying interest, document searches account for only 56 percent (not 73 percent) of catalog use.

It is interesting to note that no significant variations in the

distribution of search objectives with respect to season of the year, academic status of user, departmental affiliation, or newness to the Yale library were detected in this study.

In searches where the immediate intent is to locate a particular document, 80 percent of these desired documents are monographs, and 20 percent are articles in periodicals.

On the average, about 26 percent of desired documents are already known, through previous contacts in the same library or elsewhere, to the users who seek them (27 percent of the monographs, 22 percent of the periodical articles). Another 22 percent of the periodical seekers have had some contact with the periodical desired but not the specific article desired. Likelihood of previous contact with the desired document increases with years of library use, from 22 percent (25 percent of monographs, 10 percent of periodical articles) for users in their first year of experience with the library to 52 percent (58 percent of monographs, 26 percent of periodical articles) for users with more than twenty years of experience. There is a curious interruption in this general trend that occurs among users with seven to nine years of experience; these individuals were found to seek fewer familiar documents than any other group: 18.5 percent (16 percent of monographs, but 30 percent of periodical articles).

In all bibliographic searches and in 98 percent of the document searches, the user felt able to state whether his search was successful or unsuccessful as soon as he was finished with the catalog. However, for 40 percent of the subject searches and 30 percent of the author searches, the user stated that he would have to defer judgement on success or failure

until he had looked at specific books or browsed specific stack sections identified through the catalog. In all, about 91 percent of the catalog searches studied were evaluated by the users as successes or failures right at the catalog, and 9 percent were evaluated only after further effort elsewhere in the library.

Results of document searches showed that almost 84 percent turned up the desired item (with one or more additional useful items identified in 9 percent of these searches). Another 2 percent turned up useful supplemental documents, but not the specific document originally desired. Only 14 percent of the searches turned up nothing at all.

Although data for subject and author searches are less complete than for document searches (because it was not possible to obtain full reports on all searches continued away from the catalog), it appears that there is no great difference in success rate. Of course, successful author and subject searches tend to turn up larger numbers of documents than successful document searches, since that is almost always the intent of the catalog user.

Users who had come to the catalog to carry out a document search were asked to state their intended approach to the catalog. Results are shown in Table 5. The author approach dominates. The title approach is next. Subject and editor approaches are rare compared to author and title approaches

The decision of some catalog users to conduct a document search by looking up a subject term may seem strange (4 percent in Table 5), but

Table 5

Intended Approach to Document Searches

	Percent
Author name (personal or corporate)	62.
Title of monograph or periodical	28.5
Subject term	4.5
Editor name	4.
Author or title of portion of work (analytic)	.5
Series title	.2
Other (publisher, translator, geographic location, etc.)	.3
	100.0

it can be thoroughly rational. If the author search would involve looking through a vast number of cards (e.g., when it is a "U.S." or "Great Britain" main entry, or when the author has a common last name and the user does not know his given name), it may take less time to find the document under an obvious subject entry, providing there are not too many cards under that particular subject.

Look-up by author or title of an analytic is, in general, a fruitless approach to the catalog. This approach is found exclusively among users who have had less than two years of experience with the Yale libraries. Conversely, look-up by series title (which can be productive, but not necessarily) is observed much more frequently among experienced users than among those with less than 2 years of experience.

Users in their first two years of experience account for 55 percent of catalog use. The first year pattern is close to the over-all average with respect to author approaches, title approaches, and subject approaches, but lower with respect to editor approaches. The second year pattern is higher than the over-all average with respect to title approaches, but lower on author approaches and subject approaches; these disparities disappear in the third year.

Statistical analysis did not suggest any particular difference in average search success as a function of experience. If there is any trial-and-error learning phenomenon to be found, it cannot be very prominent. Nor was there any statistical suggestion that newcomers tend to become frustrated and tend to avoid the library toward the latter part of their first year. Newcomers who use the catalog seem to know pretty

much what to expect from the beginning. Although they may make occasional false starts, they still tend to find what they seek about as often as the more experienced users. This indication is not very surprising when one considers only those newcomers who are graduate students or faculty members--all of these are familiar with similar libraries at other colleges and universities. However, the apparent absence of a conspicuous learning phase for freshmen is more puzzling. A possible explanation is that there really is such a learning phase for freshmen but that it is hidden behind an abnormally high a priori probability of success in the types of searches undertaken by freshmen. More than one third of the catalog searches attempted by freshmen are for documents listed on printed course assignment lists. Starting with such accurate, well-formatted, locally tested reference lists, the probability of success is very high. For upperclassmen, the proportion of searches based on such lists is only one sixth; for graduate students as a group it is one ninth.

Search Success and Potential for Improvement

It was reported in the previous section that 16 percent of attempted document searches were unsuccessful. A special follow-up study of 256 unsuccessful searches was made in order to learn more about the reasons for failure: At the conclusion of the interviewing period, the research staff attempted to re-run the unsuccessful searches, using only the clues provided by the catalog users during the original interviews. In these follow-up searches, considerable use was made of such reference works as Books in Print and Union List of Serials to save time in fully identifying some of the desired documents before looking for them in the catalog.

Some 31 percent of the follow-up searches turned up the desired

document in the catalog under entries known to the original catalog user at the time of the interview. These searches had apparently failed because of faulty search technique or because of failure to persevere. Almost 60 percent of the follow-up searches had potentially adequate search clues to documents that did not happen to be in the library collection at the time of the search; one fifth of these had been added to the catalog between the dates of interview and follow-up. Less than 10 percent of the follow-up searches (about 1 percent of all attempted document searches) could not be evaluated because of inadequate or clearly inaccurate clues; it is possible that the desired documents for this group were all in the collection. Table 6 summarizes the over-all results of document searches in terms of complete success or the three types of failure described above.

An apparent conclusion from this study of failed document searches is that there is more room for improvement in catalog service through instruction of users in the proper use of the catalog (5 percent potential improvement) than there is through expansion of approaches to the catalog (1 percent potential improvement) at least for document searches. An even more obvious conclusion is that more and faster acquisition and cataloging of new books are desirable (10 percent potential improvement) as a means of improving service; but this is a truism in libraries.

The modification or expansion of catalog entries in the existing catalog apparently has the least potential of all three possible approaches to improvement of catalog service (i.e., coverage, user education, and modification). Of course, this conclusion considers only the "absolute" retrievability of the types of documents which catalog

Table 6

Success and Failure of Attempted Document Searches

	Percent
Document in catalog, located by user	84
Document in catalog, not located by user but found by research staff through user's starting clues	5
Document definitely or probably exists but was not in catalog at time of user's search	10*
Document possibly in catalog, user's clues inadequate or grossly inaccurate	1
	100.0

* One fifth of this group of documents were added to the catalog from 1 to 12 months after the user's unsuccessful search.

users seek currently. It does not consider the convenience factor in catalog searching. One could conjecture that some catalog users fail to succeed in their searches not out of ignorance, but simply because they are confronted with too many cards to look through under a valid entry. If so, catalog modification may be more desirable than it appears. This would seem especially plausible for the subject and author searches that are not represented in Table 6. One could further conjecture that making the catalog more convenient to search would tend to bring more users (back?) to the library to make searches based on clues that they know would be inconvenient or unproductive with the catalog as currently constituted.

Traffic measurements bearing on the question of catalog convenience are inconclusive. As was stated earlier, no clear evidence could be found of a frustration factor (diminishing catalog use) among freshmen and other newcomers in their first year of experience with the library. In fact, freshmen tend to increase their frequency of catalog use as they become upperclassmen (Table 3); but this could be explained as merely an involuntary requirement of their academic programs or else as a heightened awareness of the positive aspects of the existing catalog.

Some fragmentary evidence on the value of different approaches to catalog improvement can be derived from comments which were offered gratuitously to interviewers by catalog users during the data gathering phase of this study. Some 75 individuals stated complaints or suggestions about the library. These were noted and preserved. About half of the comments reflected on catalog coverage, catalog user education, or catalog design. Of this group, 8 comments indicated a desire or need for more

information regarding the use of the catalog (general orientation, interpretation of abbreviations in catalog, determination of language of cataloged work, transliteration and filing rules for non-Roman alphabets, interpretation of subject class numbers). Eight comments had to do with improving the physical convenience of catalog use. Nine comments requested more cross references in the catalog, of several types. Three requested more convenient access to periodicals by title. Two complained about the generality of subject headings; one complained about the inconsistent treatment of a particular topic that appears in various subject subsections of the catalog. Seven wanted the catalog cards to provide more collation and notes information about the cataloged works. Four wanted the catalog to provide access or better access to certain types of literature (journal articles, dissertations). These comments do not suggest any unanimity among catalog users as to a single best approach to improvement of catalog usefulness. They suggest only that there is interest in improvement along all lines of approach.

The matter of catalog coverage (size of collection, rate of acquisition, promptness of cataloging) is beyond the scope of this study and will not be discussed further. The general question of user education is also beyond our scope. The remainder of this chapter on project results will deal primarily with findings on various aspects of catalog responsiveness, both current and potential. Some of these findings suggest specific needs and opportunities for better orientation of users. But the main objective will be to try to clarify how the catalog does respond and might respond to the demands which users actually make of it currently.

Search Clues, Catalog Access, Catalog Data

As Vickery has pointed out (8), there are four functions that can be served by the data elements, or items of information, on a bibliographic record such as a catalog card: As a group, the elements serve to identify (by description) some specific publication. Each individual element can, at least in principle, serve as an entry for retrieving the collective bibliographic record from a file such as a catalog. One or more elements can indicate the physical location of the publication described by the record. Finally, the symbolic form used for recording an element (usually letters and numerals) can facilitate the sequencing of groups of records for convenient access (usually in alphanumeric order). Curran and Avram (9) have identified and listed hundreds of bibliographic data elements, of recognized or potential value, which might be included and distinguished in the bibliographic records of libraries.

The use of a bibliographic data element as an entry term for a catalog may be justifiable operationally if the users of the catalog tend to possess corresponding search clues they can match against entry terms of that type. The catalog studied in this project, like catalogs in most large research libraries, offers entry or access primarily by author (main entry) terms, title terms, and subject terms. Some observations can be made about the appropriateness and sufficiency of these approaches to the catalog.

It was reported above that document (known-item) searches account for 73 percent of the searches attempted. Virtually 100 percent of the documents have titles (which can be periodical titles as well as monograph titles); 82 percent of the documents have authors (those without authors obviously

Table 7

Author and Title Data in Document Searches

	Author Data	Title Data	Title Data x 82 %
1. Applicability to desired documents.	82 %	100 %	82 %
2. Availability in search clues (right or wrong). Sum of 3-7.	77 %	97 %	79.5%
3. Available, complete, and accurate in search clues.	42.3%	61.5%	50.4%
4. Available in search clues, but only partially complete and accurate (last name, first corporate level, or first title word correct and in place).	23.2%	15.0%	12.3%
5. Available in search clues, inaccurate.	10.5%	15.5%	12.7%
6. Available and accurate in search clues, appears in catalog card but not designated as a file entry.	0.7%	1.6%	1.3%
7. Available in search clues but not on catalog card, accuracy not determined.	0.3%	3.4%	2.8%

being periodicals primarily). Of the documents with authors, almost 92 percent have personal authors as opposed to corporate authors. There is no reason to believe that the pattern for documents in unsuccessful document searches is substantially different from the pattern for documents in successful searches.

By far the most consistently available and usable search clues brought to the catalog by users interested in document searches are clues relating to author and/or title. Details of availability of author data and title data are compared in Table 7. It is clear from the first two columns that, on the basis of absolute count, title data in general, and accurate title data in particular, are available much more frequently than corresponding author data. This finding is in qualitative agreement with results of other researchers (2, 3, 10), some of whom have interpreted them as proof that library cataloging should use titles (rather than authors) for main entries and that users should be encouraged to access catalogs by title rather than author if both clues are available.

The present study does not clearly support such a sweeping interpretation. It should be remembered that the author approach is not even a possibility for some 18 percent of the documents desired by users; for these, the main entry is by title, and that is obviously the approach which users take in searching for these documents. The comparison of author and title approaches should properly be related to the 82 percent of documents for which both approaches are possible. The third column in Table 7 gives an approximate basis for such a comparison by multiplying the absolute title values by 82 percent. (It is fully accurate if there are no differences in title data for authored documents as compared to anonymous documents--which has not been determined.) If one now compares the first and third columns, it can be seen

that the differences between author and title clues, in terms of availability and accuracy, are very much reduced. Titles come out best by a few percent in availability as clues (79.5 percent vs. 77 percent). They are considerably better than author clues where complete accuracy is the criterion (50.4 percent vs. 42.3 percent); but they are actually a bit worse than author clues where the criterion is the sum of complete and partially complete clues (62.7 percent vs. 65.5 percent).

The proper interpretation of these results would seem to be that--at least for the large academic library--neither author nor title has an overwhelming advantage as candidate for main catalog entry and for preferred search approach. Perhaps that is why the controversy on this question has persisted so long in library circles without any concensus. There seems to be an advantage to the title approach, but only a slight one.

One can question why there is such a strong tendency for library users to approach the catalog by author (Table 5) if the title approach is just about as good if not better. Part of the answer must be connected with the nature of the Sterling Memorial Library public catalog. Only three fifths of the entire Yale University collection is represented by full sets of entries in that catalog; the remaining two-fifths is represented only by main entry cards which are usually author entries. Thus, the user of the catalog casts a wider net if he searches by author. However, previous training in library use must also be a factor in preference for the author approach; there is a noticeable tendency for the author approach to be used more heavily during the early hours of the day (when experienced professionals make their heaviest use of the library) than in the evening (when use by students predominates).

Furthermore, uncertainty regarding the completeness or accuracy of search clues can often favor the author approach. It was found in this study that author clues and title clues tend to co-vary with respect to completeness and accuracy. When one is entirely accurate, the other tends to be the same. However, users generally seem to know at least the last name of an author, or a reasonable approach to it. If the name is not too common, the user stands a good chance of finding what he seeks in a brute force author search of limited section of the file. If he knows only the frequently used first word of a title and is not sure of the rest (e.g., "History of..."), he would have difficulty with a title approach. Of course, the title approach would be preferable if he has great confidence in his knowledge of the title and if he has only imcomplete knowledge of an author with a common last name. Statistics on these alternatives were not studied in this project, but the data gathered would permit such a study.

An indication of the potential for improving the success rate of author and title approaches to the catalog is found in lines 6 and 7 of Table 7. For authors, between 0.7 and 1.0 percent improvement could be gained by providing catalog access to names that are prominent in a document's description but that are not, strictly speaking, authors by current definitions. These would tend to be the names of editors, compilers, translators, and study group chairmen. For titles, there is much greater potential for improvement through filing under additional terms--from 1.3 percent to 4.1 percent. The types of title-like terms involved here include subtitles, short titles, series titles, major analytic titles, and popular designations.

It is noteworthy that all of the document searches represented in

Table 7 were successful despite the indicated major deficiencies in users' starting clues. The adaptability of the human being in his interaction with the conventional card catalog must not be overlooked or underestimated when considering the possibilities of the computerized catalog as an alternative to the card catalog. Humans using the card catalog were able to compensate for many inadequacies in completeness, accuracy, and appropriateness of their starting clues. They used several devices in compensating: brute-force searching through fairly large portions of the catalog; sampling of possible alternative spellings; or (quite infrequently) shifting to another type of search approach (including the subject approach). Success was usually determined by degree of agreement with one or more starting clues in addition to the clue used for entry to the catalog.

Achievement of near-human (or, hopefully, better-than-human) facility in compensating for inadequacies in search clues would be essential if computerized catalogs were to replace card catalogs in large research libraries. If there is any mismatch at all, no matter how minor, between search clues and file data in a computer search, the computer will fail to retrieve unless it is given some definite program that will cause it to ignore particular kinds of mismatches. Several methods have already been described in the literature by which computers have been programmed to retrieve from bibliographic files despite errors in search clues. The performance of two of these methods (11, 12) was tested, by manual simulation, on actual search clues and actual corresponding catalog data gathered in this study. Results have already been published (6, 7) and will be summarized very briefly here.

Both of the computer retrieval methods studied made use of both author and title data. One method (11, 6) truncates these data (in both

the search clue and the catalog file), taking a prescribed number of letters from the author's last name and the first and second words in the title. These truncations are matched to achieve retrieval. This method overcomes ignorance of an author's first name as well as errors in endings of words and names. The second method (12, 7) compresses author name and title words according to specific rules for casting out letters and syllables before clue and file are matched. This method overcomes certain types of common misspellings. (Both methods, of course, can cause retrieval of incorrect matches (false drops) as well as correct matches; but this aspect was not studied here.) Data gathered from a sample of 126 successful document searches were used to test the retrieval capabilities of both methods. In 77 searches (61 percent), both the author and title were known perfectly by the catalog user; these documents would therefore be retrieved by any method at all. In 49 searches (39 percent) there were inadequacies in clues for author or title or both. Ambiguities in the methods tested were always resolved in favor of the methods. Both produced the same result in over-all retrieval: 70 percent; in other words, each method was capable of producing about 9 percent more retrieval than simple character-by-character computer matching. Yet human searching had produced 100 percent retrieval with this same set of searches.

Although minor improvements might be made in each of the computer retrieval methods tested, it seems highly unlikely that either method--or any other single method--can be developed to the point where it seriously rivals human retrieval capability. On the other hand, there is reason to believe that combinations of methods can be developed that will approach or surpass human performance. If one were to accept retrieval from either of the two methods tested, the combined performance would go up to 80

percent. New computer methods could be devised to consider the data from many viewpoints at once, just as human catalogers do. However, the use of combinations of retrieval methods will probably cause unmanageable quantities of false drops also, unless devices are incorporated for suppressing retrieval as well as for promoting retrieval. Human catalog users tend to suppress retrieval on the basis of clues other than the ones they are using for file entry (e.g., subject entries, date of publication, place of publication, contents notes, author's birth and death dates). Obviously, machines could do likewise only if data elements other than author and title were available in their memories and only if such additional data elements could be accepted from the user along with his primary search clues.

This point must not be minimized. There is a strong temptation among would-be catalog computerizers to deal with the high cost of computer memory units and processing units by cutting to an absolute minimum the types of information to be included in a catalog record. For example, if a computerized catalog were to be designed for servicing document searches only, the temptation might be to store only author, title, and call number. It should be understood, however, that this would tend to make the computerized catalog inherently incapable of achieving the same retrieval efficiency (selection and rejection capability) as the conventional catalog. The inclusion of other data elements would enhance the computer catalog's retrieval potential, but at a price in storage and processing costs. Further intensive work is required on the relative trade-offs in usefulness and cost of including additional data elements in a computerized record.

Date of publication is probably the most obvious candidate for inclusion in any file intended for use in document searches. After title-like and

author-like search clues, date clues were the next most common clues among catalog users (Table 8).

Since only 59 percent of the catalog users have any confidence in their knowledge of publication date, and since their information when they have it is frequently off by several years, it is clear that date is not a very useful clue for primary access to the catalog. Nevertheless, if primary access by some other clue should not be sufficient to discriminate among many possible documents (as when author or title clues are imperfect), even a poorly known publishing date can be very useful in narrowing the field. A user who searches by author but does not know the author's first name, yet who knows merely that the book he seeks was written after World Was II, can search a large file section rapidly, rejecting at once any item published before 1945. Even more intelligently, he can quickly recognize and reject entire file sections devoted to individual authors with the same last name who are shown by the catalog to have died before 1945. With clues other than date, similar rejection processes are possible, but they can be much more subtle and require study.

So far the discussion in this section has been directed at document searches. It is equally applicable to other types of searches as well. Bibliographic searches are identical to document searches in approach. They differ only in the final object of the search--full bibliographic descriptions, rather than book locations. Author searches are also similar to document searches in that they start with author-like, or title-like entry clues (author, editor, series name, publishing institution). They differ in that their object is to identify a list of possibly pertinent references for the user's further consideration either at the catalog or away from the catalog.

Table 8

Availability and Accuracy of Publication Date Information in Document Searches

"No information" on date	41%
Information more than 5 years wrong	12%
Information 2-5 years wrong	18%
Information 1 year wrong	10%
Correct year	19%

The (overt) subject search differs in approach, since entry must be by a subject heading rather than by an author or title. The process of incremental discrimination is, if anything, much more pronounced in subject searches than in author or title searches.

Users engaged in subject searches frequently complain that subject sections in the catalog are much too large and general, rarely narrowed sufficiently to cover only the particular subject aspect of interest to the user. Consequently, the user is forced to deal with a file section containing large amounts of unwanted material. He copes by going through this section rapidly, scanning the cards for clues by which he can select or reject them. Most subject searchers are interested in retrieving only the few most pertinent items; they are usually not interested in building comprehensive bibliographies of conceivably pertinent items. Just about any category of information that appears on a catalog card can be helpful to a subject searcher at one time or another. A scope note or contents annotation can often clinch the pertinence of a document. An informative title or subtitle, or an additional subject heading, can sometimes do the same thing. The author, if his name is familiar to the searcher, can serve as a basis for selecting or rejecting an item. In many subject searches, material can be accepted or rejected on the basis of age, as indicated by publication date for example. Material is often rejected on the basis of language. The class number is often valuable as a guide to later browsing activity. When a subject search turns up numerous items of equally uncertain pertinence, users frequently make use of collation information (especially the number of pages) as a means of identifying the items that can be scanned and handled with the least effort. The extent to which a searcher employs these devices appears to be related

to both the original objective of his search and to the quantity of possibly useful references turned up by his initial search clue. Data collected in this study have not yet been studied in a quantitative way with respect to this phenomenon; however, it seems fairly clear that there would be general benefit to subject searchers if catalog cards were filed by publication date within a given subject heading rather than by main entry.

One of the questions to which this study was originally addressed is whether it would be possible to derive useful catalog records directly from newly acquired documents (e.g., as and when adequate print reading devices are developed) with little or no creative input from a professional cataloger. To facilitate study on this question, the "front matter" of retrieved documents (title pages, contents, preface, index, etc.) was scrutinized, and photocopied when justified for comparison with catalog cards and with user clues for the same documents. It is quite clear from only qualitative perusal of this material that there is at present no possibility of deriving an efficient all-purpose catalog in this manner without extensive human intervention.

There does appear to be some promise to the idea of providing for only document searches and bibliographic searches by this approach, but there are distinct disadvantages along with the advantages. From the user's viewpoint, it would be an advantage to have a machine provide for access by <u>every</u> title and subtitle and series title appearing in the front part of a book, rather than by only a human-designated "correct" title. On the other hand, a machine might not be as helpful as a human in identifying the first significant word in a title for filing purposes. At the very least, a human would have to tell the machine what language it was dealing

with. In the case of authors, it is clear that there can be definite disadvantages to simplistic cataloging. The title page of a book will sometimes contain only the initials of an author who is generally known by his full name; retrieval would still be possible, but more difficult, especially if the author's last name and initials are shared with many other authors. Furthermore, the title page of a printed document does not necessarily tell the truth or the whole truth (13), as when there are falsified publication dates, omitted edition data, pseudonymous authors, etc. Scholarly assistance with these matters by human catalogers may not be indispensible but it is certainly of tangible value. Further study of the economic factors (cost versus benefit) is needed.

Possibilities for achieving adequate subject cataloging through purely automatic processing of front matter appear to be even less promising than for descriptive cataloging. It is apparent from scanning the data assembled in this study that automatic subject cataloging (based on title page, contents, preface, index, etc.) would very rarely provide positive retrieval on as little as a two-term coordination of a subject searcher's starting clues. Sometimes the concepts are present, but are stated in terms that are synonyms or partial synonyms of the terms known to the searcher. So, at the very least, a mechanism would be required for dealing with synonyms, either at the time of cataloging or at the time of searching. Quite often books lack informative titles or informative chapter titles; and they very frequently lack indexes. When indexes are included, they can be very long and full of trivial topics which a computer could not easily distinguish from an important topic. Thus, clutter in the computer memory would be a very serious problem if automated subject cataloging were to be attempted. Despite all of the drawbacks enumerated here, however, the idea of automated

subject indexing does have some merit. It appears that such an approach would usually provide at least a one-term match with the clues of the subject searcher; if adequate secondary selection-rejection possibilities were built into the catalog system, this might conceivably suffice to achieve adequate performance. This question requires further quantitative study.

Discussion

The results of this study are described in the previous chapter; they are listed briefly in the "Summary" chapter at the beginning of this report.

Librarians can undoubtedly draw meaningful conclusions from some of these results. But it is important to note that the conclusions to be drawn will depend upon the librarian's general outlook. The librarian who believes strongly in expansion of library services will derive quite a different message from the librarian who believes very strongly in reduction of processing costs.

For example, consider the findings on immediate search objectives (73 percent document searches, 16 percent subject searches) and underlying search objectives (56 percent document searches, 33 percent subject searches). The expansive librarian might well conclude that he should put more effort into improving the subject approach to the catalog so that fewer users will be forced to sublimate their subject searches as document searches. The retrenching librarian, on the other hand, might conclude that the subject approach to the catalog should be eliminated entirely since this would do less harm to the utility of the catalog than the elimination of the author or title approaches. These two conclusions would be completely opposite, yet there is logic in each of them.

Flexibility, of course, is the rule in successful management.

Librarians pick and choose among alternative courses of action in order to achieve the best possible results with the limited resources available. The interpretation of the results of this study requires knowledge of the trade-offs of costs versus benefits--and more study is needed along that line. Nevertheless, some general observations may be in order.

The fact that 5 percent of all document searches fail even though the catalog users have adequate starting clues suggests that strong consideration should be given to improved user orientation and user assistance. User education, or self-education, methods need not be very costly; and it would be relatively easy to determine whether or not they are effective.

The fact that 10 percent of all document searches fail because the collection lacks the desired document suggests that something might be done to acquire more books in anticipation of need and to provide better notification of books that are on hand or on order but not yet cataloged. (A step in this direction at Yale was the placement of a copy of the "in-process" list in the catalog area some months ago.)

Increasing the complexity and accessibility of the catalog offers comparatively little potential for improvement of the success rate of searches currently attempted. However, improving the convenience of catalog use might attract heavier use of the catalog and the library collection. Providing access through a greater variety of title-like entries is a promising approach to improvement of document search convenience. Filing by date within subject headings is a promising approach to improvement of subject search convenience.

Data elements other than author, title, and subject are of definite value in resolving many searches in which the entry clues are ambiguous or inaccurate. Such data elements should probably not be abandoned entirely, even in a computerized catalog where data storage is very expensive. Further study is warranted on the costs and benefits of acquiring, storing, and retrieving such data elements, in order to determine their relative values or expendability.

There is promise in the idea of using automated techniques for catalog construction. However, it is unrealistic to expect impressive retrieval performance from such catalogs if they contain only information copied directly from input documents without some degree of annotation and association, whether human-supplied or computer-supplied.

References

1. Sidney L. Jackson, Catalog Use Study, Director's report, American Library Association Resources and Technical Services Division, Cataloging and Classification Section, Policy and Research Committee, Chicago: American Library Association, 1958, 86 pp. Edited by Vaclav Mostecky.

2. Mary F. Blackburn, University of Chicago Graduate Library School, private communication, 1967.

3. Renata Tagliacozzo, Lawrence Rosenberg, and Manfred Kochen, Access and Recognition: From User's Data to Catalog Entries, Communication No. 257, Ann Arbor, Michigan: University of Michigan, Mental Health Research Institute, October 1969, 40 pp.

4. Ben-Ami Lipetz and Peter Stangl, "User clues in initiating searches in a large library catalog," pp. 137-139 in American Society for Information Science, Proceedings, Vol. 5. Annual Meeting, October 10-24, 1968, Columbus, Ohio.

5. Ben-Ami Lipetz, "A quantitative study of catalog use," paper presented at Clinic on Library Application of Data Processing, University of Illinois Graduate School of Library Science, Urbana, Illinois, April 27-30, 1969. In press.

6. Peter Stangl, Ben-Ami Lipetz, and Kathryn F. Taylor, "Performance of Kilgour's truncation algorithm when applied to bibliographic retrieval from a library catalog," pp. 125-127 in American Society for Information Science, Proceedings, Vol. 6, Annual Meeting, October 1-4, 1969, San Francisco.

7. Ben-Ami Lipetz, Peter Stangl, Kathryn F. Taylor, "Performance of Ruecking's work-compression method when applied to machine retrieval from a library catalog," Journal of Library Automation, Vol. 2, No. 4, pp. 266-271 (1969 December).

8. B. C. Vickery, "Bibliographic description, arrangement, and retrieval," Journal of Documentation, Vol. 24, No. 1, pp. 1-15 (1968 March).

9. Ann T. Curran and Henriette D. Avram, The Identification of Data Elements in Bibliographic Records, Final report of the Special Project on Data Elements for the Subcommittee on Machine Input Records (SC-2) of the Sectional Committee on Library Work and Documentation (Z-39) of the United States of America Standards Institute, May 1967, 117 pp. (sections paged separately).

10. F. H. Ayres, Janice German, N. Loukes, and R. H. Searle, "Author versus title: a comparative survey of the accuracy of the information which the user brings to the library catalogue," Journal of Documentation, Vol. 24, No. 4, pp. 266-272 (1968 December).

11. Frederick G. Kilgour, "Retrieval of single entries from a computerized catalog file," pp. 133-136 in American Society for Information Science, Proceedings, Vol. 5, Annual Meeting, October 20-24, 1968, Columbus, Ohio.

12. Frederick H. Ruecking, Jr., "Bibliographic retrieval from bibliographic input; the hypothesis and construction of a test," Journal of Library Automation, Vol. 1, No. 4, pp. 227-238 (1968 December).

13. Jacob Blanck, The Title-Page as Bibliographical Evidence, School of Librarianship, University of California, Berkeley, 1966, 16 pp.

APPENDIX A

B. A. Lipetz and P. Stangl
"User clues in initiating searches in a large library catalog," pp. 137-139 in American Society for Information Science, Proceedings, Vol. 5, Annual Meeting, October 10-24, 1968, Columbus, Ohio.

USER CLUES IN INITIATING SEARCHES IN A LARGE LIBRARY CATALOG

Ben-Ami Lipetz and Peter Stangl
Research Department, Yale University Library
New Haven, Connecticut

Introduction

A library catalog is intended to make it relatively easy for a library user to identify and locate desired items in a collection. The catalog is a bridge between the information which a user brings to the library (in the form of written notes, or remembered clues) and the information or documents he hopes to carry away from the library. In order to assess the adequacy of performance of a catalog, one must study how efficiently the catalog matches the initial clues to the desired items. Such a study has been undertaken at the Yale University Library under a grant from the U. S. Office of Education.

Two purposes may be served simultaneously by a study of catalog performance. In the short run, there is the possibility of effecting minor improvements in the traditional library card catalog through the identification of desirable modifications in cataloging practice. In the long run, there is the possibility of effecting major improvements in catalog performance in the future as catalogs are converted from card files to computer files. At the very least, there is the long-range prospect of forestalling costly blunders during the transition from cards to computers. Computerization should be addressed to the needs of users and the capabilities of computers; to simply attempt to mechanize existing card catalog systems would be to institutionalize the shortcomings of card catalogs rather than eliminate them. But shortcomings of card catalogs must be identified before they can be eliminated.

Although a great many prior studies of library use have been reported, especially as master's theses in library schools, they are of little, if any, use for the purposes stated above. In almost all cases, they are based on questionaires or interviews administered after (sometimes long after) users had finished whatever they had come to the library to do. They elicited considerable information on what the users actually accomplished in the library, but very little information on what the users originally hoped to accomplish when they came to the library. In most cases, no attempt was made to determine the precise search clues, recorded or unrecorded, available to the users when the search began. Because of the frailty of human memory and because of the tendency of users to change their search objectives after they run up against the realities of a less-than-ideal system, it would seem that the only way to determine initial search clues is through interviews conducted at the beginning of a library search, not afterward. No reports of such studies were found in the literature, except for a few in which the samples were so small and poorly described as to be useless for interpretation. After the present study was begun, it was learned that a report is in preparation(1) on a study of this type which involved 100 interviews with users of the University of Chicago Library catalog, all conducted during the summer vacation period. After the present study got under way another apparently similar study was begun at the University of Michigan Library under sponsorship of the National Science Foundation(2). Every effort will be made to compare results of similar studies.

Design

The present study examines catalog user objectives, starting clues, and catalog responsiveness in a large academic and research library. The library under study is Sterling Memorial Library, the largest library unit at Yale University. This library has a book collection of 3 million volumes. Its card catalog is of the single-alphabet type, covering all of the volumes in the building with full cataloging and the balance of the 5-million volume collection of the Yale University Library system with main-entry (author) cards. The catalog is used by undergraduates, graduate students, faculty, university staff, and visitors. Although there are a number of departmental library collections at the university, the Sterling Memorial Library collection is the major research collection for many departments, and an important back-up collection for all other departments.

Five distinct data-gathering activities are included in the study:

1. Gross Statistics on Catalog Use

Parameters of catalog use are determined by simple observation, counting, and timing of traffic in the catalog area. This is necessary in order to provide a sound basis for selection of a representative sample of catalog users to be interviewed. Statistics being collected include the number of users entering the catalog area versus time of day, day of week, and day of year. Although not essential for sample determination, statistics are being collected also on the amount of time spent in the catalog by users, the number of card drawers consulted per catalog use, the frequency of use of individual card drawers, the number of cards searched per catalog use, and the number of charge-out slips filled in per catalog use. Such statistics are of immediate administrative interest as well as of more general research interest.

2. Interviews Preceding Initiation of Catalog Searches

Using an objective selection technique, interviewers approach catalog users at the moment of initiation of a search and attempt to determine the users' objectives and the precise clues with

which they start. This is the most critical phase of the entire study. Selection of interviewees is made objective by tying it only to the clock; each interviewer must interview the first person entering a given portal to the catalog area after a time specified in advance on a prepared schedule. The schedule reflects the observed density of catalog traffic for each day and time of day.

Interviews are conducted according to a rigid plan which was shaped and tested over a period of months. A nondirective approach is used at the beginning of the interview, to minimize interviewer bias and to encourage free expression by the interviewee. As the interview progresses, questions become more direct in order to fill in important details. Responses are recorded by the interviewer on a form, then coded and transcribed at a later date for computer-aided analysis. To save time and preserve accuracy (or inaccuracy) of spellings, a photocopier is used to copy notes and bibliographies which users bring to the catalog.

3. Follow-Up Interviews

Each person interviewed at the start of a catalog search is observed unobtrusively until the search seems to have been completed or terminated. The catalog sections searched by the user are noted. He is then approached once more and questioned briefly about his success. If he has retrieved call numbers which satisfy his requirements, these numbers are recorded to permit later inspection of catalog data and corresponding documents. In cases where the user indicates that his search will be continued through inspection of potentially pertinent documents located through the catalog, a further follow-up interview is requested to determine the results.

4. Examination of Catalog Cards and Catalog Structure

Using actual clues and actual research results obtained from interviews as starting points, the existing card catalog can be examined at leisure. The cards can be examined to see whether they actually contain the types of clues that are brought to the catalog by the users, or, indeed, whether they contain clues that are not wanted. The file arrangement and file headings can be compared with the search approaches taken by catalog users. (This is particularly interesting at the start of the school year, before the new students and faculty become familiar with the existing catalog structure.) Hypotheses regarding possible changes in cataloging rules or catalog arrangement can be tested against observed user requirements.

5. Examination of Retrieved Documents

Books identified as pertinent as a result of catalog searches can be borrowed for examination after the catalog user is finished with them. By examining such a book, especially its front matter (cover, title page, contents, preface, etc.), one can judge whether a hypothetical change in cataloging practice could have provided a more convenient match to the clues brought to the search by the interviewee. One can also judge the extent to which mechanized techniques, e.g., automatic print readers, could have satisfied the search requirement.

Work Accomplished

Data collection began in early October 1967, with the initiation of observation of the gross characteristics of traffic in the catalog. This form of data collection will continue for approximately 18 months, to provide a full-year profile and an indication of year-to-year fluctuation and trend.

Interviewing began on a pilot basis in early October 1967. A total of 200 pilot interviews were conducted by three interviewers before interview structure and technique were standardized. Interviews for actual data analysis were begun in March 1968. At this writing (late April) some 200 "production" interviews have been conducted, and virtually all of these had associated follow-up interviews. Refusals of interviews have been negligible; user cooperation with interviewers has been outstanding. Names of interviewees are not asked, but academic status and departmental affiliation are recorded. It is anticipated that 2000 or more production interviews will be completed in the 12-month period during which interviews will be conducted.

Data gathering on catalog cards, catalog structure, and content of retrieved documents is in the earliest stages at this writing.

Results

The small amount of information collected so far does not warrant any firm or provocative conclusions. It has been observed, however, that the initial search objectives of catalog users can be divided into four distinct types. These are, in order of decreasing frequency:

1. Specific document search, in which the objective is to locate a document already known to exist.

2. Subject search, in which the objective is to identify a document or documents corresponding to a specific subject.

3. Document group search, in which the objective is to identify a document or documents corresponding to a specific bibliographic description (e.g., any books by a given author, any books in a given series).

4. Bibliographic data search, in which the objective is only to retrieve specific information from the catalog card itself and not to identify or locate a document (e.g., completing a reference for a bibliography).

Of the production interviews conducted so far, 80 percent were concerned with the first type of search. It is possible that this percentage may change as more data are collected; there may be large fluctuations in catalog use patterns throughout the academic year. Detailed findings will be given in future reports.

Acknowledgments

Thanks are due to Frederick G. Kilgour for the initial suggestion of the desirability of a study of this kind. Herbert A. Menzel, New York University, provided valuable advice on interviewing technique. Carolyn S. Potts and Katherine M. Skor have given able assistance in collection and analysis of data.

References

(1) Blackburn, Mary F., thesis research project, University of Chicago, Graduate Library School.
(2) Kochen, Manfred, private communication.

APPENDIX B

B. A. Lipetz
"A quantitative study of catalog use," 8 pp. Paper presented at Clinic on Library Applications of Data Processing, University of Illinois Graduate School of Library Science, Urbana, Illinois, April 27-30, 1969.

A QUANTITATIVE STUDY OF CATALOG USE

Ben-Ami Lipetz

Head, Research Department
Yale University Library
New Haven, Connecticut

Presented at Clinic on Library Application of Data Processing, conducted by the University of Illinois Graduate School of Library Science, Urbana, Illinois, April 27-30, 1969.

Among people who are concerned with the management of libraries, it is now almost universally accepted that the traditional manual card catalog must sooner or later be replaced by an on-line computerized catalog of some sort. This is accepted almost as an article of faith; there is almost never any questioning or disputing of its inevitability. I have no intention of questioning or disputing its inevitability in this paper. But there are questions regarding the computerizing of library catalogs which ought to, and indeed do, trouble conscientious library managers. These are the crucial questions of how to computerize and when to computerize. The work I will report on was prompted mainly by concern with these questions.

The notion of computerized library catalogs has been with us for many years. Computerized library catalogs were, in fact, set up at libraries here and there as far back as a dozen or more years ago--which means during the era of the first generation of large computers. They operated in batch mode, of course, and on rather restricted document collections; but they operated. And, as the years have passed, the catalogs or indexes of more and more document collections have been committed to computers.

The appeal of computers is obvious. There is, first of all, the speed and accuracy with which they can perform basic functions, such as filing in of new data, compiling statistics, transcribing data for human reading, and transmitting data for use by other machines. There is the ability of computers to perform complex logical searches, at least on pre-designated elements of the stored data. And, very important, there is now the ability of computers to serve numerous users simultaneously at diverse locations, by means of time-shared terminals, to obviate the need for the users to be in physical attendance at the catalog storage location.

Nevertheless, the use of computerized catalogs today is still highly restricted. It tends to be confined to applications where the document collection is relatively small, where the catalog information is very simple and limited, where there is an unusually high value attached to rapid or remote catalog service, where large computing capacity is already available for purposes unrelated to the library. This is because of the negative aspects of computers: the high cost of converting existing catalogs to machine-readable form; the high cost of computers; the unavailability of really large-scale rapid-access memory; the limited reasoning capacity of existing computer programs.

Because the negative aspects of catalog computerization have been particularly serious for the very large general-purpose library, of which the Yale University Library is a prominent example, there has long been a tendency for management in these libraries to regard catalog computerization as probably inevitable but clearly remote. Therefore, it could be dismissed from serious attention. That attitude can no longer be justified. Recent events have indicated that the time when conversion will be practical for large libraries may not be so remote after all--indeed may be only a few years away. Events contributing to this change have included: the steady growth of rapid memory capacity of computers; the falling cost of computing capacity; the improvement of equipment and of programs for remote-terminal time sharing; the establishment of the MARC system to make new catalog data available in machine readable form at low cost; the development of regional library groups which have the potential to make

existing catalog data available in machine readable form at low cost through cooperative effort; the development of standard machine formats which will make data interchange possible and economical.

So the decision on when to computerize the catalog of the very large libraries may soon become a matter of tactics, rather than strategy. At this point, the question of how to computerize the very large catalog is in need of urgent attention. The natural tendency, of course, would be to create a computerized catalog in the image of the existing manual card catalog, preserving all features of present-day catalog content and file organization. Tradition tends to be very strong among catalogers in large libraries. Yet tradition must be resisted, or at least questioned. Existing card catalogs are not necessarily the ultimate in human wisdom and ingenuity. Certainly some of the features in their design are attributable to the inherent limitations of cards and card drawers. There is no need to perpetuate the weaknesses of present catalogs in future catalogs. Before computerizing our catalogs, it would be very desirable for people in large libraries to take a hard look at what we would want from an ideal catalog, then see what sort of design in a computerized catalog would most closely approach that ideal. The key question is "What do we want from a library catalog?" One of our research projects at the Yale University Library is endeavoring to provide an answer to this question.

The approach we have taken is very direct. We are trying to learn what a future catalog should be by studying, quantitatively, what our library patrons are trying, successfully or otherwise, to get out of our present catalog. This study is supported, in part, by the Office of Education (1). The basic idea of a catalog use study is not at all new. There are quite a few such studies already reported in the literature, mostly master's thesis projects. Unfortunately, almost none of them inspire any confidence in the results because of gross deficiencies in experimental design, sample size, or both. Our own study was carefully designed to anticipate and obviate any foreseeable criticism. It is a two-year study which began in late 1967 and will be completed late this year.

Actually our study is much broader than I indicated in my introduction: It attempts to find out what our users want from a catalog, but it does not stop there. It also attempts to find out the extent to which our present card catalog satisfies the needs of the users. And, furthermore, it attempts to find out whether there are practical methods, manual or mechanized, to satisfy needs that are not now being met. Thus, even if we do not computerize our catalog for many years, the study should be useful in perfecting our traditional card catalog in the meanwhile.

Because the study is still in progress, I am unable to give any final results. The collection of data is more or less complete, but many of the projected analyses of the data have not yet been accomplished. Therefore, I will confine myself mainly to describing how the study has been carried out and stating what we should be able to learn from it. I will state some of our preliminary findings, but I must emphasize that all figures to be quoted here are based on incomplete data and are subject to possible revision in our final report.

The public catalog of the Yale University Library is located in the main

entry hall of the Sterling Memorial Library. It contains some 7 million cards, housed in some 7000 file drawers. It is a single-alphabet catalog. It contains full catalog card sets for the more than 3 million volumes housed in Sterling Memorial Library and only main-entry cards for the 2 million volumes housed in other libraries at Yale. Since the numerous school and departmental libraries have more complete catalogs for their respective collections, users of the main catalog are generally in search of books that are housed in the collection at Sterling Memorial Library. The stacks of Sterling Memorial Library are open to all Yale faculty and students, and to a rather large number of authorized outside users of the library. The catalog, as you can imagine, takes up a rather large area, and is the scene of constant activity throughout the hundred hours a week that the library is normally open.

A catalog search is basically a word-matching procedure. The searcher seeks to match some known clue, which is commonly a word or a phrase or a name, against the headings in the file; if he succeeds in finding a file item which matches his clue, he can expect to find some associated information in the file (e.g., a call number) which is the object of his search. In a nutshell, the aims of our study are to find out: 1) what clues the catalog users possess when they begin a catalog search; 2) how well our present catalog responds to (i.e., matches) the clues that the user brings; and 3) whether the responsiveness of the catalog might be improved through some change(s) in catalog design.

We are finding out what clues the users bring to their catalog searches through interviews with a representative sample of catalog users. The interviewees are approached at the instant that they reach for a catalog drawer to begin a search; they are asked a number of carefully worked out questions designed to elicit very precisely what the searcher is trying to accomplish through the catalog and what information he has brought to the search. We also collect background information about the searchers (but we do not ask for their names). The interviewers are all trained to follow a standard interview outline. At the beginning, the questions are very general and nondirective, to avoid leading of the subject. ("Could you please tell me what you were about to do here at the catalog when I interrupted you?") Only after the subject has had ample opportunity to say whatever he wants to, in his own way, do the questions become more direct and specific. Clues available to the searcher are recorded in full detail. If he carries them in the form of a printed bibliography or as handwritten notes, they are photocopied by the interviewer. If he carries them in his mind, they are transcribed by the interviewer, taking pains to determine and preserve the searcher's personal version of the spelling of author names and unusual words.

An average interview takes about ten minutes; but it may take as little as two minutes or more than fifteen minutes, depending on the nature of the searcher's problem and the amount of information which he brings to the search. When the interview is concluded, the subject is left alone to carry out his search, but is observed discreetly from a distance. The catalog drawer which he uses is noted. When he appears to have finished, he is approached again and asked if he was successful. If so, the interviewer notes the call number(s) of the item(s) which satisfied the search. Later on, we can examine the catalog cards for these call numbers, and we can examine the books themselves, to see how well the existing catalog matched, and how well it might have matched, the

clues which the user had when he began his search. This follow-up activity to examine the catalog cards and the books they represent is considerably less glamorous and exciting than face-to-face interviewing, but it is every bit as important to our study; and it actually takes more time and effort than the interviews.

The interview program, concluded only this month, was conducted over a full calendar year. We gathered data from some 2,000 interviews. The catalog users were cooperative beyond our wildest dreams. Fewer than 1 percent of the people approached refused to be interviewed--generally it was because they had to rush off to a class. Most interview subjects were delighted to be asked about their activities and eager to respond to all questions. Because of the accidents of random sampling, some people were interviewed two or three times during the year, and they still remained fully cooperative. To put it simply, the library users were very happy to learn that somebody actually cared about them.

At this point, I should explain how the interviewees were selected in order to provide a representative sample. Long before we began any interviewing, we had already begun collecting gross statistics on observed traffic in the catalog area and on various activities which occur in the catalog area. There happen to be five different entrances to our catalog area. By counting the number of people entering through each doorway at various times on different days, we constructed a preliminary projection of expected traffic by day of week and time of day. We then decided how large an interview sample we wanted (at least 1 percent). To get this, we worked out a precise interview schedule for each doorway in which the interview times and dates are in proportion to the expected traffic. Thus, each of our interviewers (2 full time, with a third available to help in emergencies) was assigned to be at a specific doorway at a specific hour and minute; and the first catalog user who entered through that doorway before a fixed interval elapsed was the person to be interviewed. Then the interviewer would go on to his or her next assignment, which would generally be at a different doorway. Assignments were spaced to allow reasonable time for completion of one interview before starting the watch for the next one. Sometimes no one would come through the doorway during the scheduled interval and so there was no interview; however, this is a random event which does not affect the value of the sampling technique.

What <u>can</u> affect the value of the sampling technique we used is seasonal variation in traffic pattern. Therefore, we continued the gross traffic counting program for more than a year in order to detect such variations. Differences between the observed pattern and the preliminary projection on which the interview scheduling was based will be compensated by applying appropriate weighting factors to the results of interviews conducted at different times and times of the year, so as to make the statistical results entirely representative of observed traffic.

By now you probably have a fairly clear idea of what we have been doing. Now we can discuss the ultimate question: What do we expect to get out of the study that can do anyone some good?

Let us start with our gross observations of traffic and other activities in

the catalog area. We can plot traffic by time of day, day of the week, and time of the academic year, and can thus produce a clear picture of expected volume and variation of catalog use. This can be of immediate value to the library administration--particularly in planning for the provision of reference assistance, and in scheduling of catalog maintenance--and it can be important in helping to determine the peak simultaneous access capacity which must be provided in any future computerized catalog facility. Of course, librarians already know quite a lot about traffic patterns from long years of experience, so we do not expect any earth-shaking revelations from this particular result of the study.

Other aspects of our observation of catalog traffic are more novel. We have collected much information on the amount of time which users spend at the catalog. What proportion of users spends one minute per use, two minutes, five minutes, fifteen minutes, etc? From this we can tell what kind of queuing to expect in the catalog area, not only with the present level of activity but with increased future activity as our user population grows. This should give us a sort of yardstick against which to measure the performance of contemplated computerized systems, to see whether they are worthy of serious consideration. We have collected extensive data on the number of catalog cards which users actually look at in the course of a catalog search, and on the number of references which they tend to copy from the catalog cards during a search. These data may or may not prove useful in furthering our understanding of the catalog user. We have collected data on precisely which catalog drawers were consulted by searchers at times when traffic was being observed. This should tell us whether all catalog drawers tend to be consulted equally or whether there are high-activity areas and low-activity areas in the catalog. This will have an important bearing on the level of queuing to be expected in a computerized catalog for any given memory access arrangement. All of these results will be based on very simple objective observations of the catalog area--merely counting people, and timing people, counting their hand motions in writing down references or flipping cards, noting and recording catalog drawer numbers. These measurements require no interviewing at all.

The interview data will yield a wealth of potentially useful results. For one thing, they will add some useful details to our picture of catalog traffic. Since we record the academic status of persons interviewed, we will be able to describe separate traffic patterns for students, faculty, staff, outsiders--and see whether they differ significantly. We will be able to do the same for newcomers to the University (students or faculty), as opposed to old-timers. We will be able to do the same for different departmental affiliations or areas of study.

Secondly, the interview data will yield quantitative insights into what it is that catalog users are seeking, and will tell us whether different categories of users tend to bring different types of problems to the catalog. Fairly early in the study, it was observed that the objectives of catalog searches tend to fall into four rather distinct categories. One category, the "document search," is where the user has a specific published work in mind and is using the catalog in order to locate a copy of that work. A second category, imperfectly called the "author search," is where the user knows of a source of publication--usually but not necessarily an author or corporate author--and wants to find out what works

are available from that source (e.g., what are some books by Thomas Mann?). A third category is the "subject search," where the user seeks to identify publications on a known abstract topic. The fourth category is the "bibliographic search," where the user has no intention of borrowing any book, but is only interested in finding the catalog card for a known publication so that he may get some specific information from the catalog card itself (e.g., to complete the bibliographic citation in a paper he is writing).

The document search is by far the most common. Analysis of a portion of our data suggests that about 75 percent of the uses of our catalog are for the purpose of locating a specific known publication (which, to our surprise, is almost always available in our collection). The other three use categories are more or less equally divided among the remaining 25 percent.

These results are preliminary, of course. Even if they were final, they would be suspect, however. There is a strong possibility or presumption that the actions of a library user are shaped by the nature of the catalog facility that is available to him. Do library users tend to accommodate themselves to what our catalog can do very well, such as locate known works? We are getting an answer to this from a very innocuous sounding but highly revealing question that we ask in our interviews. It reveals that a significant number of the document searches performed at the catalog are really subject searches in disguise. Presumably there would be a smaller proportion of overt document searches if our library catalogs were b tter suited for subject searching. We hope to get at the question of accommodation in yet another way, by looking for any difference in searching patterns between newcomers to the University and old-timers, or between newcomers at the beginning of the school year and later in the school year (when they have had a chance to adjust to reality).

A third, and also very important, type of result expected from our interview data will be the compilation and analysis of the search clues which catalog users possess at the start of their searches. By comparing the clues with the information available in the retrieved catalog cards and the documents they represent, we can assess the accuracy of the clues. For example, we can tell how often the catalog users start out with author names or titles that are inaccurate or misspelled, and we can analyze the frequency of different types of inaccuracies. This is fairly important for designing card catalogs, but it could be crucial for computerized catalogs. Computers make no concessions to misspelling unless designers take great pains to program around their punctilious and unyielding accuracy. The data collected from the interview program can be used to test the effectiveness of computer algorithms which are intended to produce matches despite inaccurate input from the searcher. We have already made quantitative evaluations of the effectiveness of two different data compression algorithms described in the literature by testing them on real data from our interview program.

Last, but by no means least, we will be able to use data from the interviews and from the retrieved catalog cards, and from the works corresponding to those catalog cards, to seek means to improve the quality and efficiency of cataloging rules and catalog structure. We will be able to see whether there are categories of data included on cards which are rarely wanted, or categories

which are frequently wanted but rarely included. We will be able to throw some light on the wisdom of dividing a catalog into sections segregated by date of publication or by other unconventional distinctions. We should learn whether machine-like subject indexing which makes use of the key words occurring in book titles, or prefaces, or chapter headings, or indexes, etc., would match actual user clues as well as our conventional subject indexing (based on authority lists) does now. Or whether it would be even better.

Of course, we are only studying one library at one university. Will our results be useful to people outside of Yale? We believe that they will be; but I would caution in advance against blind acceptance of any of our results as universally relevant. There are bound to be local differences among libraries and universities. To find out how significant these differences can be, it would be prudent to conduct studies similar to ours at a considerable number of large libraries of different kinds. I was very gratified to learn recently that a study of this type will soon be undertaken at the Library of Congress. But more studies are needed. I hope that they will not be long in coming. After all, the computers are nearly upon us. With all the effort that has been going into research and development work on how to computerize catalogs, it would be nice to have more guidance on how to do it right.

Reference

1. Ben-Ami Lipetz and Peter Stangl, "User clues in initiating searches in a large library catalog," pp. 137-139 in American Society for Information Science, Proceedings, 5, Annual meeting, October 20-24, 1968, Columbus, Ohio, Greenwood Press, New York, 1968.

APPENDIX C

P. Stangl, B. A. Lipetz, and K. F. Taylor
"Performance of Kilgour's truncation algorithm when applied to bibliographic retrieval from a library catalog," pp. 125-7 in American Society for Information Science, Proceedings, Vol. 6, Annual Meeting, October 1-4, 1969, San Francisco.

PERFORMANCE OF KILGOUR'S TRUNCATION ALGORITHM WHEN APPLIED TO BIBLIOGRAPHIC RETRIEVAL FROM A LIBRARY CATALOG*

Peter Stangl, Ben-Ami Lipetz, and Kathryn F. Taylor
Research Department, Yale University Library
New Haven, Connecticut 06520

Abstract

F. G. Kilgour's truncation algorithm for machine retrieval from large bibliographic files (1) was tested for performance in matching user-supplied, unedited search clues to bibliographic data contained in a library catalog. Kilgour had previously tested the algorithm to identify duplicate book orders in the in-process list of Yale University Library, and found recall to be about 90%. We have now tested the algorithm, by manual simulation, on data derived from 126 case studies of actual searches of the catalog at Yale University Library. The algorithm achieved 70% recall when compared to results of conventional manual searching. Precision was not determined.

Frederick G. Kilgour (1) has proposed an algorithm for machine retrieval of bibliographic entries from very large files, including library catalogs. The algorithm is designed to cope with misspellings and other discrepancies in the user's input when searching a file that contains entries of high editorial quality. The algorithm truncates and matches the user's version and the file's version of author-and-title data in a bibliographic entry. Kilgour reported on a test of his algorithm in which it was used to check for duplicate book orders in a 20,000-entry in-process (acquisition) list at Yale University Library. In this paper we report on a test of his algorithm as applied to a library catalog, rather than an in-process list.

The opportunity to test Kilgour's method when applied to retrieval from a library catalog was provided by the ready availability of data derived from a current study (2) of catalog use at Sterling Memorial Library (3.5 million books) at Yale University. This study collects, from a rigidly randomized sample of catalog users, precise information on the clues available to them at the moment of initiating a search. Search clues are recorded exactly as known to the catalog user, employing his own spelling--right or wrong. For each catalog user studied, the outcome of the search is ascertained; complete catalog information is recorded for documents identified as pertinent in successful searches.

In our test, search clues known to catalog users who seek specific documents, and catalog data corresponding to documents identified by these users, were truncated and matched, by manual simulation, according to Kilgour's algorithm. We were thus able to test its recall performance with real catalog searches. A test of the method's precision was not immediately feasible, because it would require comparison of input data with the entire catalog or a substantial portion of it. However, it is felt that the determination of recall performance should at least indicate whether the method shows sufficient promise in catalog searching to warrant evaluation of its precision in such an application.

Data used in our evaluation came from 126 searches in which the catalog user was successful in locating the specific document he was seeking. The two most successful versions of Kilgour's truncation algorithm were tested, those with formulae 3-3-1 and 5-5-1 (where the three figures stand for the number of initial characters to be retained from the author's last name, the title's first word, and the title's second word). Both user data and catalog data were truncated; where truncated versions matched, the entry was considered retrieved.

It should be noted that certain allowances which favored the algorithm were made in our test. Kilgour applied his method to only those entries in the file having a personal or corporate name main entry, thus excluding title main entries. Some title main entries were included in our sample of 126 catalog searches, and all but two were considered retrieved, since the user's clue corresponded perfectly to catalog data; thus any algorithm would have retrieved them. In two title main entries the user's clue did not match perfectly, so we eliminated them from our test, reducing the sample to 124. Further, in our test, all cases where a user had information on any name entry (not just the main entry) in the catalog, that information was considered as though it were a main entry. Thus a user's clue which matched only a joint author and title was still considered retrieved by us, although in Kilgour's test it could not have been, since his test was performed on a single-entry file. Finally, where the only difference was one of punctuation, or where there was a difference because translated or transliterated data were supplied by the user, full credit was given and the item was considered retrieved.

In his test on the 20,000-entry in-process list, Kilgour found that his algorithm produced a precision of 97.3%; that is, 97.3% of the "duplicate" references retrieved by the algorithm were indeed duplicates. (It should be noted, however,

*This work was supported in part by a grant from the U. S. Office of Education.

Table I

Results of Applying Kilgour's Method in Cases Where User's Clues and Catalog Data Did Not Match Completely

Imperfections in User's Clues	Documents Retrieved Method 3-3-1	Method 5-5-1	Documents Not Retrieved*
Neither author nor title			2
Author's last name, no title			9
Title, no author			3
Wrong author			1
Misspelled author	2		6 (8)
Wrong words in title	2	2	5
Misspelled words in title	3	1	1 (3)
Transposed words in title	1	1	
Incomplete title:			
a. First word correct	4	4	
b. First word incorrect			4
Entire subtitle, no title			1
Part of subtitle:			
a. First word correct			1
b. First word incorrect			3
Total documents	12	8	36 (40)

*Numbers in parentheses apply to items not retrieved by 5-5-1; all other numbers in this column apply to items not retrieved by either method.

that precision performance is in part a function of file size, and would be expected to drop off when applied to much larger files.) Kilgour tested recall by visual inspection of a sample of the in-process list, and estimated it to be between 85% and 93%. In other words, about 90% of all real duplicates were retrieved from the file by his algorithm.

Results of our test showed that of the 124 documents which were located successfully by manual search in the existing card catalog, 88 were retrieved by algorithm 3-3-1 and 84 by algorithm 5-5-1, amounting to recall rates of 70.9% and 66.1%, respectively. In 76 out of our sample of 124 searches, the user's author-and-title information matched corresponding catalog data character by character, while in 48 cases there were some discrepancies (misspellings, missing and wrong words, etc. (see Table I)). Algorithm 3-3-1 was able to "heal" discrepancies and retrieve the item in 12 cases; algorithm 5-5-1 in 8 cases. On the other hand, human beings were able to heal these discrepancies in a far greater number of instances, namely 48.

The viewpoint from which we determined a recall rate of 70% should be clearly understood. We are considering real document searches that happen to have been concluded successfully in an actual library with a manual catalog and we have determined the proportion of these searches which would be concluded successfully in a hypothetical, computerized library where the only means of searching the catalog would be by Kilgour's method. In a real library with a manual catalog, wanted documents can be located many ways, not merely through a knowledge of author and title (e.g., through subject entries, series entries, cross references). We do not disqualify any manual approach from consideration. We are comparing the real world with a specific potential alternative. Obviously, the use of Kilgour's method in combination with other computer programs could result in a recall rate higher than 70% by our method of calculation, and conceivably higher than 100% (because some document searches on manual catalogs that now end in failure might become successful using new search methods).

It is interesting to note that our results presented here very closely correspond to another similar simulation test we conducted (3) on Frederick H. Ruecking's method (4), where we found recall performance to be also about 70%, compared to his report of 90%. Ruecking matched unedited, user-supplied purchase requests against a MARC I tape according to an ingenious word-compression algorithm. Kilgour matched entries of a largely unedited in-process list against each other, according to an elegantly simple truncation algorithm.

We would caution readers against assuming that the same algorithm is likely to be equally effective in solving problems associated with acquisitions control and library catalogs. The differences between our results and those reported by Kilgour and Ruecking demonstrate that the situations with regard to catalog use and acquisitions control are very different, and that tests made in one situation cannot be regarded as very reliable to predict what will happen in the other. It appears that user clues vary significantly with different types of application; those brought to the catalog and those supplied on purchase requests have little in common. The name of the author and the title seem to be much more consistently (and correctly) supplied by users on purchase requests than in catalog searches.

On the basis of our tests, it is difficult to regard machine retrieval by means of word compression or truncation algorithms as a satisfactory substitute for conventional manual searching of library catalogs. However, improvement in the performance of machine techniques might be expected from modification of such algorithms or from the use of combinations of algorithms for retrieval. Further work on such approaches is highly desirable.

References

1. Kilgour, Frederick G., "Retrieval of Single Entries from a Computerized Library Catalog File," in American Society for Information Science, Proceedings, 5, Annual Meeting, October 20-24, 1968, Columbus, Ohio, pp. 133-136.

2. Lipetz, Ben-Ami, and Stangl, Peter, "User Clues in Initiating Searches in a Large Library Catalog," in American Society for Information Science, Proceedings, 5, Annual Meeting, October 20-24, 1968, Columbus, Ohio, pp. 137-139.

3. Lipetz, Ben-Ami: Stangl, Peter; Taylor, Kathryn. Submitted for publication.

4. Ruecking, Frederick H., Jr., "Bibliographic Retrieval from Bibliographic Input; the Hypothesis and Construction of a Test," Journal of Library Automation, 1, No. 4, pp. 227-238, December 1968.

APPENDIX D

B. A. Lipetz, P. Stangl, and K. F. Taylor
"Performance of Ruecking's word-compression method when applied to machine retrieval from a library catalog," Journal of Library Automation, Vol. 2, No. 4, 266-271 (1969 December).

PERFORMANCE OF RUECKING'S WORD-COMPRESSION METHOD WHEN APPLIED TO MACHINE RETRIEVAL FROM A LIBRARY CATALOG

Ben-Ami LIPETZ, Peter STANGL, and Kathryn F. TAYLOR:
Research Department, Yale University Library, New Haven, Connecticut

Reprinted from the *Journal of Library Automation*
Volume 2, Number 4: December, 1969
Printed in the U.S.A.

PERFORMANCE OF RUECKING'S WORD-COMPRESSION METHOD WHEN APPLIED TO MACHINE RETRIEVAL FROM A LIBRARY CATALOG

Ben-Ami LIPETZ, Peter STANGL, and Kathryn F. TAYLOR:
Research Department, Yale University Library, New Haven, Connecticut

F. H. Ruecking's word-compression algorithm for retrieval of bibliographic data from computer stores was tested for performance in matching user-supplied, unedited bibliographic data to the bibliographic data contained in a library catalog. The algorithm was tested by manual simulation, using data derived from 126 case studies of successful manual searches of the card catalog at Sterling Memorial Library, Yale University. The algorithm achieved 70% recall in comparison to conventional searching. Its acceptability as a substitute for conventional catalog searching methods is questioned unless recall performance can be improved, either by use of the algorithm alone or in combination with other algorithms.

Frederick H. Ruecking has published a report (1) of a method for improving bibliographic retrieval from computerized files when searching on unverified input data supplied by requestors. The method involves compression of author-and-title information before comparison. The rules for compression cause certain types of spelling errors and word discrepancies to be ignored by the computer. Ruecking reported 90.4% recall and 98.67% accuracy (precision) in a test of his method in which unverified book order requests were matched against a MARC I data base that contained 1392 of the references searched. This paper reports on a small-scale manual simulation test undertaken to assess the value of the method when applied to bibliographic retrieval from a library catalog.

The opportunity to test Ruecking's method when applied to retrieval from a library catalog was provided by the ready availability of data derived from a current study (2) of catalog use at Sterling Memorial Library (3.5 million books) at Yale University. This study collects, from a rigidly randomized sample of catalog users, precise information on the clues available to them at the moment of initiating a search. Search clues are recorded exactly as known to the catalog user, employing his own spelling—right or wrong. For each catalog user studied, the outcome of the search is ascertained; complete catalog information is recorded for documents identified as pertinent in successful searches. Search clues known to catalog users who seek specific documents correspond to the "unverified input data" which Ruecking's method would match against catalog holdings. Catalog information on those documents identified as pertinent corresponds to the portion of the data base that Ruecking's program seeks to match. It was possible, therefore, to apply Ruecking's method by manual simulation, and to test its recall performance in real catalog searches. A test of its precision was not immediately feasible because such a test would require comparison of input data with the entire catalog (or a substantial portion of it). However, the determination of recall performance would at least indicate whether the method shows sufficient promise in catalog searching to warrant evaluation of its precision.

An aside on precision is in order, however. It should be noted that precision of retrieval with a given method tends to vary inversely with the size of the file being searched. Although Ruecking did not specify the number of records included in his MARC I data base, it could not have exceeded 48,000. Had he run his test on a data base, ten, or fifty, or one hundred times larger, the measured precision would certainly have been much lower than the figure reported. Any librarian who is contemplating the adoption of a retrieval technique which has been tested on a data base similar to, but smaller than, his own should realize that precision performance must inevitably drop as the data base is increased. The degree of lowered precision to be expected may be predicted theoretically or estimated from tests on files of several different sizes.

The data used in the evaluation of recall performance reported in this paper came from 126 searches in which the catalog users had been successful in locating the specific documents that they were seeking. The compression coding method described by Ruecking was applied in each instance to the author-title search clues supplied by the catalog user and to the author-title information available on the catalog card. Threshold values were computed for the catalog card data, and retrieval values were computed for the user data. When the retrieval value was at least as large as the threshold value, the document was considered "retrieved."

Ruecking's method was designed for use with English-language titles only. Of the 126 catalog searches in the study sample, 20 involved foreign-

language titles. Recall was determined on both the full sample and the English-language subset of 106 searches. Surprisingly, there is not a great improvement in performance when foreign-language references are excluded.

It should be noted that several difficulties were encountered in applying Ruecking's method because of ambiguities in the rules stated in his paper. In fact, in his Figure 2 (page 236), of the seventeen illustrations of compression-coded data retrieved by his program, at least eight appear to contain departures from the compression-coding rules as stated in the paper. His Table 5 (page 235) is scantily described: "Individual Code Test" and "Full-Code Test" are not defined; neither are column headings. And, contrary to the text (page 234), values in columns five through seven are obtained by adding two to the calculated thresholds in only the top half of Table 5; in the bottom half, no such regular correlation exists. In all cases of ambiguity, the alternative was selected that would tend to increase probability of retrieval. For example, Ruecking states (page 234) that the search program provided for matching of titles on the basis of rearrangement of title words, and that the threshold value required for retrieval is raised at the same time. Raising this value decreases the probability of retrieval, but it is not clear by how much the value is to be raised. For purposes of the test, the threshold value was not raised at all in cases where title words were out of correct sequence, thus retaining maximum probability of retrieval based on the number of matched words alone, regardless of their sequence.

Results of the test showed that, of the 126 documents in the full sample which were located successfully by manual search in the existing card catalog, only 88 were retrieved by the compression-code method—a recall rate of 70%. Considering only the 106 English-language references, 77 were retrieved by the compression-code method—a recall rate of 73%.

The premise for the preceding calculation of recall rate should be clearly understood. The test considered real document searches that were concluded successfully in an actual library using a manual catalog; recall is defined here as the proportion of such searches that would be concluded successfully in a hypothetical, computerized library where the only means of searching the catalog would be by Ruecking's method. In a real library with a manual catalog, wanted documents can be located in many ways, not merely through a knowledge of author and title (e.g., through subject entries, series entries, cross references). The test did not disqualify any manual approaches from consideration; it compared the real world with a specific potential alternative. Obviously, the use of Ruecking's method in combination with other computer programs could result in a recall rate higher than 70% or 73% by the method of calculation employed, and conceivably higher than 100% (because some document searches of manual catalogs that now end in failure might become successful using new search methods).

Table 1 provides detailed information on the discrepancies between user data and catalog data in the test. With respect to the full sample (126 documents), there were 49 documents for which mismatches of data were observed. Of these, the compression-code method was able to "heal" mismatches in 11 instances to cause retrieval; on the other hand, manual searches had achieved retrieval in all 49 instances. With respect to the English-language sample (106 documents), there were 37 documents for which mismatches of data were observed. Of these, the compression-code method was able to "heal" mismatches in 8 instances to cause retrieval; on the other hand, manual searches had achieved retrieval in all 37 instances.

Contrary to expectations, the compression-code method performed somewhat worse, or at least no better, in "healing" actual mismatches in English references (8 out of 37) than it did with foreign-language references (3 out of 12). The higher overall recall percentage with the English-

Table 1. Results of Applying Ruecking's Method in Cases where User Clues and Catalog Data Did not Match Completely

	Full Sample (126 documents)		*English Subset (106 documents)*	
Type of Mismatch in User Data	*Retrieved*	*Not Retrieved*	*Retrieved*	*Not Retrieved*
Had neither author nor title		2		1
Had author's last name, no title		9		5
Had title, no author	1	2	1	2
Had wrong author		1		1
Had misspelled author	4	4	2	1
Had wrong words in title	1	9*	1	6
Had misspelled words in title	2	2	1	2
Had words transposed in title	2		2	
Had incomplete title:				
a. First word correct	2	5**	2	5**
b. First word incorrect		6		5
Had entire subtitle, no title		1		1
Had part of subtitle				
a. First word correct		1		1
b. First word incorrect		2		2
Total documents***	11	38	8	29

*1 case of correct word stems not matched because of wrong endings.

**2 cases of long or composite titles with maximum threshold values contained in input words but not among the first four significant words.

*** Figures shown are lower than totals of figures in columns because some documents had two or more types of mismatch.

language subset is attributable entirely to the fact that users had complete and correct data more frequently for English references (69 out of 106) than they did for foreign-language references (8 out of 20). Thus, regardless of original intent, the method words equally well (or equally poorly, depending on one's viewpoint) on foreign-language and English references. If foreign-language references had been systematically ignored in applying the test to catalog searches, some 16% (20 out of 126) of the searches would have been excluded, with no real gain in performance.

The block of interviews from which the searches used in this test were drawn included 10 unsuccessful document searches in addition to the 126 successful searches. One could speculate on whether the compression-code method would have been able to "heal" these failures, resulting in a higher performance rating. The indications are, however, that the chances of such healing are close to zero. In a majority of these unsuccessful searches, the available data were incomplete or were not of the type that the method is intended to utilize. In the few remaining cases, it is very likely that the searches were unsuccessful simply because the desired documents were not in the library collection.

Recall performance as measured by the test could have been improved by modifying Ruecking's rules to some extent. For example, five more titles would have been retrieved had the assigned retrieval value been increased by two units in cases where the first title word matched correctly; this would have increased overall recall performance from 70% to 74%. A further increase to 76% would have resulted from matching the user's version of the title with the catalog's subtitle, or with portions of titles which follow a punctuation mark (in addition to matching with the actual title in the catalog).

Extension of the compression code to include publisher and date as well as author and title would do little or nothing to improve the performance of this method. The test data, although admittedly a small sample, indicate that users who do not have accurate author and title information when they begin a search very rarely have accurate information on any other descriptive data element.

It is, of course, a matter for individual judgment as to whether the performance of the compression-code method, as indicated by the test reported here, is sufficiently good to make it attractive for use in some computerized alternative to the manual library catalog. In the authors' opinion, Ruecking's method does not in itself supply an adequate solution to the problem of searching a computerized catalog. However, further investigation seems warranted along two lines. First, the method might be modified to give better performance in this application. Second, it might be used in combination with some other computer methods to give searching performance approaching that which is attained today by the manual searching of card catalogs.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by a grant from the U.S. Office of Education, OEG-7-071140-4427.

REFERENCES

1. Ruecking, Frederick H., Jr.: "Bibliographic Retrieval from Bibliographic Input; The Hypothesis and Construction of a Test," *Journal of Library Automation, 1* (December 1968), 227-38.
2. Lipetz, Ben-Ami; Stangl, Peter: "User Clues in Initiating Searches in a Large Library Catalog," in *American Society for Information Science, Proceedings, 5.* Annual Meeting, October 20-24, 1968, Columbus, Ohio, p. 137-139.

APPENDIX E

FORMS USED FOR INTERVIEWING AND FOLLOW UP

INTERVIEW SCHEDULE Number: _____

Interviewer _____ Date _____ Time _____ Door _____ Time began _____ Door done_____

Pardon me, Sir, (Madam), we are conducting a study of the use of the card catalog, and I would like to ask you a few questions if I may.

1.1 Have we interviewed you here at the catalog before?

[ ] yes [ ] no

If yes, number of times _____

2.1 Could you please tell me precisely what you were about to do here at the catalog the moment I interrupted you?

2.2 Do you have anything written down, anything to guide you in your current search?

[ ] yes [ ] no

If yes, copy obtained?

[ ] yes [ ] no

2.3 Do you know anything about this work in addition to what's written down?

2.4 Suppose that I would offer to do this search for you; how would you instruct me to ensure that I would do it right?

A. [ ] Please show me just what you are doing, as though you were showing a new assistant whom you are training to do this.

B. [ ] We need specific data on what people are looking for in the catalog, so we can compare this with what the catalog contains.

C. [ ] Are you looking for a specific publication you know about?

D. [ ] Are you looking for information on a topic, rather than a specific publication?

E. [ ] Are you just trying to find out something about a specific publication?

[ ] Document search
[ ] Subject search
[ ] "Author" search
[ ] Bibliographic search

Document and Bibliographic search Followup

Note: for Bibliographic search skip questions 2.3 and 2.4.

When finished with catalog drawer.

1. Have you found the item?
[ ] yes [ ] no

2.1 If yes, call no.:

2.21 Under what did you find it?

2.22 Are you satisfied that this is the item you were looking for?
[ ] yes [ ] no
Comment:

2.3 Did you pick out something in addition to what you were looking for (that is, something in the same connection)?
[ ] yes [ ] no

2.4 If yes, call no.:

How did you come across it?

3.1 If no, what did you look under?

3.2 What are you going to do next?

Time ended: _____: _____ a.m. p.m.

BACKGROUND

1. In what connection do you need this work?

Student	Non-student
[ ] course work	[ ] lecture preparation
[ ] paper	[ ] publication
[ ] undergraduate thesis	[ ] research project
[ ] research project	[ ] looking for ideas
[ ] orals	[ ] professional reading
[ ] dissertation	[ ] interest
[ ] interest	[ ] other
[ ] exam	
[ ] other	

2. How or where did you find out about this work?

[ ] someone mentioned it	[ ] remembered it
[ ] someone wrote it down	[ ] hand-copied it
[ ] was in footnote or bibliography in book	[ ] machine-copied it
	[ ] have original source
[ ] found in reference book	
[ ] course list	
[ ] saw work itself	
[ ] own list	
[ ] owns work itself	
[ ] has always known	
[ ] other_____	

Statistical data

1.1 Affiliation:

[ ] Undergraduate

11 [ ] Freshman	4 [ ] Faculty
12 [ ] Sophomore	5 [ ] Staff
23 [ ] Junior	6 [ ] Non-Yale student
24 [ ] Senior	7 [ ] Non-Yale faculty
3 [ ] Graduate student _____ year	8 [ ] Faculty or student wife
	9 [ ] Other _____

1.2 How many years have you been using this library? _____ years

2.1 If not a freshman or sophomore: What department are you with? _____

or

What is your major? _____

2.2 If faculty or staff: What is your title or the name of your position? _____

3. Sex: 1 [ ] Male 2 [ ] Female

4. How often do you use the Catalog at Sterling Library:

[ ] several times a day
[ ] daily
[ ] 3-4 times a week (every other day)
[ ] 2 times a week (every 3-4 days)
[ ] weekly
[ ] 2-3 times a month
[ ] monthly
[ ] less frequently

5.1 Which other Yale libraries do you use regularly, if any? [ ] none

5.2 Do you use these just to study in, or do you use their collections?

Libraries:	Study	Use collections
	[ ]	[ ]
	[ ]	[ ]
	[ ]	[ ]

6. Interview ended: _____: _____ a.m. p.m.

DOCUMENT SEARCH

[ ]

2.1 Please describe the item to me as fully as possible.
A. [ ] If you had something written down, what would it say?
B. [ ] If I offered to get this item for you, what more would you tell me to enable me to find it?

2.2 Is that a book or a periodical?
[ ] book [ ] periodical

If periodical, is reference to
[ ] article [ ] title only

2.21 If book, is reference to the whole work?
[ ] whole work [ ] section

2.22 If whole work, do you need the work as a whole, or do you just need to use some specific part of it?
[ ] whole work [ ] section
Which specific part?

2.3 Have you had any previous contact with this publication?
[ ] yes [ ] no
If serial, do you read this journal regularly, or just an occasional article?
[ ] regularly [ ] article

2.31 If yes, was that the Library's copy?
[ ] yes [ ] no

2.4 If yes, what do you remember about the physical appearance of the publication?

2.5 Under what are you going to look in the catalog?

2.6 How many items would you like to obtain on this occasion? _____

1. Post from P.1 and "notes"

P.1	Notes	
		1.1 If monograph:
[ ]	[ ]	author or editor
[ ]	[ ]	title
[ ]	[ ]	subtitle
[ ]	[ ]	edition and date of publication
[ ]	[ ]	translation from
[ ]	[ ]	publisher
[ ]	[ ]	place of publication
[ ]	[ ]	series
[ ]	[ ]	sponsoring organization
[ ]	[ ]	size, paging, color, binding
[ ]	[ ]	illustrations
[ ]	[ ]	bibliography
[ ]	[ ]	anything else
		1.2 If serial:
[ ]	[ ]	title of journal
[ ]	[ ]	subtitle of journal
[ ]	[ ]	editor
[ ]	[ ]	sponsoring organization
[ ]	[ ]	publisher
[ ]	[ ]	place
[ ]	[ ]	starting date and frequency
[ ]	[ ]	change in title
[ ]	[ ]	size and appearance
[ ]	[ ]	author of article
[ ]	[ ]	title of article
[ ]	[ ]	volume, issue number, date
[ ]	[ ]	anything else

3. For any category not covered, ask:
"Do you know the _____?"

Interview number: _____

SUBJECT SEARCH

Note: If it is apparent early in the interview that final choice of items will be made outside the catalog, skip questions 2.6 and 2.7.

[ ] 1.1 What is the subject on which you are searching material?

1.2 How would you describe the topic, if I were to offer to do the search for you?

1.3 Do you think you would want to tell me more, so that I would not pick out things you don't want?

1.4 Do you think you would want to tell more, so that I would not overlook anything?

2.1 How do you plan to find material on your topic?

2.2 O.K., under what are you going to look in the catalog?

2.3 Are you planning to look under anything else in the catalog?

2.4 And if that doesn't produce enough, are you then going to look under anything else?

2.5 For each heading given, ask "You mean, just like that?", or equivlent, to get an accurate form of the heading. Give numbers to headings elicited.

2.6 Supposing again, that I would do this search for you. When I look under these headings, is there anything I should look out for, so I wouldn't pick out things you don't really want?

2.7 When you look under these headings, will any of the following influence your choice?

[ ] author or editor
[ ] title (as description of topic)
[ ] date of publication
[ ] language
[ ] author or sponsoring organization
[ ] contents note
[ ] call number
[ ] size and paging
[ ] illustrations
[ ] bibliography
[ ] anything else

3. Approximately how many items do you wish to obtain?

4.1 When you are finished with the catalog here, do you think you will know exactly which publications you want to obtain, or will you need to look or ask anywhere else befor you make your final choice?

[ ] will know
[ ] look elsewhere Where?

4.2 Why?

[ ] books not there Other:
[ ] can't tell from catalog

4.3 What are you going to base your final choice of publications on?

4.4 (When you look in the stacks) Will any of the following influence your choice?

[ ] author or editor
[ ] title (as description of topic)
[ ] date of publication
[ ] place of publication
[ ] language
[ ] author or sponsoring organization
[ ] table of contents
[ ] index
[ ] preface
[ ] illustrations
[ ] bibliography
[ ] size and appearance
[ ] anything else not covered?

Request followup.

Interview number: _____

AUTHOR SEARCH

[ ] 1.1 Please describe your current search to me as accurately as possible.

1.2 Supposing that I were to offer to do the search for you, would you tell me anything more, so that I wouldn't pick out anything you don't want?

1.3 Would you tell me anything more, so that I would not overlook anything?

2.1 How do you plan to find the material you want?

2.2 O.K., under what are you going to look in the catalog?

2.3 (If it applies) Are you planning to look under anything else?

2.4 When you look in the catalog, would any of the following influence your choice? (ask all that applies)

[ ] author or editor
[ ] title (as description of topic)
[ ] subject headings
[ ] edition and date of publication
[ ] place of publication
[ ] language
[ ] publisher
[ ] series
[ ] sponsoring organization
[ ] size and paging
[ ] illustrations
[ ] bibliography
[ ] anything else

3. Approximately how many items do you wish to obtain? _____

4.1 When you are finished with the catalog here, do you think you will know exactly which publications you want to obtain, or will you need to look or ask anywhere else before you make your final choice?

[ ] will know
[ ] look elsewhere Where?

4.2 Why?

[ ] books not there
[ ] can't tell from catalog other:

4.3 What are you going to base your final choice of publications on?

4.4 (When looking in stacks) Will any of the following influence your choice?

[ ] author or editor
[ ] title (as description of topic)
[ ] edition and date of publication
[ ] place of publication
[ ] publisher
[ ] language
[ ] series
[ ] sponsoring organization
[ ] size
[ ] illustrations
[ ] bibliography
[ ] Table of contents
[ ] index
[ ] preface

Subject and Author search Followup

If it was indicated that choice was going to be made at catalog, approach user as he is leaving catalog area.

1. Did you find your material?

   [ ] yes [ ] no

2.1 If yes, call numbers:

2.2 Under what did you find these?

2.21 (If not the intended headings) What made you look there?

2.3 Are you satisfied that these are the works you need, will you take them?

2.4 Is this enough, are you now finished?

If 2.3 and 2.4 indicate that search is not over, request followup.

3.1 If no, did you find the headings you were looking for?

[ ] Found nothing
[ ] Found: [ ] Not found:

3.2 (If headings found are not the intended headings) What made you look there?

3.3 What was wrong with the material under those headings? How did you decide they were not appropriate?

3.4 What are you going to do next?

If indicated, request followup.

On return from stacks.

1.1 How far did you go at the catalog?

Get call numbers or portions obtained at catalog:

1.2 Under what did you find these?

2.1 What are your final choices?

Call numbers:

2.2 How did you choose these?

Interview number: _____

BIBLIOGRAPHIC SEARCH

[ ]

2.1 Please describe the item to me a s fully as possible.

A. [ ] If you had something written down, what would it say?
B. [ ] If I offered to find this item for you, what more could you tell me to enable me to find it?

2.2 Is that a book or a periodical?
[ ] book [ ] periodical

2.3 Exactly what do you need to find out about this publication?
[ ] verification Comment, details:
[ ] specific fact
[ ] other

2.4 Have you had any previous contact with this publication?
[ ] yes [ ] no

2.41 If yes, was that the Library's copy?
[ ] yes [ ] no

2.5 If yes, what what do you remember about the physical appearance of the publication?

2.6 Under what are you going to look in the catalog?

2.7 How many items are you looking up on this accasion? _____

1. Post from p. 1 and Notes		3. For any category not covered, ask: "Do you know the _____?"
p.1	notes	
1.1 If monograph:		
[ ]	[ ] author or editor	
[ ]	[ ] title	
[ ]	[ ] subtitle	
[ ]	[ ] edition and date of publ.	
[ ]	[ ] Translation from, lanfg.	
[ ]	[ ] publisher	
[ ]	[ ] place of publ.	
[ ]	[ ] series	
[ ]	[ ] sponsoring org.	
[ ]	[ ] size, paging, color, binding	
[ ]	[ ] illustrations	
[ ]	[ ] bibliography	
[ ]	[ ] anything else	
1.2 If serial:		
[ ]	[ ] title of journal	
[ ]	[ ] subtitle of journal	
[ ]	[ ] editor	
[ ]	[ ] sponsoring org.	
[ ]	[ ] publisher	
[ ]	[ ] place of publ.	
[ ]	[ ] starting date and frequency	
[ ]	[ ] change in title	
[ ]	[ ] size and appearance	
[ ]	[ ] author of article	
[ ]	[ ] title of article	
[ ]	[ ] vol., issue no., and date	
[ ]	[ ] anything else	

Please complete this form for each item chosen at the catalog or by browsing through the stacks as part of the search covered by the interview. Use the examples below as a guide. Please do this as you go along, because you are not likely to remember details later.

When finished and back from the stacks, please hand this form to one of the interviewers, or fold it in half and drop at the Circulation Desk, one of the Reference Librarians or in Campus Mail.

By promptly completing the form you are greatly contributing to the success of our study whose goal is the eventual improvement of the catalog. We very much appreciate your cooperation.

Research Department, Sterling Memorial Library

SPECIFIC CALL NUMBERS CHOSEN AT CATALOG

CALL #:	FOUND ON SHELF: (circle one)	IF NOT FOUND: (check "claim" if having book searched or recalled)	IF FOUND BOOK ON SHELF: TOOK BOOK (please check)	DID NOT TAKE BOOK: (NOT USEFUL AS REVEALED BY:) TITLE PAGE	PREFACE	TABLE OF CONTENTS	INTRO-DUCTION	OTHER REASON (SPECIFY)
XX10 967x	(yes) no	[ ] will claim [ ] will choose other book [ ] forget it	✓					
XX10 901y	(yes) no	[ ] will claim [ ] will choose other book [ ] forget it		Title too general; also too old		Super-ficial		
XX10 967Z	(yes) no	[ ] will claim [ ] will choose other book [ ] forget it			Too technical			Much too lengthy

BOOKS CHOSEN BY BROWSING IN STACKS

CALL #:	CHOSEN ON THE BASIS OF INFORMATION ON: TITLE PAGE	PREFACE	TABLE OF CONTENTS	INTRO-DUCTION	OTHER REASONS (SPECIFY)
XX10 966X	Author familiar		Chapter IV		Concise
XX10. 966Y		Original work			Good index
XX10 966Z	Recent			State-of-the-art report	Recommended reading lists

SPECIFIC CALL NUMBERS CHOSEN AT CATALOG

CALL #:	FOUND ON SHELF: (circle one)	IF NOT FOUND: (check "claim" if having book searched or recalled)	IF FOUND BOOK ON SHELF: TOOK BOOK (please check)	DID NOT TAKE BOOK: (NOT USEFUL AS REVEALED BY:) TITLE PAGE	PREFACE	TABLE OF CONTENTS	INTRO-DUCTION	OTHER REASON (SPECIFY)
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						
	yes no	[ ] will claim [ ] will choose other book [ ] forget it						

STACK AREAS CHOSEN AT CATALOG

AREA:	FOUND UNDER: AUTHOR	TITLE	SERIES	SUBJECT (SPECIFY):

BOOKS CHOSEN BY BROWSING IN STACKS

CALL #:	CHOSEN ON THE BASIS OF INFORMATION ON: TITLE PAGE	PREFACE	TABLE OF CONTENTS	INTRO-DUCTION	OTHER REASONS (SPECIFY)

APPENDIX F

FACTORS COMPARED IN COVARIANCE ANALYSIS

Factors Compared in Covariance Analysis

1. Day of the week
2. Time of day (half hours)
3. Season of the academic calendar
4. Entryway to catalog area
5. Reason for an interview not being done as scheduled
6. Previous interviews undergone by person interviewed
7. Interviewer
8. Duration of interview (minutes)
9. Promptness of follow-up on search results (immediately after catalog search or deferred pending further user activity.)
10. Completeness of follow-up
11. Issuance and return of questionnaires for follow-up
12. Academic status of person interviewed
13. Departmental affiliation or major subject of person interviewed
14. Years of experience with Yale University libraries
15. Sex
16. Frequency of past use of the catalog
17. Number of other Yale libraries used by person interviewed
18. Types of other Yale libraries used
19. Number of items desired in this use of the catalog
20. Type of catalog search (document, subject, author, bibliographic)
21. Underlying intent of document search (document, subject)
22. Connection in which material is needed (course work, paper, thesis, project, personal, etc.)
23. Source of reference to a desired document

24. Type of reference to a desired document (remembered, hand copied, duplicated)

25. Type of document desired (monograph, periodical)

26. Nature of previous contact with document

27. Intended search approach

28. Language of desired document

29. Is a translation involved?

30. Personal author clue (availability and accuracy)

31. Corporate author clue (availability and accuracy)

32. Title clue for which catalog contains an entry (availability and accuracy)

33. Title clue for which catalog contains no entry (availability and accuracy)

34. Added entry clues (availability and accuracy)

35. Date clues (type specificity and whether "known" or guessed)

36. Accuracy of date clues

37. Actual document dates (in ranges of years)

38. Other search clues (availability and accuracy)

39. Result of search (nothing found, desired material only, desired material plus additional material, additional material only)

40. Was intended search entry successful? (yes, no)

41. Type of heading under which desired material was located if different from intended approach.

42. Number of entries looked up

43. Number of useful call numbers found in addition to desired material

44. User intent following unsuccessful search (abandon, continue within Yale library system, continue elsewhere)

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

(TOP)

ERIC REPORT RESUME

ERIC ACCESSION NO.	CLEARINGHOUSE ACCESSION NUMBER	RESUME DATE	P.A.	T.A.	IS DOCUMENT COPYRIGHTED?	ERIC REPRODUCTION RELEASE?
	001	06 — 70			YES ☐ NO ☒	YES ☒ NO ☐

100 TITLE
User Requirements in Identifying Desired Works in a Large Library. Final Report.

200 PERSONAL AUTHOR(S)
Lipetz, Ben-Ami

300 INSTITUTION (SOURCE)
Yale University Library, New Haven, Connecticut

SOURCE CODE

310 REPORT/SERIES NO.

320 OTHER SOURCE

SOURCE CODE

330 OTHER REPORT NO.

340 OTHER SOURCE

SOURCE CODE

350 OTHER REPORT NO.

400 PUB'L. DATE Jun — —1970 CONTRACT/GRANT NUMBER OEG-1-7-071140-4427

500 PAGINATION, ETC
120 p.

600 RETRIEVAL TERMS
academic libraries; access; cataloging; catalogs; computerized cataloging; computer retrieval algorithms; information retrieval; library use; research libraries; search clues; traffic studies; user studies.

607 IDENTIFIERS
Sterling Memorial Library, Yale University

800 ABSTRACT

Utilization of the card catalog in the main library (Sterling Memorial Library) of Yale University was studied over a period of more than a year. Traffic flow in the catalog was observed, and was used as the basis for scheduling interviews with a representative sample of catalog users at the moment of catalog use. More than 2000 interviews were completed. Data were collected on user objectives and starting clues. Follow-up studies were done on the matches among user clues, catalog card data, and information available in the front matter of cataloged documents. Reasons for search failures were determined. In terms of immediate intent, 73 percent of searches are document ("known item") searches and 16 percent are subject searches; in terms of underlying interest, 56 percent are document searches and 33 percent are subject searches. Remaining searches are 6 percent author searches (to find out what is on hand from a known author or institution) and 5 percent bibliographic searches (to complete or verify a reference on the basis of catalog-card data). The importance of secondary search clues in achieving retrieval despite incomplete or inaccurate primary clues is discussed.

OE 6000 (Rev. 9-66)

INSTRUCTIONS FOR COMPLETING ERIC REPORT RESUME

The resume is used to identify summary data and information about each document acquired, processed, and stored within the ERIC system. In addition to serving as a permanent record of the document in the collection, the resume is also a means of dissemination. All fields of the form must be completed in the allotted spaces, but inapplicable fields should be left blank. The following instructions are keyed to the line numbers appearing in the left margin of the form:

TOP LINE. ERIC Accession No. Leave blank. A permanent ED number will be assigned to each resume and its corresponding document as they are processed into the ERIC system.

LINE 001. Clearinghouse Accession No. For use only by ERIC Clearinghouses. Enter the alpha code and 6-digit document number.

Resume Date. In numeric form, enter month, day, and year that resume is completed. *(Example: 07 14 66)*

P.A. Leave blank.

T.A. Leave blank.

Copyright. Check appropriate block to denote presence of copyrighted material within the document.

ERIC Reproduction Release. Check appropriate block to indicate that ERIC has permission to reproduce the document and its resume form.

LINES 100–103. Title. Enter the complete document title, including subtitles if they add significant information. Where applicable, also enter volume number or part number, and the type of document *(Final Report, Interim Report, Thesis, etc.)*.

LINE 200. Personal Author(s). Enter personal author(s), last name first. *(Example: Doe, John J.)* If two authors are given, enter both. *(Example: Doe, John J.; Smith, Ted.)* If there are three or more authors, list only one followed by "and others."

LINE 300. Institution *(Source)*. Enter the name of the organization which originated the report. Include the address *(city and State)* and the subordinate unit of the organization. *Example: Harvard Univ., Cambridge, Mass., School of Education.*

Source Code. Leave blank.

LINE 310. Report/Series No. Enter any unique number assigned to the document by the institutional source. *(Example: SC-1234)*

LINE 320. Other Source. Use only when a second source is associated with the document. Follow instructions for Line 300 above.

Source Code. Leave blank.

LINE 330. Other Report No. Enter document number assigned by the second source.

LINE 340. Other Source. Use only when a third source is associated with the document. Follow instructions for Line 300 above.

Source Code. Leave blank.

LINE 350. Other Report No. Enter document number assigned by the third source.

LINE 400. Publication Date. Enter the day, month, and year of the document. *(Example: 12 Jun 66)*

Contract/Grant Number. Applicable only for documents generated from research sponsored by the U.S. Office of Education. Enter appropriate contract or grant number and its prefix. *(Example: OEC 1-6-061234 (0123))*

LINES 500–501. Pagination, etc. Enter the total number of pages of the document, including illustrations and appendixes. *(Example: 115p.)* USE THIS SPACE FOR ADDITIONAL INFORMATION PERTINENT TO THE DOCUMENT, such as publisher, journal citation, and other contract numbers.

LINES 600–606. Retrieval Terms. Enter the important subject terms *(descriptors)* which, taken as a group, adequately describe the contents of the document.

LINE 607. Identifiers. Enter any additional important terms, more specific than descriptors, such as trade names, equipment model names and numbers, organization and project names, discussed in the document.

LINES 800–822. Abstract. Enter an informative abstract of the document. Its style and content must be suitable for public announcement and dissemination.

U.S. GOVERNMENT PRINTING OFFICE: 1966 O—231-551