THOMAS T. HEWETT
Associate Professor of Psychology
Drexel University

CHARLES T. MEADOW
Professor of Information Science
Drexel University

A Study of the Measurement of User Performance

This paper reports on an attempt to measure the performance of users of interactive information retrieval systems.[1] The subjects studied were end users of the information who were doing their own interactive searches. The measures consist of a set of computerized diagnostic procedures applied to the sequences of commands used in querying the database. These diagnostics trigger various kinds of messages to the user. Presumably, the frequency with which a diagnostic is triggered is an index of the difficulties which the user may be having in doing a search. Although the utility to the user of the information retrieved is assumed to be the best overall measure of search outcome, it is the manner of using the system, not the search outcome, which is the focus of this report.

The studies described here were part of a larger project called Individualized Instruction for Data Access (IIDA). The goal of this project has been to provide a method for allowing direct user access to bibliographic searching. Thus, the attempt has been to develop a set of computer software packages which can provide on-line assistance to occasional users. This collection of programs is also intended to provide instruction, if needed, in the commands used in searching and search strategy. When originally conceived, the expected utility of IIDA lay in the area of what might be referred to as "problem-solving searches." These are searches where the end user of the information does not know exactly what the characteristics are of the desired set of references until they have actually been found. Consequently, it is very difficult for the user to describe the problem to an intermediary. There is no reason, however, why the IIDA user could not and should not make use of the system for all kinds of searches, if desired.

One intended IIDA user is a working scientist or engineer who may need access to the database only a few times a year, and consequently is not interested in training oriented toward those who become professional intermediaries. This person is assumed to be comfortable using computers, but not necessarily trained in their use. In addition, it is assumed that this user is a serious, well-intentioned searcher who is trying to use the system to solve a problem. The IIDA software and diagnostic procedures were created to help remove the barriers to access for these users. Minor modifications of the system, however, could make it available to a much wider audience.

DESCRIPTION OF IIDA

Physical Configuration

A schematic of the current physical arrangement of the IIDA system appears in figure 1. While it is now possible to package the IIDA software in a dedicated minicomputer, the experimental version of the IIDA software resided in a general-purpose, time-shared computer at the Massachusetts Institute of Technology. An IIDA user at a terminal communicates with the MIT computer which houses the IIDA software. This computer also houses the CONIT software, developed by Marcus and Reintjes, which performs some vital functions for the IIDA software.[2] When actively searching, the user is connected with DIALOG through either TELENET or TYMNET. The user employs the DIALOG language, and receives standard DIALOG responses. IIDA adds additional information and offers its own help facilities, but does not offer an alternative to using the DIALOG language. For the studies reported here, the database used was Compendex, but experimental work of various kinds has also made use of ERIC, NTIS and ONTAP ERIC. A detailed description of the IIDA software can be found in an IIDA report and in Toliver.[3]

User Introduction to IIDA

One of IIDA's design principles was to make its use as much like the direct use of DIALOG as possible, except when the user needs help. After logging on, the user is offered a brief, optional introduction to IIDA which, if accepted, provides a summary description of the system and of the available IIDA services. The user is also told how to ask for help and how to quit at any time.

While interacting with IIDA and the database, the user does so in either of two modes: instruction or assistance. In the assistance mode,

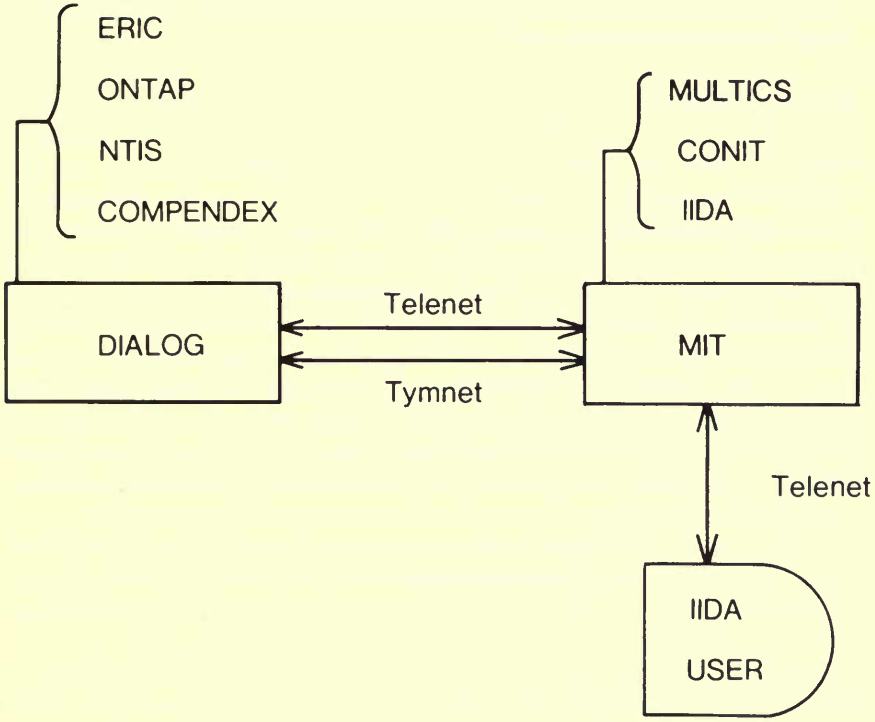

Figure 1. Schematic of IIDA physical arrangement

described later, the user begins searching, enters a series of DIALOG commands, and receives the same responses as he would without IIDA, if the commands are syntactically valid or do not trigger any of the diagnostic messages. If the user is not previously trained or desires a refresher course, he might begin with one of the computer-assisted training courses in the instruction mode.

Instruction Mode

The three exercises which constitute the instruction mode were designed to provide the user with: (1) the basic information necessary to do a search, (2) an opportunity to test these skills in a hands-on searching context, and (3) an opportunity to be exposed to advanced training in search commands and search strategy. While the user is allowed to begin

use of IIDA at any point, and to proceed in any order, the recommended path for the novice is to do the three instructional exercises, in order, before going into the assistance mode.

"Canned" Search

In the first exercise, the user is introduced to the basic commands of DIALOG—BEGIN, SELECT, EXPAND, COMBINE, PAGE, TYPE, LOGOFF—through the medium of a "canned" search. The user is instructed to enter the commands which are given to him (i.e., he does not choose them; he is told, for example, to enter BEGIN 8). Then he sees a DIALOG response to the command, and the command is then explained. The intent is to maximize self-discovery. All of the commands are used in the context of an actual search, and so are used in the way they might be in a "real" search situation. In addition to this exposure to search commands and search strategy, the user is also introduced to the concept of a two-cycle search. A two-cycle search is, basically, one in which the user creates a set of references, and then, based on set size, browsing, or even intuition, cycles back to refine that set.

The user at this stage is also provided an introduction to the IIDA "help" facilities. These facilities enable a user to get: (1) the definitions of search commands, (2) advice on current problems, (3) information about commands given in the current search, (4) a list of the sets created up to the point at which help is requested, (5) a summary of the records viewed up to the point at which help is requested, (6) a list of the errors made, (7) a list of the descriptors used, and (8) instruction on how to change from the present exercise or mode to another. A more detailed description of this and the other instructional exercises can be found in an IIDA report.[4]

Practice Search

The second exercise is a practice search in which the novice user is asked to try out the use of some of the things learned in the first exercise. The user is also introduced to goal setting in searching, in that he is asked to set a search goal for a specific number of citations. There is no stipulation that he actually meet this goal, but several of the diagnostic messages that he might receive refer to it.

During exercise one, the computer is not connected to DIALOG, and the search which is done always produces the same results, since the information is stored in IIDA. During exercise two, the IIDA computer is linked with DIALOG, and the search performed is completely under the control of the user. Normally, it is expected that only the commands to which the user has been introduced will be encountered in this exercise, but there is no restriction on usable DIALOG commands. Similarly, IIDA

suggests to the user some sample topics for searching in this exercise, but the user is free to search any desired topic. Those topics suggested to the user are generally simple ones for which there is a relatively high probability that the user will be able to do a successful search. An early success is desired to provide the satisfaction of seeing the system work correctly and productively.

Reference Walk

The third exercise in the instruction mode consists of a required look at summary descriptions of the available advanced instructional materials with the option for self-directed advanced training. These options are illustrated in figures 2 and 3. The contexts of this exercise include: (1) a review of the basic commands, (2) information about advanced commands, (3) an introduction to text searching, (4) further information about search strategy, (5) database description(s), (6) information about beginning and ending, and (7) a discussion of the IIDA facilities. As illustrated in figure 2, the user is introduced at a general level to topics on which considerably more detailed information is also available. Once this introduction is completed, users have a choice: they can either go on to the assistance mode, or return to do self-selected advanced training in various portions of exercise three (see fig. 3).

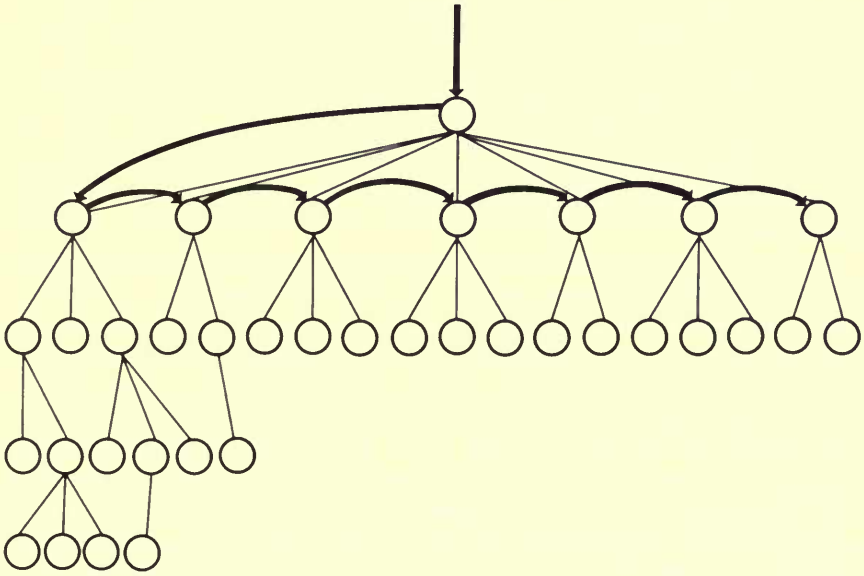

Figure 2. Reference walk in exercise three

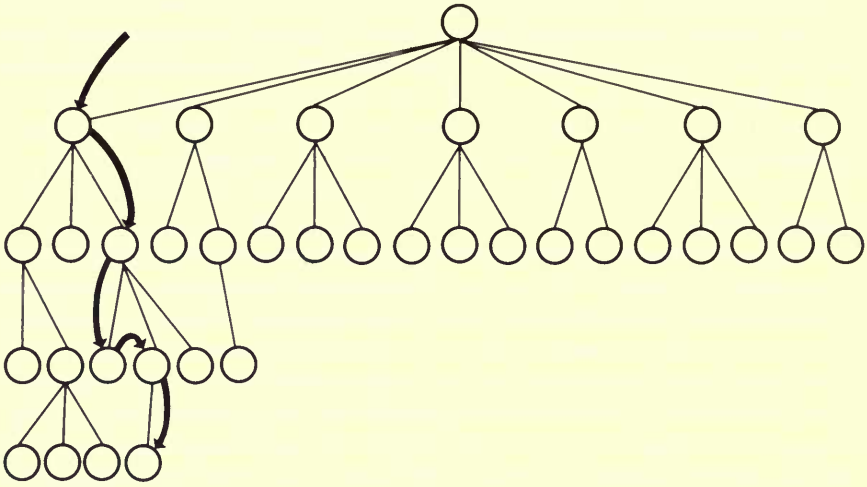

Figure 3. Self-selected advanced training in exercise three

Assistance Mode

In the assistance mode, the searcher performs his own search. The purpose of IIDA in the assistance mode is to offer assistance, to enable the user to do a search in DIALOG with a minimum of difficulty and interruption. Besides actually doing a search in the assistance mode, the user can, at any time, leave the search and enter any of the instructional exercises, or start the search over. Although it has not in fact been tested, in principle the user should be able, if patient enough, to do a search by going directly to the assistance mode without first using the three exercises in the instruction mode.

The three major features of the assistance mode, in addition to the flexibility it offers the user, are: (1) the reference help library, (2) quick advice, and (3) the set of diagnostics which monitor searcher behavior. As shown in figure 4, the reference help library contains all the information of exercise three in the instruction mode, but here the user is allowed free access to any information in the library at any level of detail selected. In addition, IIDA will also provide quick advice. Quick advice involves suggestions to the user on how to proceed in problem situations. In some cases, the nature of the problem is unambiguous, and IIDA can reference a particular bit of information in the help library. In other cases, the suggestion may simply be to use the help library. Once the desired information has been obtained, the user can signal for an automatic return to the search.

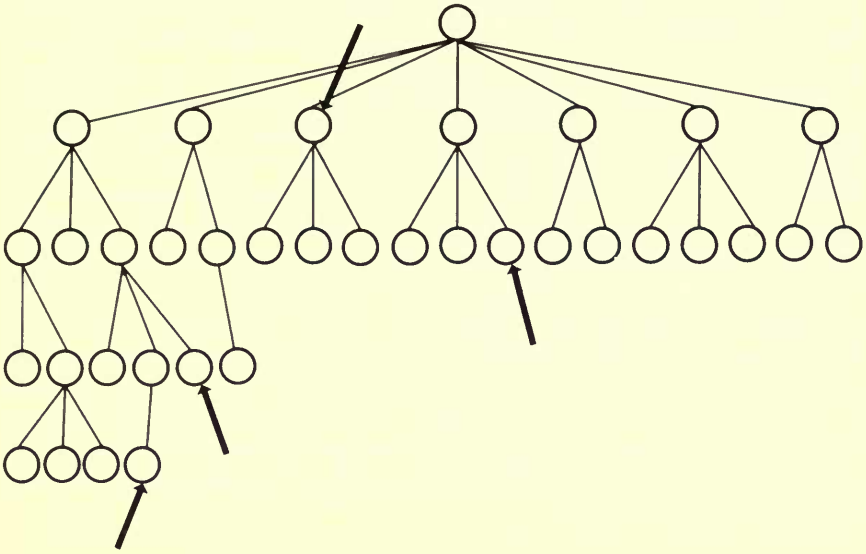

Figure 4. Free access to the reference help library during assisted searching

The most important aspect of the assistance mode, however, is the set of diagnostics which monitor searcher behavior and offer advice to the user. These diagnostics consist of a set of rules, clustered into categories, which are used by the computer to decide which messages to send to the user. Thus, IIDA is primarily a reactive system—the diagnostics are used to detect errors or tendencies which may lead to problems. If a diagnostic routine detects a problem, IIDA intrudes on the user with a statement of what the problem appears to be, and an indication of how to get advice or reference information.

DIAGNOSTIC PROCEDURES

Once a search begins, the user enters a series of DIALOG commands and receives the same responses as he would without IIDA, if the commands are syntactically valid. However, each command and its DIALOG response are intercepted by IIDA, parsed, and the components analyzed by the diagnostic procedures. If errors are discovered, IIDA intrudes into the communication between the user and DIALOG. In the case of a straightforward syntax error, the command is rejected (as it would be by DIALOG), and a specific statement of the problem is given the user.

The term *error* is used guardedly for other diagnostic procedures, since the nature of the problem is not one susceptible to algorithmic solution. Except for syntax errors, it is not possible to be entirely certain that a given command is "wrong." This judgment requires a knowledge of the goal and intentions of the user within a completely closed or bounded problem in which it is possible to specify all actions as either "right" or "wrong" on the basis of a set of well-defined rules. In this case, "wrong" and "error" are relative, since the diagnostics operate on the basis of certain heuristic principles. Thus, the diagnostic routines are intended to identify various behavior patterns which might lead to, or already be, a problem. The diagnostic messages are generally words of advice or caution to a user. They identify a problem or potential problem, and offer the user access to advice or instruction specific to that problem. The user is completely free to assume he knows what is going on, and to continue doing whatever triggered the diagnostic message. Given the open and unbounded nature of the problem involved, it may well be that the more experienced user knows exactly what he is doing.

Design Principles

The IIDA diagnostic system was based on the following principles:

1. The diagnostics have no access to any information about a search other than the search history; that is, IIDA can have no prior knowledge of how a given search "should" be performed.
2. Hence, IIDA is *reactive*; it helps a user decide how to proceed, given his performance to a point. It does not initiate directions in the sense of telling the user how to proceed with the search.
3. Diagnostics analyze user commands and the search service's response to them, determine problems, and point generally toward solutions. Emphasis is on encouraging self-discovery of solutions, e.g., when an error is detected, IIDA will suggest where advice or instruction is available, but will not force it on the user.
4. In making up the messages to be sent to the user, there was a conflict between the desire to offer the fullest explanation of any error (or apparent error) and the desire to avoid distracting the user by too much verbiage. The choice made was that all messages to the user reporting on error conditions should be as short as possible. Further, they should be unemotional. It is particularly important to avoid a conversational tone which users might interpret as pejorative ("You are wrong"), patronizing ("Good for you"), or cute ("Well, [name of user], you seem to have done this before"). This decision arose both from the designers' experiences with other systems which became abrasive quickly, and from the

belief that serious users will be content with straight information and will derive satisfaction from task accomplishment with a minimum of "chatter."

5. Certain errors must be brought to the user's attention each time they occur. For example, a syntactic error in formulating a command must be corrected if the command is to be executed. Other types of errors need be brought to the user's attention only if repeated often. This leads to the need for *thresholds, suppression* and *enhancement.*

   For some errors, a *threshold* may be established allowing the error to be repeated a certain number of times before it is brought to the user's attention. In this case, the error message is *suppressed* until the threshold is reached. Once the threshold is reached, the message is sent, the index is reset, and when the new value is exceeded, a stronger or *enhanced* message may be sent to the user.
6. IIDA diagnostics do not, except indirectly, assist in the selection of appropriate search terms. Use of DIALOG's dictionary is stressed in the instructional material. Failure to use the EXPAND command (which searches the dictionary) may result in a diagnostic message, but IIDA does not suggest what actual terms to use.
7. Throughout the design of IIDA diagnostics, the intent was to focus the user's attention on the concept that a search is a structure to be designed and executed as a whole. At the same time, the user's easy access to help should relieve anxiety about mechanical detail by assuring him that forgotten details can easily be retrieved.

IIDA Concept of a Search

Conventionally, a DIALOG search starts with a BEGIN command and progresses through EXPANDs and SELECTs to one or more COMBINEs, and then to TYPE or DISPLAY.[5] After browsing, a user typically goes back through one or more of the commands—EXPAND, SELECT, COMBINE—and then browses again. Each excursion through EXPAND, SELECT, COMBINE, and TYPE, even if not all of these commands are used, constitutes a *cycle.* A sequence of commands of the same type constitutes a *string.* A cycle, then, consists of one or more strings, each of which consists of one or more commands of a given type. This is illustrated in figure 5, where strings of length one, two and three are shown, and these form two cycles. The second cycle begins when, following a TYPE command, the user reverts to SELECT. A number of the IIDA diagnostic procedures make use of this concept of strings and cycles.

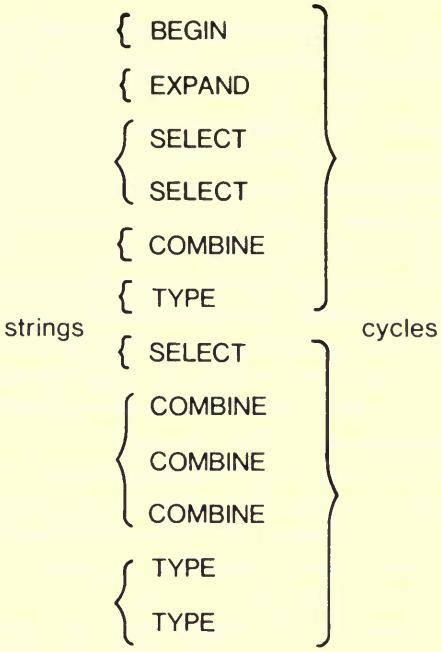

Figure 5. Strings and cycles in a search

Classes of Terms

The diagnostic procedures fall into the following three broad classes

1. *Syntactic.* A DIALOG command may be viewed either as a simpl sentence consisting of a verb and (usually) a noun phrase, or as compound word. It is conventional computer science terminology t refer to the governing structural rules as *syntax*, although *morpholog* may be more accurate, since a command lies somewhere between compound word and a sentence. At any rate, the first class of diagnostic is concerned with the validity of a command, and this consideration i entirely context-free. The guiding operational rule is that a command i valid if and only if it would be accepted by DIALOG.
2. *Individual command usage.* Given a command that satisfies context free structural rules, there may yet be problems that range from fata errors (such as the use of an undefined set number in a COMBINE command) to mere inelegance of usage (such as repeating a previou command). The context of the analysis is limited. The diagnostics i this class consider the most recent command in the context of th

accumulated history of previous commands. In a sense, this form of analysis is syntactic in that it is more or less analogous to analysis of sentence structure in natural language.

3. *Command string usage.* Command string diagnostics are concerned with a set of commands as an entity, rather than with individual commands. These diagnostics are, in effect, concerned more with style than with mechanical exactitude.

Diagnostic Subclasses

In addition to performing diagnostic functions, the syntactic diagnostics are also an inseparable part of the IIDA command and response parsers, which maintain the history of the search and interact with communications control software. The nonsyntactic diagnostic procedures are, however, implemented as a series of rules, not unlike rules in a formal decision table.[6] Below, the diagnostics are described in terms of subclasses of diagnostics which represent groups of rules. The actual statement of the rules is very much dependent upon the specific database and search service used, while these subclasses remain fairly general.

1. *Syntactic.* All syntactic diagnostics begin with a left-to-right scan of the command text. If an unacceptable element is detected, the user is informed of what part of the command text is acceptable and at what point it becomes unacceptable. A response is made to all syntax errors.
2. *Individual commands.* These diagnostics consider a single command in the context of previous search history.
   a. *Repeated commands.* In DIALOG searching, in general, repeating a command with its argument is wasted motion. There are exceptions to this, such as if the searcher is working at a CRT terminal and does not have a printed record of his previous work, or if he wishes to use the PAGE or DISPLAY SETS commands. But to repeat a SELECT, COMBINE or TYPE generally is an indication of careless work. Further, IIDA can recognize what we call *essential repetition*, which is a repetition of the logical essence of a command without it being exactly the same command (e.g., COMBINE 1 * 2 is logically equivalent to COMBINE 2 * 1). A single instance of a command repetition fault is not serious.
   b. *Null sets.* Generally, a null set resulting from a SELECT command is an indication of poor choice of terms. If it results from a COMBINE command, it may represent no more than an infelicitous choice of Boolean expression. Null sets created with the SELECT command indicate that the user should take some steps, probably thesaurus browsing, to find better terms. Nulls created with the COMBINE command

simply indicate the need for a different choice of expression. Repeated null set errors are assumed to be serious, and positive steps should be taken when they occur.

c. *Unused sets.* Sets are "used" when they are referred to in commands such as COMBINE or TYPE. If sets in any appreciable numbers are being created but never used in later commands, the assumption is that they are perhaps ill considered. This may be an indication of thrashing, i.e., the rapid change of direction of searching. IIDA counts the number of unused sets created in each cycle, but comments to the user only when, after a cycle, the number of unused sets is greater than it was at the end of the previous cycle, indicating a possible trend toward creating an increasing number of sets never used. Null sets are not charged to the unused set count. This is considered to be a moderately serious problem.

d. *Print format.* There are two types of print control errors. The first is use of a format that is likely to be uninformative. The extreme case here is a DIALOG format which gives accession numbers only, and is of no help to a searcher for browsing purposes. This is an example of a problem which might not occur often, but could unnerve a user when it does happen.

e. *Excessive printing.* The second type of print control error involves excessive printing. Novice users often print excessive numbers of records on-line because they are not used to sampling. They do not know the use of the off-line print commands, and they may be unaware of the costs. In addition, the experimental IIDA system terminals operated at 300 baud. At this speed, economy of time was important because of the limited number of hours during the day when the research studies could be conducted. Consequently, it was necessary to limit the number of citations a user could print with a single command. This limit varied with the format used.

f. *Response time.* The time between a machine's message and the input of the next command by the user is measured. Excessive consumption of time is, of course, expensive, and is also an indication of searcher difficulty. In addition to its pedagogical value, the time measurement can be used to terminate a user's session if the delay is extreme, indicating that the person has probably left the terminal but has not logged off. Time is not critical to search logic, but long delays inhibit the extent to which a person can remember all he has done previously. Time, in early experimental work, has been found a good discriminator between experienced and inexperienced searchers.[7]

3. *String diagnostics.* These are the diagnostics that look at sequences of commands.

a. *String and cycle length.* Tests are made for length of a string, by type

of string, and for length of a cycle. The basic assumption is that any unusually long segment of a search might be an indication of a user problem.

b. *Thrashing.* This is changing the "direction" of a search rapidly or often, without pursuing any given direction far enough to see if it can work out. Searcher motivation for thrashing might be overly easy discouragement with preliminary results, such as getting a null set or an extremely large set from one search formulation; rather than trying to modify that formulation, the user turns to a completely different approach. At its worst, thrashing is indicative of random, uncontrolled behavior. It is, of course, not possible to measure "direction" precisely. It is done in somewhat arbitrary fashion by taking a measure of the similarity of the Boolean expressions in successive COMBINE commands. One consequence of this is that thrashing can occur only within a string of COMBINEs. The measure, called the *similarity index*, is made up from the percentage of implied terms common to the two expressions. "Implied" here means that an expression is expanded by replacing set numbers with their defining terms, and then the similarity count is done on terms, not set numbers. IIDA computes the number of terms in common, divides that first by the number of terms in one expression, then by the number in the other, and then takes the mean. This gives the similarity index between any two COMBINE commands. Thrashing is a mildly serious problem. A little bit of it does no particular harm. A great deal of it indicates a searcher is in trouble.

c. *Dwelling.* This is behavior opposite to thrashing. It is remaining with a search concept when it may be time to give up and try another approach. As with thrashing, dwelling is defined only within a string of COMBINEs. The similarity index is used to recognize dwelling, but this time it is in cases of similar COMBINE expressions, rather than dissimilar, that an IIDA message is sent to the user. Also used is the concept of convergence or divergence. The searcher will have been asked at the beginning of an IIDA-mediated search to state the search goal as a numeric step function. If a string of similar COMBINEs shows progress toward that goal, there is less concern about dwelling; if there is progress away from the goal, but the user remains with the same basic search concepts as indicated by the similarity index, he is dwelling. Regardless of convergence or divergence, if a similar string is too long, the user is dwelling. Dwelling can be a serious fault in that it might indicate a lack of understanding on the part of the searcher as to what a mechanical search, performed by someone of his skill level, is capable of producing. Perhaps the best example is a case in which the database simply does not have much information on the subject sought. Excessive dwelling does

not help much; the problem is to realize the meagerness of the lode.
d. *Relevance.* Much has been written about the nature of relevance, all of which is sidestepped in IIDA. Relevance in this context is a judgment made by a searcher about the value of search results to him. There is no way to control how he thinks about the concept, and there is no diversionary teaching on the subject. IIDA does ask the searcher to make a relevance judgment after each printing of a record (using the command TYPE or DISPLAY), unless the display was done using format 1 (which shows only accession number). The judgment is made on a five-point scale, from irrelevant to highly relevant, with "1" representing "irrelevant." Diagnostics are performed on the average relevance figures for a group of records displayed with a single command, as long as at least three records are typed. Relevance diagnostics are not based upon faults or search errors, but they are used to direct the searcher's attention to the fact that a particular set seems not to be fruitful, or that a previously examined set yielded higher relevance scores. Although low average relevance figures give no hint as to the remedy, the problem detected is a serious one—the searcher has come to the end of a cycle and is dissatisfied with the results according to his own definition of relevance. If this condition recurs repeatedly at the end of cycles, the entire approach to the search comes into question.

TESTING

The testing of the diagnostics as a way of assessing user performance necessarily involves the testing of users as well. There have been essentially three major phases of this testing. These consisted of the pilot testing of the entire system to ensure that it functioned as intended, a baseline or benchmark study conducted to establish a set of reference points for the operation of the diagnostics, and a training method study done to determine whether the diagnostics operate differently with people trained in differing ways.

Pilot Testing

The pilot testing conducted with IIDA was oriented primarily toward the discovery of either programming or conceptual problems with the operation of the system. Among the groups working with the system during this early testing were undergraduate computer science majors who acted as users and who were challenged to find the problems. In addition, a number of faculty and graduate students from Drexel's School of Library and Information Science were invited to review the operation of the system

and give their advice and comments. All of the members of this latter group were trained searchers, and some of them are regularly involved in the training of searchers. Another pilot test group consisted of a number of undergraduate engineering majors, who did the search training as part of a course on technical writing. The experiences and feedback of each of these user groups resulted in successive changes and refinements in the operation of the system. For example, extensive cross-comparison of the search transcripts from the technical writing students and the records kept by IIDA provided a check which ensured that the machine record-keeping programs did what they were programmed to do.

This kind of pilot testing is an indispensable part of the development of any system as large and complex, and as user-oriented, as IIDA. Because of their closeness to the system and the problems of conceptual design, the system designers are at times not the best judges of what material should be viewed by, or must be explained to, the user. The following examples, which were easily corrected by revision of user instructions or of the instructional material, illustrate this point well. Even though their initial classroom description of IIDA was of a computer system that would assist them in retrieving references for use in writing required course papers, a few of the technical writing students started the training initially expecting to retrieve facts about their search topic rather than references relevant to the topic. Other students did not seem to realize that the commands encountered in exercise one should be learned for future use. Another problem which showed up in the early stages of training the technical writing students involved the effects of search commands in the context of a search. The version of exercise one used with these students illustrated a search on the topic of library automation. At one point in this search, the user was told to EXPAND LIBRARY. A few of the users subsequently thought that the next time they wanted to EXPAND a term, they were to enter EXPAND LIBRARY, no matter what term was to be expanded.[8]

Baseline Study

The baseline study was conducted in 1979 at the Exxon Research and Egnineering Company facility at Florham Park, New Jersey.[9] The study design called for collecting data on the searches performed by twenty-five searchers who were trained and assisted by IIDA. In addition, twenty-five comparable searches were to be done by the information retrieval staff on site. The intent here was to use the intermediated searches to provide a profile of diagnostic usage for comparison with the diagnostics usage of the IIDA users. It was anticipated that significant differences between the two groups of searches would reveal deficiencies in the IIDA training and

assistance, which would then be subsequently corrected by appropriate modifications of the system.

Procedure

Participants in this study consisted of fifty Exxon research engineers employed at the Florham Park site who had been recruited by mail. Approximately 2200 letters were sent out offering the opportunity for technical personnel to do on-line searching on their own via the IIDA system. Approximately 150 responses were received. The names were randomized so that date of response, level of position within the company, and the department and section would not be considerations for participation in the study. From this randomized list, twenty-five participants were then randomly assigned to the IIDA training group, and twenty-five were randomly assigned to the intermediated group.

Participants assigned to the IIDA training group were scheduled for two sessions which involved the IIDA training followed by two IIDA-assisted searches on topics of their own choosing. Each of these participants filled out a post-search questionnaire after completing the second search. Participants assigned to the intermediated group brought their next search topic to one of the Information Center staff, as they would normally do. This search, however, was done by the intermediary through IIDA with all IIDA responses and messages suppressed. Thus, the only aspects of IIDA working for the intermediated searches were the record-keeping functions, which kept track of the search history and of the number of times various diagnostic rules were triggered by the searching behavior of the intermediary. A post-search questionnaire was given to each of the participants in the intermediated group, along with the results of the search requested. Since they had been recruited with an offer of search training, these participants were scheduled for IIDA search training after completing the post-search questionnaire.

Results

The post-search questionnaire for participants in both the IIDA and the intermediated groups included a question which asked what percentage of the references retrieved were: (1) very useful, (2) useful, (3) useless. For purposes of analysis, the "very useful" and "useful" categories were combined into the single category, "useful." The difference between the two groups on this measure was not significant (Mann-Whitney U Test, $p > 0.05$), with the IIDA-trained group reporting an average of 52.5 percent useful references, and the intermediated group an average of 49.3 percent useful references retrieved. Consequently, it appears that the results of the searches were valued equally by those who did their own searches and those who had a search done for them.

Turning to the diagnostics, the pattern is one in which there appear to be no differences between the two sets of searches. The mean frequencies for each group of searches for each of the three major categories of diagnostics—syntax, the individual command usage or "local" diagnostics, and the command string usage or "global" diagnostics—are illustrated in figure 6. For each of these classes of diagnostics, the differences betwen the two sets of searches were not statistically significant (Mann-Whitney U Test, $p > 0.05$). This finding also holds true for each of the subcategories of diagnostics contained within each major category.

In light of the intent of this study, i.e., to provide a set of diagnostic bench-mark criteria against which to assess the performance of the IIDA-trained and assisted searchers, it was surprising to discover a lack of significant differences between the two groups of searches. A significant difference on one or more of the diagnostics would have pointed toward some deficiency in the IIDA training, or in the usefulness of the diagnostic messages during IIDA-assisted searching. This would have led to revision of one or both aspects of the system. The finding of no significant differences poses a potential problem in that this could suggest either that the IIDA training and IIDA diagnostics work well, or that they are totally irrelevant. One reason for arguing that the training and diagnostic assistance worked as intended is that the end user evaluation of the utility of the information retrieved in the searches did not differ significantly between the two groups. This means that a group of individuals who had never before done on-line searching were able, with IIDA training and assistance, to do searches which produced results containing as much useful information as contained in searches done by professional searchers. Under these circumstances, the notion that the diagnostics are irrelevant as measures of searcher performance seems implausible.

It should be mentioned in passing that there is one subcategory of diagnostic information which is not included in the results reported in figure 6. Since the intermediaries did their searches with IIDA suppressed, they were not asked by the system to make relevance judgments on the references typed out. Consequently, this diagnostic category was not included in the data of the IIDA users when comparisons were conducted between the IIDA-assisted and the intermediated searches.

Training Method Study

The second major test of IIDA and the diagnostics took place at the Exxon Research and Engineering Company facility in Linden, New Jersey, during the winter months of 1979-80. This study involved a comparison of the searches done by IIDA-trained and assisted searchers with those

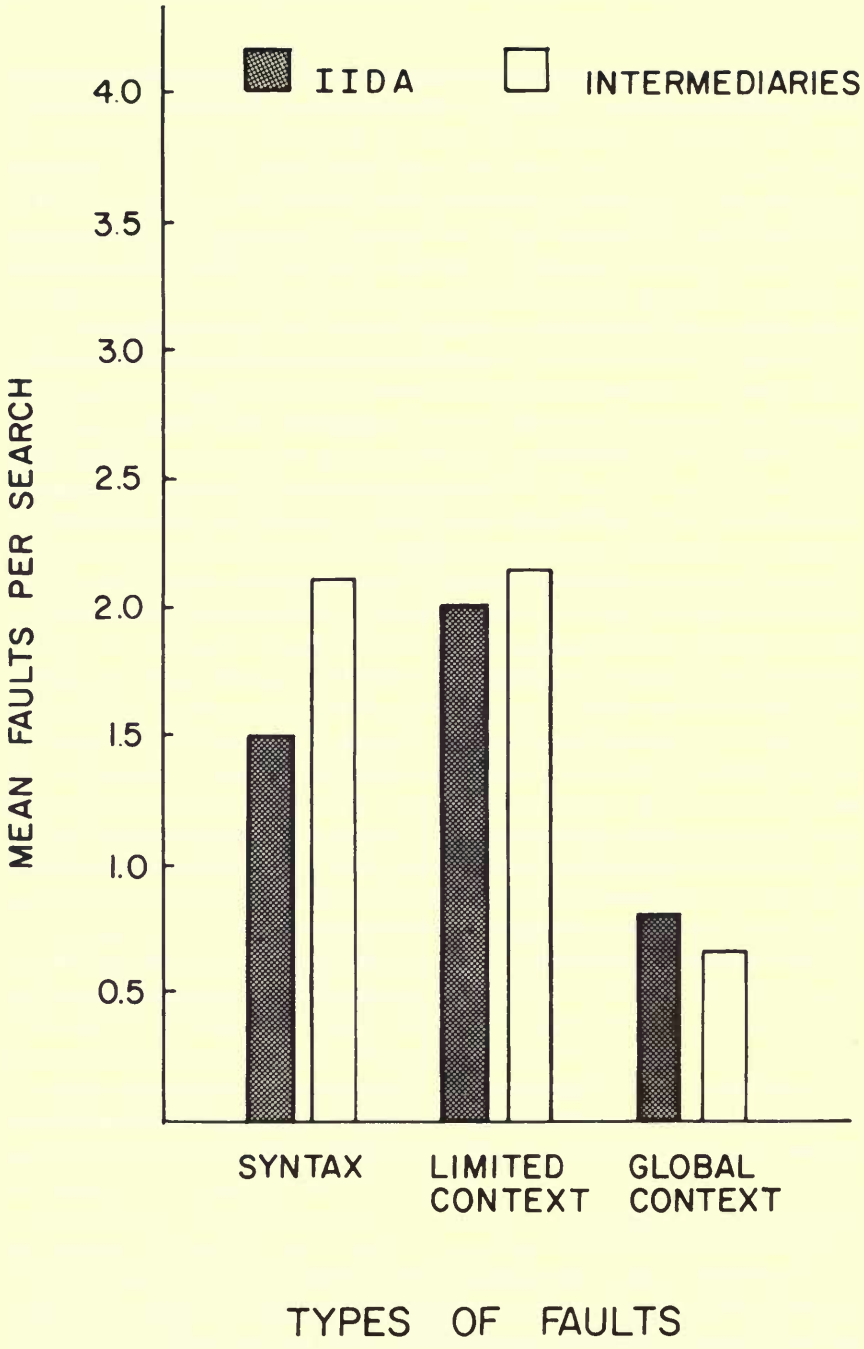

Figure 6. Average frequencies for three major classes of faults of Florham Park

done by a group of searchers who were also assisted by IIDA, but who had been trained in a half-day classroom training session. The intent of this study was to test the instructional procedures and materials involved in IIDA training against a more frequently used and conventional training method. Presumably a human instructor is more responsive to the problems which a student may encounter, and presumably the student has a greater range of questions which can be asked of a human instructor, especially since IIDA is not designed to be a question-answering system. Consequently, it was anticipated that significant differences between the two groups of searchers would reveal deficiencies in the IIDA training, which would subsequently be corrected by appropriate modifications of the system.[10]

Procedure

Participants in this study were primarily research chemists and other scientists. Participants were again recruited by mail, with letters sent to the entire research staff describing the opportunity for bibliographic search training and asking for volunteers. As with the earlier study, about 150 responses were received. From the list of volunteers, fifty were randomly selected to participate. Twenty-five volunteers were randomly assigned to receive IIDA training and then do two IIDA-assisted searches, while the other twenty-five were assigned to the conventional training sessions, with two IIDA-assisted searches following training. All participants in the conventionally trained group received training by Exxon Information Center staff in one of four morning sessions. All of the training sessions used the same training materials (handouts, etc.), covered the same material, and allowed some on-line time for the users to practice what they had learned.

Results

In this study, the post-search questionnaire, which was administered after completion of the second IIDA-assisted search, asked the user to decide what percentage of the references retrieved had been: (1) very useful, (2) useful, (3) useless. Again, the first two categories were combined into the single category, "useful." The difference between the two groups, while apparently large, was not significant (Mann-Whitney U Test, $p > 0.05$), with the IIDA-trained group reporting an average of 46.6 percent and the conventionally trained group an average of 62.9 percent useful references retrieved.

Concerning the diagnostics, again there appear to be no differences between the two sets of searches. The mean frequencies for each group of searches for each of the three major categories of diagnostics are illustrated

in figure 7. It should be noted that the data in this study do include the subcategory of relevance diagnostics in the "global" diagnostics, since both groups did their search with IIDA fully operational rather than suppressed (as was the case for the intermediated searches done in the study described earlier). For each of the three categories of diagnostics, the differences between the two sets of searches are not statistically significant (Mann-Whitney U Test, $p > 0.05$). Furthermore, individual comparison of the subcategories of diagnostics indicated no significant differences between the two groups.

Considering that the purpose of the study was to compare IIDA training with a more conventional method, the lack of significant differences between the two groups was surprising. A significant difference on one or more diagnostic measures would have indicated either a flaw in the IIDA training materials or in the diagnostic messages during assisted searching. The result would be a revamping of one or both parts of the system. Since no significant differences were discovered, there is the problem of interpreting the results, except for the fact that the groups were able to retrieve a significant percentage of useful references during their searches. Presumably this represents something which neither group would have been able to accomplish if simply turned loose with a terminal without any training. Further, the two groups did not differ in their estimates of the percentage of useful references retrieved. This pattern of results argues strongly that IIDA training, as it is presently structured, represents a viable alternative to the type of conventional training with which it was compared.

Looking at the data from both studies and comparing both sites by diagnostic category, there were no significant differences in the usage of the diagnostics frome one site to the other. In addition, judgments made by users about the percentage of useful references retrieved did not differ significantly from one site to the other. Thus, the two groups trained at Linden did not produce results appreciably different from those obtained for the group of intermediated searches done at Florham Park. It should also be mentioned that the variability in the measures, including percentage of useful references retrieved, did not differ significantly from one group to another. One final point to be noted is that the users seemed to like IIDA. Approximately 90 percent of the users, either trained and assisted by IIDA or just assisted by IIDA, said they would recommend use of IIDA to their friends.

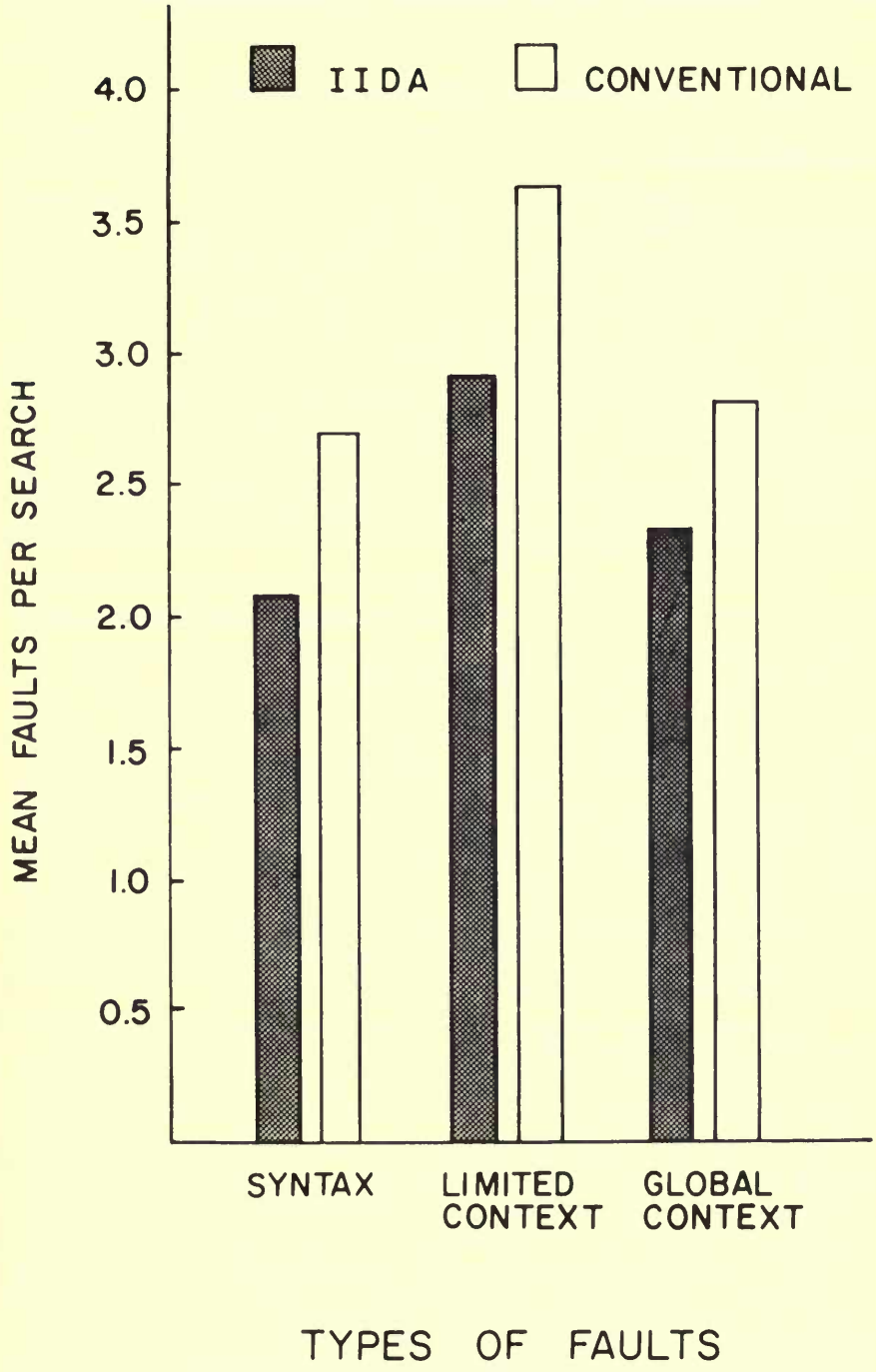

Figure 7. Average frequencies for three major classes of faults at Linden

DISCUSSION AND CONCLUSIONS

User Performance

It is not possible to separate completely the performance of the IIDA diagnostics from the performance of the IIDA users. Focusing first on the users, however, the data suggest that IIDA users did searches which were, on two kinds of measures, done as well as those done by a professional searcher. First, in the ratings of the utility of the overall results of the searches, there was no significant difference between the results obtained by the professional searcher and those obtained by the user performing his own search. Second, no significant differences were found in the number of IIDA-detected errors or faults between the newly trained, IIDA-assisted searchers and the professional searchers. Together, these findings imply that the results of the searching were about equal, and that the performance of the searchers was about equal.

It was initially a surprise to find that the error/fault rate seemed about equal between the professional and neophyte searchers. IIDA was designed to produce acceptable results for its users, but it was never expected that the step-by-step performance would match that of professionals. There are, however, some mitigating circumstances which may have handicapped the professionals somewhat. For example, the Exxon professional searchers are used to working with 1200-baud terminals. With these terminals, more on-line typing is acceptable than with the 300-baud terminals used for IIDA searching. Hence, the professional staff was accustomed to doing a great deal of on-line printing, often exceeding the limit allowed for IIDA users working with the slower terminal. Also, IIDA's design was fixed just before Lockheed announced the new SUPER SELECT command.[11] Thus, this command was not included in the IIDA training materials, and its syntax was not recognized by IIDA's parser.

Although the intermediaries knew about these limitations, an inadvertent transfer of habits from the usual search context into the IIDA search context could have introduced some faults or errors which might not otherwise have occurred. For example, if a searcher used SUPER SELECT for set combination, then there was an increased likelihood of the appearance of a large number of SELECT commands without any COMBINE command. This could then trigger an "excessive string length" diagnostic. Consequently, although it is not possible to assign numerical values, there is reason to believe that the professional searchers might have had a higher error/fault rate than normal. However, it is certainly the case that the IIDA searchers performed respectably well, and, in terms of final outcome, the users who did their own searching using IIDA achieved results equivalent in utility to those who worked through intermediaries.

The other surprise in the results on user performance came in the findings that the two different user groups trained by different methods did not differ appreciably either on the diagnostic measures or in the percentage of useful references retrieved. The original design specifications for IIDA did not envision a system which would be competitive with more conventional training approaches. Rather, the intent was to provide an avenue of access for individuals who could not or would not take a conventional training course, but who still wanted personal (rather than intermediated) access to a database. It had been assumed that comparisons between the two different training methods would highlight difficulties with IIDA. One reasonable guess as to the reason for the pattern of results reported here is that, while the human instructor may well have been more flexible and responsive in assisting the student and in answering questions, the design of IIDA does, as intended, enable the user to discover where to go to get the information needed to answer questions.

Performance of the Diagnostics

Earlier, the possibility was mentioned that the IIDA diagnostics, rather than indexing various aspects of searcher behavior, might be irrelevant to the process of searching. This would account for the lack of significant differences between IIDA users and professionals, and between IIDA-trained and conventionally trained users, on these measures. This explanation seems, however, to be less reasonable than the idea that the system is working as it was designed to work. One reason for not assuming that the diagnostics are irrelevant as indices of searching behavior is that the diagnostics were all empirically developed. In other words, they were all designed to index and deal with problems encountered by searchers which have been observed by the designers and/or reported in the literature by others.

Furthermore, the relative frequencies of the various types of errors or faults correspond reasonably well with the intuitions which originally led to the development of the diagnostics. Figure 8 shows the mean number of occurrences per search of the various subcategories of diagnostics, aggregated over the entire study. Also indicated are the 95 percent confidence limits for each type of diagnostic. The diagnostics which were triggered most commonly were syntax, relevance and null set generation. A second group consists of string length faults and excessive printing. The third group, of six fault types, is characterized by relatively low frequency of occurrence; in fact, some of the confidence intervals include zero. These last six are: response time, command repetition, print format, unused sets, dwelling, and thrashing. Even though these diagnostics were triggered

relatively infrequently, they are probably worth keeping, since the response to the open-ended questions on the post-search questionnaire suggested that for an occasional user they were useful.

In general, the diagnostics are mutually independent in that the triggering of one has no implication as to whether the others will be triggered. There are, however, some exceptions where there is an interrelation. For example, a string length diagnostic is triggered whenever a certain number (n) of successive commands of any given type is received. A dwelling or thrashing diagnostic requires both a string of COMBINE commands and certain conditions with respect to the arguments of the COMBINEs. In other words, there must be a string of length m, plus other conditions. If the threshold on the string length diagnostic were set short enough, that is, if $n < m$, a user entering a string of COMBINEs would be warned against continuing that pursuit before the dwelling or thrashing rules and diagnostic messages could take effect. The various thresholds were set heuristically, and these settings have some effect on which rules a searcher is deemed to have violated. Hence, a slight change in the thresholds might have resulted in the occurrence of one type of diagnostic rather than the other.

Conclusions from the Diagnostics

Overall, it seems that the various categories of diagnostics do represent a reasonable set of measures of the performance of on-line searchers. The fact that the novice searchers studied were able to achieve a degree of satisfaction with the final results equal to that achieved through professional searchers suggest that, with the help of the IIDA diagnostics, these searchers performed the search as well as the professionals. Enabling them to do so was the goal of the IIDA assistance mode program.

This equivalence would imply that when a searcher has an error or fault rate widely divergent from those reported, there also exists a difference in performance in some area of searching behavior. In the case of excessive errors or faults, this is presumably self-correcting over time with repeated exposure to the diagnostic messages and practice. In the case of a low error rate, the deviation may indicate a person ready for a language of greater complexity. That is, sustained error-free performance probably means that the user is ready to take on a language capable of greater logical power. Such languages are typically more complex and demanding in terms of what the user must know.

One important caution is that the IIDA diagnostics were designed for use in a certain limited context, that of the training and assistance of the kind of users described in the introduction to this paper and in the user

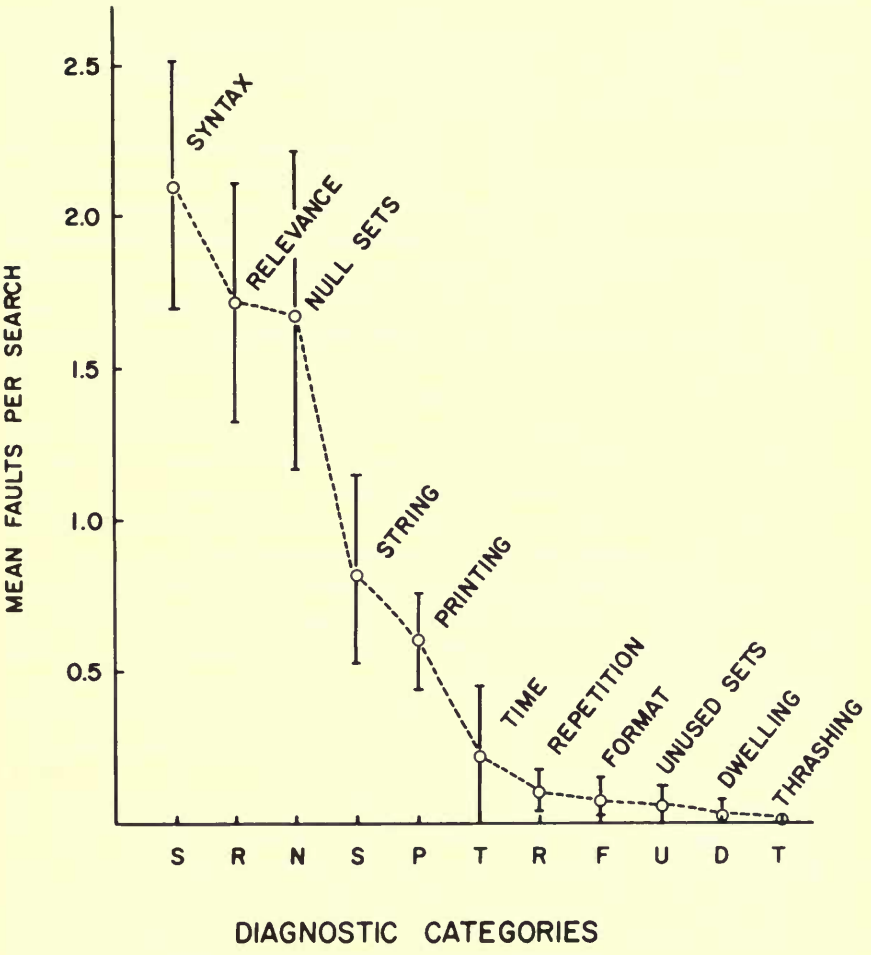

Figure 8. Average frequencies for the diagnostic categories for all users

studies conducted. There is some indication in the experiences of the professionals that when the diagnostics were employed with highly experienced searchers, they may not be the most appropriate procedures, i.e., there are times when the diagnostics may have been triggered by relatively sophisticated searcher behavior which is appropriate in one context, but not in situations with which IIDA was designed to cope.

Future Research

Inasmuch as the conclusions drawn about the effectiveness of the diagnostics are not as clear as might be desirable, several additional directions for future research should be followed:

1. *Diagnostic evaluation.* The overall requirements for evaluation of the project as a whole, as well as the available resources, dictated that some studies have a higher priority than others. This led to concentration on the evaluation of IIDA as a system rather than a more direct focus on the diagnostic component of the system. A future evaluation devoted solely to the question of the performance of the diagnostic system, with and without various kinds of training in advance, would be both useful and desirable.
2. *Longitudinal study.* A real limitation of studies such as those described here, and of many others on searcher behavior, is the relatively short amount of time between the participants' first exposure to the system and the evaluation of their performance or of the system performance. In addition, there seems to have been little or no testing of information system users at several points to investigate development or change over an extended period of time. In the studies described above, the volunteer subjects generally wanted to do the whole set of exercises, and then do their searching in a relatively short time. It seems particularly desirable that a long-range study be done, over a period of a year or so, focusing on how different people adapt to a new system, how quickly they adapt, how their behavior changes over time, and how it changes as the result of multiple search experiences.
3. *Adaptation to user skill level.* As noted above, it is not expected that the same diagnostics will be equally useful or desirable for all user skill levels. It would be desirable not only to test the diagnostics on persons of differing skill levels, but to test different variations of the diagnostics. In particular, it seems desirable to test whether, by varying the thresholds in the existing set of diagnostics as a function of actual prior performance, it is possible to get them to perform adequately with persons of different skill levels. That is, can they be made adaptive?
4. *Various user groups.* Although the original target user group for IIDA consisted of technically trained individuals interested in a particular class of search problems, there now seems to be no reason not to attempt extension of IIDA. The technically or scientifically trained users may be only one of several groups who would find IIDA attractive and useful. In particular, it seems desirable to determine whether or not a system such as IIDA can be used to provide direct database access for a wide variety of possible end users interested in a wide variety of search problems.

In summary, it seems that a new idea has been fairly tested in the very environment for which the concept was intended. Indeed, one of the important characteristics of the two studies discussed is that there appeared to be no differences among the results produced by the various user groups when there had been every reason to expect a number of differences. While some aspects of these results are not entirely conclusive, they support the idea that the IIDA diagnostic procedures did adequately measure important aspects of user performance. What is more certain, however, is that the IIDA system represents a way of training and assisting novice users to search databases with a level of performance that matches that of more experienced searchers. Furthermore, IIDA clearly represents a viable alternative to gaining direct database access for those end users who cannot or will not engage in more conventional forms of serach training.

REFERENCES

1. The work reported here was sponsored in part by the National Science Foundation, Division of Information Science and Technology, under Grant No. DSI 77-26524. The two user studies were also supported in part by a grant from Exxon Research and Engineering Company, which also made space available for the conduct of the studies. Special thanks are owed to Barbara Lawrence of Exxon, who played a major role in enabling the user studies to be set up, and to the staff members and the participant users of the Exxon information retrieval services at the sites where the studies were conducted. Finally, the authors thank Robert Rich, the IIDA consultant on evaluation, whose insightful comments were always valued even when not heeded.

2. Marcus, R.W., and Reintjes, F.J. "Experiments and Analysis on a Computer Interface to an Information Retrieval Network" (NSF Grant No. IST 76-82117). Cambridge, Mass., MIT Laboratory for Information and Decision Systems, 1979.

3. *Individualized Instruction for Data Access (IIDA). Quarterly Report No. 6* (NSF Grant No. DSI 77-26524). Philadelphia, School of Library and Information Science, Drexel University, and Franklin Research Center, Sept. 1979; and Toliver, David E. "A Program for Machine-Mediated Searching," *Information Processing & Management.* (In press.)

4. *Individualized Instruction for Data Access (IIDA). Quarterly Report No. 5* (NSF Grant No. DSI 77-26524). Philadelphia, School of Library and Information Science, Drexel University, and Franklin Research Center, June 1979.

5. *A Guide to DIALOG Searching.* Palo Alto, Calif., Lockheed DIALOG Information Retrieval Service, 1979.

6. For a full description of these rules, *see: Individualized Instruction for Data Access (IIDA). Quarterly Report No. 4* (NSF Grant No. DSI 77-26524). Philadelphia, School of Library and Information Science, Drexel University, and Franklin Research Center, March 1979. (ED179 195)

7. Fenichel, Carol H. *Online Information Retrieval: Identification of Measures that Discriminate among Users with Different Levels and Types of Experience.* Philadelphia, School of Library and Information Science, Drexel University, 1979.

8. *Individualized Instruction for Data Access (IIDA). Quarterly Report No. 7* (NSF Grant No. DSI 77-26524). Philadelphia, School of Library and Information Science, Drexel University, Dec. 1979.

9. *Individualized Instruction for Data Access (IIDA). Quarterly Report No. 8* (NSF

Grant No. DSI 77-26524). Philadelphia, School of Library and Information Science, Drexel University, March 1980.

10. Ibid.
11. *Guide to DIALOG Searching*, op. cit., pp. 3-6.