In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

3/17/65
b

A METHOD FOR INVESTIGATING THE BEHAVIOR OF ATTRIBUTES WHICH BELONG TO INFORMATION STORAGE AND RETRIEVAL SYSTEMS

A THESIS

Presented to

The Faculty of the Graduate Division

by

Ralph Paul Heckman

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Information Science

Georgia Institute of Technology

August, 1965

A METHOD FOR INVESTIGATING THE BEHAVIOR OF ATTRIBUTES WHICH BELONG TO INFORMATION STORAGE AND RETRIEVAL SYSTEMS

Approved:

____

____

____

Date approved by Chairman: Aug. 31, 1965

ACKNOWLEDGMENTS

The author wishes to express his thanks to Dr. David E. Fyffe who served as his advisor. Thanks are also extended to Dr. Vladimir Slamecka and to Dr. Arthur T. Kittle for serving on the reading committee.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

The purpose of this study is to develop and apply, by the way of illustration, a method for investigating the behavior of attributes which belong to information storage and retrieval systems. Although several attributes are common to many information systems, their values differ according to the conditions which are present in a given system. An investigation of the relationships between the conditions and the attributes can enlarge the operational understanding of the concept "information storage and retrieval system." An operational understanding of this concept is necessary in order to design these systems because it provides an a priori knowledge about the probable state that a system will assume. This state is defined as the values which the attributes will possess under specified conditions, once the system is in operation.

An operational understanding is currently being obtained and enlarged by the experience gained in analyzing problems and designing systems. Generally, the purpose of an analysis is to establish the objectives and constraints. Once these are established, the solution is devised with the aid of previous experience with similar objectives, constraints, and their solutions. This induced input to the design of a system is at best conceptual. Therefore, the knowledge from previous experience is currently being conveyed in an ambiguous fashion from one problem to another by the same individual, and from individual to individual. Thus, the value of experience to the design of a system is extremely restricted because of the lack of a method for collecting and representing the knowledge gained by past experience.

In order to collect data, it is necessary to establish a representative sample of information storage and retrieval systems and to identify the important attributes (parameters) of these systems. It is recognized that there are several types of information systems and that each system has its own objectives. Therefore, a representative sample of systems is difficult to establish.

It is also difficult to identify the important attributes of information systems. R. M. Hayes[1] has provided an outline which he calls "parameters of operation," and he suggests that the development of a defined set of parameters is a basic problem which must be solved. The lack of a defined set of parameters does not preclude a survey or an investigation; it only reduces its scope to those attributes which are currently being measured.

The National Science Foundation published, in 1962, a survey containing descriptions of several types of operating information storage and retrieval systems. While it was presumably not the purpose of this study to identify all of the important attributes or to establish a representative sample of all types of information systems, its results, while not being ideal for an investigation, are useful enough to be employed here.

Since the publication of this survey, several authors have indicated a need for performing a comparative analysis of the systems, but they have not suggested a method. For example, B. C. Vickery[2] proposed that the system descriptions may be examined to determine the conditions to which certain parameters are best suited, but he cautioned that more specific sets of parameters than those provided by the survey are needed before any criteria of retrieval and economic efficiency can be established.

The method developed in this study is based on an investigation directed at the operations of these systems. An operation of a system, such as retrieving, has attributes associated with the input, the process, and the output. By investigating the relationships between the attributes of one operation at a time, the problem can be broken into manageable segments.

The results of this study consist of functional producer-product relationships which are composed of attributes. For example, one relationship consists of the producer attribute, rate of growth of collection, and the product attribute, number of clerical personnel. The regression equation for this relationship is

$$y = 1.71 + .0001\ x$$

where x is the producer attribute and y is the product attribute. This relationship provides an a priori knowledge about the probable state that a system will assume if the value for the producer attribute is specified. (The specification of an attribute value constitutes the establishment of a condition.) If the value for the rate of growth of the collection is specified during the course of analyzing a problem, the value for the number of clerical personnel which the new system will possess can be estimated by solving for the unknown in the regression equation. The producer-product relationships indicate what is currently being practiced; they are not criteria for determining retrieval or economic efficiency.

CHAPTER II

A DEFINITION OF AN INFORMATION STORAGE AND RETRIEVAL SYSTEM

The present investigation of information storage and retrieval systems is directed at the intermediate outcomes of their operations. In order to identify the system elements and their arrangement, a general definition of systems will be considered, and the components of the systems under investigation established and fitted to a structure consistent with the definition. This establishes a standard from which to develop the model and investigate the sample systems.

General Definition

The general definitions of a system and its elements are as follows:[3]

A. System - A system is a set of objects, existing within a defined boundary, operating toward a common objective.

B. Objects - Objects are the parameters of systems: inputs, processes, and outputs.

C. Attributes - Attributes are the external manifestations of the ways in which an object is observed.

D. Relationships - Relationships are the associations of objects and attributes.

E. Component - A component consists of one process and at least one input or output.

F. Environment - The environment consists of factors outside of the system's boundary which affect or are affected by the system.

The nature of a system can be established by representing these elements as component parts (Figure 1).

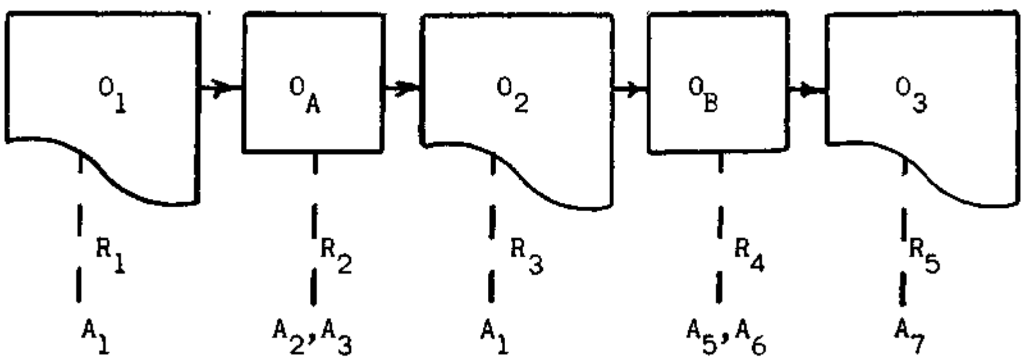


Figure 1. Nature of a System.

In Figure 1, O_A and O_B are processes, and O_1, O_2, and O_3 are the inputs and outputs of these serially-related processes. The A's are attributes associated with the objects by virtue of relationships R's. The solid lines represent the direction of flow, the dashed lines represent relationships.

The following equations can be obtained from the graph in Figure 1:

$$O_A(O_1) = O_2 \tag{1}$$

$$O_B(O_2) = O_3 \tag{2}$$

and by substitution restricted to the direction of flow (since the operations are not reversible),

$$O_B(O_A(O_1)) = O_3. \tag{3}$$

The processes of a system are either algorithmic or intuitive. An algorithmic process is a process that is performed according to instructions which are sufficiently detailed so that for a given input, the output remains constant when repeating the operation. An intuitive process is not performed according to detailed instructions with the result that for a given input, the output may vary when repeating the operation. Clearly, equations (1), (2), and (3) hold only for algorithmic processes.

When analyzing the operations given in Figure 1 with respect to success, failure,and the cause of failure, two approaches are available, depending on the nature of the process. Consider the component O_2-O_B-O_3 and assume that O_B is an algorithmic process. If O_3 is observed and found to be other than the expected outcome, it may be concluded that the cause of error is related to O_2, since this is the only possibility. Now again consider the component O_2-O_B-O_3, but assume that O_B is an intuitive process. Then, at best, the expected outcome can only be established within a range. If O_3 falls outside this range, the source of error could be related to O_2, O_B, or both. O_B could be in error because it is an act of judgement, and "bad" judgement is possible. In order to determine the source of error, the attribute values may provide hints which can be used to narrow the problem.

Just as the values of an attribute may help to determine the cause of failure, they may also give insight into the causality of success within the successful range. For example, assume that there are two objects which are adjacent in the direction of flow. If an attribute value of the first object is transferable in effect to an attribute value of the second object, then a functional relationship exists: this relationship may be of either deterministic or probabilistic causality. Consider the first attribute to be X and the second attribute to be Y. Then if X is necessary and sufficient for Y, the functional relationship is a deterministic causality relationship. If X is necessary but not sufficient for Y, then the functional relationship is one of probabilistic causality.

It is entirely possible that the values of more than one attribute are transferable in effect to the value of another attribute. Therefore, the existence of X_1 and X_2 implies that the existence of Y, and X_1 and X_2 are necessary for Y. But it is also possible that an X_3 actually exists, although not observed, and that X_3 would imply Y. Since the observed X's do not necessarily define the closed set of all X's that are necessary for Y, it cannot be concluded that the observed X's are sufficient for Y. Thus, the functional relationships between the observed X's and Y are probabilistic causality. Ackoff[4] calls these producer-product relationships: the X's are the producers of Y, and Y is the product of the X's.

Specific Definition

A specific definition of an information storage and retrieval system requires the indentification of its attributes and objects. It is also necessary to fit the attributes and objects into their component parts. The descriptions of systems contained in Nonconventional

Technical Information Systems in Current Use[5] were examined, and the following attributes were selected, based on their occurrences in a significant number of descriptions.

A. Numerical Attributes

1) Size of Collection

The number of items (documents) included in the system.

2) Rate of Growth of Collection

The annual rate of item addition.

3) Depth of Indexing

The average number of subject entries assigned per item.

4) Size of Terminology Authority

The total number of terms contained in the terminology authority.

5) Rate of Addition to the Terminology Authority

The number of new terms added per 1,000 items indexed.

6) Number of Professional Personnel

The number of persons involved in selecting the subject entries and located within the physical confines of the given system.

7) Number of Clerical Personnel

The persons involved in processing the selected terms into the file.

8) Input Processing Time

The average time necessary for the human aspect of processing one item for input and storage. This time includes both professional and clerical effort.

9) Search Time

The average time required for searching the index file. This does not include the time required for processing a request or the the time required for processing the results of a search.

10) Terms per Question

The average number of terms which are required to define the search question.

B. Non-numerical Attributes

1. Degree of Mechanization

a. Manual

b. Uniterm

c. Peek-a-boo

d. Edge-Notched Card

e. Simple Sorter

f. Collative

g. Photographic

h. Computer

2. Nature of the Contents of the Index File

a. Reference

b. Data

c. Search Aids

The general purpose of an information storage and retrieval system is to store items for future use. In order to accomplish this, processors are necessary for performing the following functions: selection, acquisition, indexing, storing, retrieving, and dissemination. A processor for performing a function consists of a person or piece of equipment that operates according to an algorithm or intuition.

In order to fit the objects and attributes into their component parts, it is necessary to combine the above mentioned functions to the degree required by the definitions of the attributes. Selection and acquisition are eliminated because none of the attributes of the sample system could be directly associated with them. Indexing and storing are combined to input processing because of the generality of the definition of input processing time. Retrieval is retained, but dissemination is eliminated because none of the attributes can be directly associated with it.

Firgure 2 represents the conceptual definition of an information storage and retrieval system as used in the context of this study. The dotted line is the system boundary. The boundary separates the system from the environment. This system has two components which are called the input processing phase and the retrieval phase. The objects and attributes are identified within the components. The relations, R's, associate the attributes with the proper objects. The attributes of the environment are associated with the entire system, that is, with each object of the system.

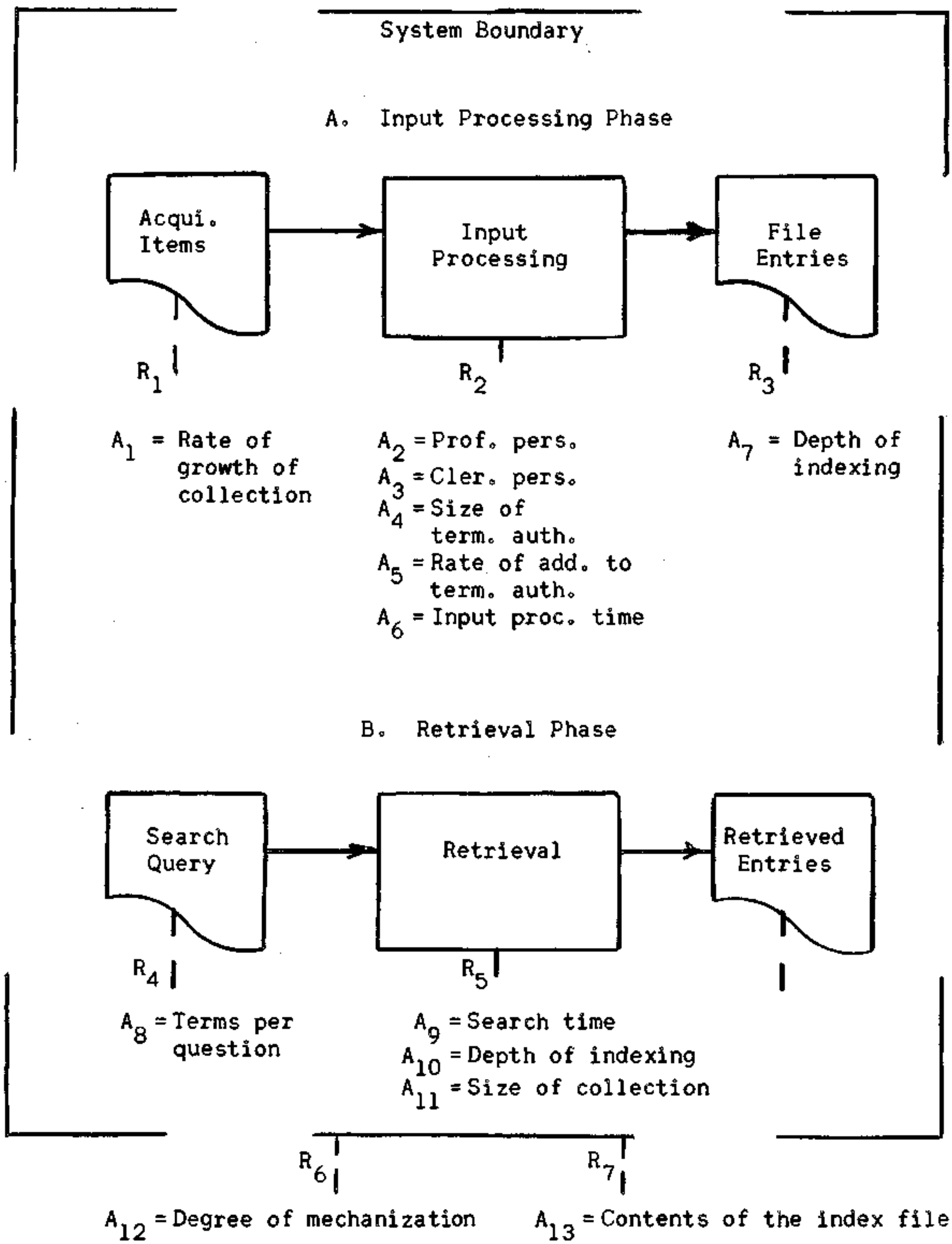


Figure 2. Information Storage and Retrieval System.

CHAPTER III

DEVELOPMENT OF THE MODEL

The model for use in this study is developed by identifying the potential producer-product relationships and determining a method for representing the values of the producer and product attributes. The representation of the attribute values must allow an insight to the causality of the product attribute behavior. This insight is obtained by determining the degree of association between the producer attribute and a selected range of the product attribute. The degree of association is used in an analytical procedure in order to determine if a functional relationship does in fact exist. If a functional relationship does exist, then it is a property of an information storage and retrieval system.

Producer-Product Relationships

The task of determining the potential producer-product relationships among the attributes is a matter of judgement supported by the condition of the system and the procedure for examining the system. The condition of the system identifies all of the attributes to be considered; these are A_1 through A_{13}, as contained in Figure 2. The procedure for examining the system is obtained by considering the magnitude of error that is inherent when calculating the parameters of the relationships, the structure of the system, and the flow direction of input processing and retrieving.

The results of the mathematical calculations will contain an experimental error, but since these calculations are based on empirical data, the degree of error is proportional to the number of variable attributes contained in a function. For example, assume that there are four variable attributes (w,x,y,z), and that the values for each attribute are obtained from a sample of systems. The objective is to determine the number of variable attributes that should be contained in the function so that the result of the calculations contains the minimum amount of experimental error. Assume that the errors for each variable attribute are: e_w, e_x, e_y, e_z. Then the following equations may be considered:

$$a = w + x + y + z \tag{4}$$

$$b = w + x \tag{5}$$

$$c = y + z \tag{6}$$

$$e_a = e_w + e_x + e_y + e_z \tag{7}$$

$$e_b = e_w + e_x \tag{8}$$

$$e_c = e_y + e_z \tag{9}$$

and by substitution

$$e_a = e_b + e_c \; . \tag{10}$$

Therefore,

$$e_a > e_b \tag{11}$$

$$e_a > e_c \tag{12}$$

Thus, equations (5) and (6) contain less experimental error than equation (4). Since the attribute values to be used in this study were established by several persons, the experimental error is assumed to be "large"; therefore, the producer-product relationships will contain only two attributes.

Each producer-product relationship will consist of two attributes in any one of the following situations:

A) an attribute of the input and an attribute of the process, or

B) an attribute of the process and an attribute of the output, or

C) two attributes of the same object.

In the first two situations, dependency is based on the direction of flow; that is, the producer attribute precedes the product attribute. In the third situation, dependency is based on the direction of flow which would exist if the object could be established as more specific in nature. The attributes, "number of professional personnel" and "number of clerical personnel" provide an example of the third situation: both are associated with input processing. However, if input processing could have been established as two functions, e.g., indexing and input processing, the number of professional personnel could have been associated with indexing, and the number of clerical personnel with input processing.

All attribute combinations were considered, and those suspected to result in nonsense correlations were eliminated. An example of one eliminated nonsense correlation is the relationship "terms per question-size of collection." The values of terms per question do not cause the size of the collection to assume its values. The following are the resulting potential producer-product relationships of the system; the producer attribute precedes the product attribute:

A) Rate of growth of collection - Professional personnel

B) Rate of growth of collection - Clerical personnel

C) Rate of growth of collection - Input processing time

D) Rate of growth of collection - Size of collection

E) Professional personnel - Clerical personnel

F) Rate of addition to terminology authority - Size of terminology authority

G) Input processing time - Depth of indexing

H) Professional personnel - Depth of indexing

I) Terms per question - Search time

J) Depth of indexing - Search time

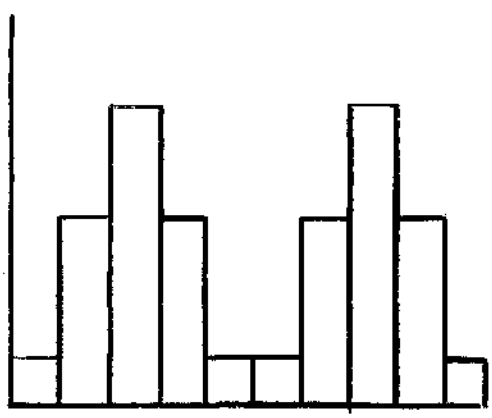
K) Size of collection - Search time

In this study, both of the two environment attributes are considered to affect all the system attributes. This approach is selected because the environment attributes assume a constant, non-numerical value over a significant operating time interval. During this time interval, the variable attributes of the system can be considered to be dependent on the environmental attributes. As an example, consider a set of systems, belonging to the sample contained in the source, that have computer as their value for the degree of mechanization. When they are compared (that is, when they are thought of as one group), the value of the degree of mechanization is constant; it is computer. The value of an attribute of the group, say the depth of indexing, is a variable as a result of the individual values that each system possessed before they were combined.

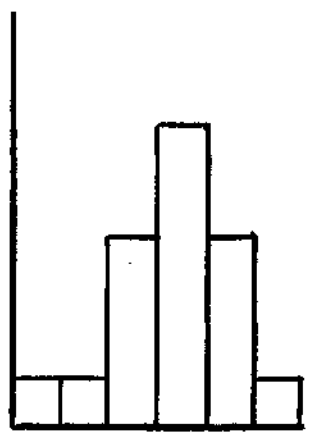
Representation of Product Attribute Values

The investigation of product attributes consists of analyzing the behavior of their values as they are affected by the producer attributes. A product attribute possesses values which occur within a range. These values vary in such a manner that they occur more frequently in some segments of the range than in others. Therefore, the range can be broken down into segments based on the frequency of occurrence and called behavior sets. Each behavior set is the result of a set of reasons which determine that the values occur within the behavior set as opposed to some other point in the range. A set of reasons is preferred (rather than one reason) because more than one reason may govern the behavior set in a similar manner.

The representation of product attribute values must be such that these behavior sets can be identified. This identification may be accomplished by observing histograms of a product attribute. Consider Figure 3,



Histogram A



Histogram B

Figure 3. Histograms and Behavior Sets.

in which the purpose of the two histograms is to aid in determining the nature of a probability density function that is a governing result of the behavior of a product attribute. The nature of the function refers to the number of nodes and not the applicable theoretical function. If the histogram implies a pathological function, as Histogram A does, it may be inferred that several distinct sets of reasons govern the behavior of the product attribute. The number of reasons is equal to the number of nodes. For Histogram A, there are two nodes and, thus, two sets of reasons which govern the behavior of product attribute A. If the result is not a pathological function, as for Histogram B, it may be inferred that there is one set of reasons that govern the behavior of the product attribute.

The histograms are constructed by calculating the frequency of occurrence of attribute values within class intervals. The magnitude of the class intervals is established by trial and error so that the selected magnitude results in the best accentuation of the behavior sets. Thus, this trial and error procedure determines those intervals which accompany the node intervals in constituting behavior sets.

The values within a behavior set may be correlated independently with the possible producer. If it is established that a correlation exists between the values of a behavior set and the corresponding values of the producer attribute, then the identification of the correlation is an explanation of a reason governing the behavior set. The recognition of this restricted correlation between the producer attribute and a behavior set of a product attribute increases the accuracy of the inferences about the properties of information storage and retrieval systems.

The reason why a particular attribute possesses a value which occurs in a given behavior set is one of the following:

A) the attribute is dependent on one or more of the producer attributes defined to exist within the sytem, or

B) the attribute is dependent on one or more of the producer attributes defined to exist within the environment, or

C) the attribute is dependent on one or more producer attributes which are not included in the scope of this study, or

D) any combination of the above.

In order to determine which of the preceding reasons is the actual case, it is necessary to investigate the first two possibilities, draw conclusions about the third, and assume that the fourth is possible.

Consider the first two possibilities. If it is established that there is a significant degree of association, base on a specified level of significance, between values belonging to a behavior set and the corresponding values of a producer attribute, then it can be inferred that the producer attribute, does, in fact, contribute to the behavior of the product attribute. Therefore, the producer attribute is a reason for the product attributes behavior within the behavior set. There is some risk involved in making this inference because a third factor may actually govern the producer and product in such a way that it just appears that the producer contributes to the behavior of the product. In this study, the existence of a third factor is unlikely.

Consider the third possibility. If an insufficient degree of association is found to exist between all of the defined dependencies, it may be concluded that attributes other than those defined within the scope of this study dominate the behavior of the product attribute.

On the other hand, consider the fourth possibility. If it is found that a significant degree of association exists between two attributes, this does not mean that the association is the only producer-product relationship. It is certainly possible that there are other producers and their identification is only possible by an actual observance of the relationship.

The degree of association is determined by calculating an estimate of the sample correlation coefficient, r, which is defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{13}$$

A test of the hypothesis that the correlation coefficient is equal to zero is given by rejecting when

$$|t| = \left| \frac{r}{\sqrt{1 - r^2}} \right| \sqrt{n - 2} \geq t_{\alpha/2;n-2} \tag{14}$$

where $t_{\alpha/2;n-2}$ is the 100 $\alpha/2$ percentage point of Student's t distribution with n-2 degrees of freedom.

If the null hypothesis; that is, the correlation coefficient is equal to zero, can be rejected, then the underlying physical relation will be calculated by the method of the least squares estimates of the slopes and the intercept. The general equation for the physical relationship is

$$y = a + b_x \tag{15}$$

The expression for the slope, b, is

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{16}$$

and the expression for the intercept, a, is

$$a = \bar{y} - b\bar{x} \tag{17}$$

The confidence interval estimate of the slope is

$$b \pm t_{\alpha/2;n-2} \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{18}$$

and the confidence interval estimate of the intercept is

$$a \pm t_{\alpha/2;n-2}\, s_{y/x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{m} (x_i - \bar{x})^2}} \tag{19}$$

where $s_{y/x}$ is an estimate of the variability about the line and is given by

$$s_{y/x} = \sqrt{\sum_{i=1}^{n} (y_i - y')} \tag{20}$$

and y' is a point estimate of the expected value of y for a given value of x^* which is given by[6]

$$y' = a + bx^* . \tag{21}$$

<u>Representation of Producer Attribute Values</u>

Producer attributes are defined to belong to the system and the environment. Those belonging to the system are numerical in nature and those belonging to the environment are non-numerical in nature. The representation of the producer attribute values depends on the procedure for determining which values of the producer and product to associate and how the degree of association is to be obtained.

Since all of the product attribute values are numerical, a procedure for determining the corresponding numerical producer attribute values does not present a problem. It consists simply of selecting the product attribute values from a behavior set and noting the systems, in the sample under investigation, from which they came. The corresponding producer attribute values can then be obtained from the same systems and the producer-product values may be associated. The method for determining the degree of association consists of calculating an estimate of the sample correlation coefficient.

On the other hand, the procedure for determining which values of the non-numerical producer attributes and the product attributes to associate, and the method for determining the degree of association do present a problem. The producer attributes of the environment are divided into mutually exclusive subsets. For example, the degree of mechanization consists of the subsets: manual, uniterm,..., computer. Every system in the sample under investigation belongs to one of these subsets. These systems also have product attributes whose values belong to the behavior sets. Therefore, the situation consists of associating the members of a behavior set with the subsets of a producer attribute. If it can be established that the members of a given subset have product attributes values which are significantly present in a behavior set when compared with members of the other subsets of the producer attribute, it may be inferred that this subset contributes to the behavior of the product attribute. Unfortunately, the term "significantly" cannot be defined in a quantitative manner, nor can an analytical method be devised to determine which are the significant subsets and which are not. This problem can be circumvented by relating the purpose for the association of a subset and a behavior set. If it can be established that the members of a subset do not have any product attribute values in a given behavior set, it may be concluded that the subset does not contribute to the behavior of that behavior set.

The procedure for determining which values of a non-numerical producer attribute and a product attribute to associate begins with the selection of a behavior set. Each value contained in this behavior set belongs to a different system in a sample of systems. Each of these systems possesses one value for a given producer attribute and this value is the name of a subset of the producer attribute. Therefore, the systems can be examined to determine the number of times each subset occurs as the value for the producer attribute. A display of some hypothetical results of this procedure is illustrated in Figure 4. Figure 4 is an illustration of one product attribute with two behavior sets, and one

Product Attribute			
Behavior Set 1	Behavior Set 2		
A_1	B_1	Subset 1	Producer
A_2	B_2	Subset 2	Attribute

Figure 4. Subset Manifestation.

producer attribute with two subsets. A_1 is the number of times that the systems of a sample possess values that occur in behavior set 1 and subset 1; and likewise for A_2, B_1, and B_2. The addition of A_1 and B_1, $(A_1 + B_1)$, results in the total number of times that subset 1 is manifested in the behavior sets of the product attribute of systems. Therefore, $A_1/(A_1 + B_1)$ is the ratio of the number of times that subset 1 is manifested in behavior set 1. If A_1 is zero, it may be concluded that subset 1 does not contribute to the behavior of the product attribute within behavior set 1.

CHAPTER IV

EXPERIMENTAL APPLICATION OF THE MODEL

Establishment of Behavior Sets

The values for all of the attributes were gathered from the system descriptions contained in the source, Nonconventional Technical Information Systems in Current Use. These values are presented in Appendix A, Data Tabulation, pp. 39 - 45. Each entry consists of the system identification number, as contained in the source, followed by the extracted data. The symbol "unk." means that the particular item of data is either not available or is too vague to be included in the sample. A vague item of data is one that is given in the form of a range, such as 3 - 50. The mean of this range can certainly be calculated but it would be different than the mode. The mode is the important value to select since it represents the most probable value that an attribute of a system will possess at any given time. For small ranges, such as 3 - 7, it is assumed that the mode and the mean are equal. Appendix A contains the sample population values, and three iterations of the identification numbers are required to present the data from the 87 system descriptions.

Frequency tables were tabulated for each of the product attributes, and are presented in Appendix B, Frequency Tabulations, pp. 47-51. Each table consists of an enumeration of the class intervals followed by the frequency of occurrence of sample values within that interval, and then the occurrence ratio for each interval. The sample values which were not included in the frequency calculations are individually listed in the fourth column. It was necessary to eliminate these values in order to prohibit the occurrence of several empty class intervals within a histogram. The actual "cut-off" value was established by observing the magnitude between each adjacent pair of values in numerical sequence. The larger value of an adjacent pair, and all those values beyond, were eliminated when the magnitude was observed to increase sharply.

The results of the frequency calculations were used to construct histograms, one for each of the product attributes. The histograms are presented in Appendix C, Histograms, pp. 53-62. These bar graphs are a plot of the frequency of occurrence of sample members within a class interval against that class interval.

The structures of the histograms are observed and the behavior sets are established. This procedure is a matter of judgement whereby the nodes are identified and the adjacent intervals are included on their probable membership in the behavior set under construction, as opposed to an adjacent behavior set. No overlap is allowed, that is; a class interval can only belong to one behavior set. In the histogram for search time, class interval eight is not included in any of the behavior sets. In this case, it is judged that is is not close to either of the behavior sets, nor does it have sufficient members to be a behavior set of its own.

All of the producer-product relationships are now re-established on the basis of the behavior sets. The producer attribute is related to each behavior set of the product attribute. These relationships are contained in Appendix D, Numerical Attribute Associations, pp. 64-65, which establishes all of the correlations that are made between the numerical attributes. For example, the sample population values for professional personnel are correlated with the corresponding values of the rate of growth of collection. A correlation is made for each behavior set of the product attribute. For professional personnel, behavior set one contains 64 members within the range of one to six people.

It is not necessary to include the relationships between the producer attributes of the environment and the behavior sets of the product attributes in Appendix D. This information can be obtained by observing the histograms in Appendix C.

Each system description contained in the source was examined. If the system description contained a value belonging to a class interval and to a subset, a tally was made. These tallies are listed beneath the abscissa of the histograms. For example, consider the histograms of the rate of growth of collection in Appendix C. The appearance of the number "1" in the first row beneath class interval 4 means that one system with manual as the value for degree of mechanization also has a value for the annual rate of growth of the collection which is between 3,500 and 6,000 items.

Scatter-grams were also plotted for each of the numerical producer-product relationships. These are contained in Appendix E, Scatter-Grams, pp. 67-77. The product attribute is contained on the abscissa and the producer attribute is contained on the ordinate. Points that are not plotted as a result of the scale are listed below the graphs. The ranges of the behavior sets are noted at the top of each graph. The purpose of the scatter-grams is to give an indication of the existence of a correlation.

Calculations

The correlation coefficients, calculated according to equation (13), are used to calculate a "t" test value according to equation (14). A table of percentage points of the t distribution is entered using a level of significance of 10 per cent.[7] If the "t" test value is greater than the value provided by the table, the hypothesis that the correlation coefficient is equal to zero is rejected. This procedure is performed on each of the associations contained in Appendix D. The calculated coefficients for all of the relationships are contained in Appendix F, Results, pp. 79-88. Those relationships whose correlation coefficients are proved to be not equal to zero are also contained in Table 1, Results of Numerical Attribute Associations, p. 31.

The physical relationships are calculated for those relationships identified in Appendix 4 whose correlation coefficients are not equal to zero. This is accomplished according to equation (15) with substitutions from equations (16), (17), (18), (19), (20), and (21). A level of significance of 10 per cent is used in equations (18), and (19). The results are presented in Table 3, Regression Equations, pp. 34-35.

The ratio of the number of times that a subset is manifested in a behavior set was calculated for each subset and behavior set. These ratios are also presented in Appendix F.

CHAPTER V

RESULTS

Presentation of Results

The results of the calculations can be grouped into five categories. These categories are:

1) the correlation coefficients,

2) the correlation coefficients which are not equal to zero,

3) the regression equations between the attributes where the correlation coefficient is not equal to zero.

4) the occurrence ratios, and

5) the occurrence ratios which are equal to zero.

Categories one and four contain the general results. Categories two, three, and five contain the results from which inferences are made about the properties of an information storage and retrieval system.

The results of categories one and four are contained in Appendix F, containing ten tables, one for each of the product attributes. The subdivisions of a table consist of all the producer attributes which are intuitively believed to contribute to the behavior of the product attribute. The columns of data are the degrees of association between the producer attributes and the behavior sets of the product attributes. The degree of association is the correlation coefficient for the numerical producer attributes, and the occurrence ratio for the non-numerical producer attributes. Of the 26 correlation coefficients contained in Appendix

F, seven were proved to be not equal to zero. Of the 253 occurrence ratios contained in Appendix F, 33 were observed to be equal to zero. These seven correlation coefficients constitute category two and the 33 occurrence ratios constitute category five.

The results of category two are contained in Table 1, Results of Numerical Attribute Associations. Table 1 has two levels of subdivision. The first level, with alphabetical notation, consists of the numerical producer attributes. The second level, with numeric notation, consists of those product attributes with which a probabilistic causality relationship exists. These relationships only hold for the range of the product attribute in the column of data.

The results of category five are contained in Table 2, Results of Non-Numerical Attribute Associations, having two levels of subdivision. The first level, with alphabetic notation, consists of the non-numerical producer attributes. The second level, with numeric notation, consists of those product attributes with which an association does not exist.

The results of category three are contained in Table 3, Regression Equations, consisting of seven linear regression equations. The producer attributes are identified and denoted by "x." The product attributes are also identified and denoted by "y." The numbers in parentheses are the confidence intervals based on a level of significance of 10 per cent. Below each equation is the range of the product attribute for which the equation holds.

Table 1. Results of Numerical Attribute Associations

	Non-Zero Correlation Range
A. Rate of growth of collection	
1. Clerical personnel	1 - 9 persons
2. Size of collection	1 - 81,000 items
B. Professional personnel	
1. Clerical personnel	1 - 9 persons
C. Terms per question	
1. Search time	1 - 6 and 27 - 45 minutes
D. Size of collection	
1. Search time	1 - 6 minutes

Table 2. Results of Non-Numerical Associations

	Zero Correlation Range
A. Manual	
1. Rate of growth of collection	9,000 - 12,000 items
2. Input processing time	1 - 30 minutes
3. Size of collection	45,000 - 81,000 items
4. Rate of addition to terminology authority	175 - 300 terms
5. Depth of indexing	24 - 56 terms
6. Search time	27 - 45 minutes

(Continued)

Table 2. (Continued)

	Zero Correlation Range
B. Uniterm	
1. Rate of growth of collection	13,500 - 21,000 items
2. Professional personnel	7 - 12 persons
3. Input processing time	30 - 65 minutes
4. Depth of indexing	24 - 56 terms
5. Search time	1 - 6 minutes
C. Peek-a-boo	
1. Rate of growth of collection	13,500 - 21,000 items
2. Professional personnel	7 - 12 persons
3. Input processing time	30 - 65 minutes
4. Rate of addition to terminology authority	175 - 300 terms
D. Edge-Notched Card	
1. Rate of growth of collection	9,000 - 21,000 items
2. Professional personnel	7 - 12 persons
3. Input processing time	1 - 50 minutes
4. Rate of addition to terminology authority	175 - 300 terms
5. Terms per question	1 - 5 terms
E. Simple sorter	
1. Professional personnel	7 - 12 persons
2. Depth of indexing	24 - 40 terms

(Continued)

Table 2. (Continued)

	Zero Correlation Range
F. Collative	
1. Rate of growth of collection	9,000 - 21,000 items
2. Size of collection	45,000 - 81,000 items
3. Depth of indexing	44 - 56 terms
4. Search time	1 - 6 minutes
G. Photographic	
1. Rate of growth of collection	1 - 9,000 items
2. Professional personnel	7 - 12 persons
3. Input processing time	30 - 65 minutes
4. Size of collection	45,000 - 81,000 items
5. Size of terminology authority	2,000 - 8,500 terms
6. Rate of addition to terminology authority	175 - 300 terms
7. Depth of indexing	24 - 56 terms
8. Search time	6 - 45 minutes
9. Terms per question	5 - 15 terms
H. Computer	
1. Input processing time	50 - 65 minutes
I. References (represented in all sets of all dependent attributes)	
J. Data	
1. Input processing time	30 - 50 minutes
2. Rate of addition to terminology authority	175 - 300 terms
K. Search aids	
1. Rate of addition to terminology authority	175 - 300 terms

Table 3. Regression Equations

1. x = rate of growth of collection

y = clerical personnel

$$y = 1.71\ (\pm .74) + .0001\ (\pm .00008)\ x \qquad (22)$$

for $y \leq 9$

2. x = rate of growth of collection

y = size of collection

$$y = 8218\ (\pm 44440) + 1.46\ (\pm .91)\ x \qquad (23)$$

for $y \leq 45{,}000$

3. x = rate of growth of collection

y = size of collection

$$y = 52{,}618\ (\pm 12{,}500) + .84\ (\pm .88)\ x \qquad (24)$$

for $45{,}000 < y \leq 81{,}000$

4. x = professional personnel

y = clerical personnel

$$y = 1.34\ (\pm .89) + .27\ (\pm .18)\ x \qquad (25)$$

for $y \leq 9$

5. x = terms per question

y = search time

$$y = 2.22\ (\pm 2.54) + .40\ (\pm .55)\ x \qquad (26)$$

for $y \leq 6$

6. x = terms per question

y = search time

$$y = 20.1\ (\pm 26.8) + 4.41\ (\pm 6.95)\ x \qquad (27)$$

for $27 \leq y \leq 45$

(Continued)

Table 3. (Continued)

7. x = size of collection

y = search time

$$y = 4.43\ (\pm\ 1.14) - .0000043\ (\pm\ .0000061)\ x \qquad (28)$$

for $y \leq 6$

Discussion of Results

The method used to obtain these results is a technique for identifying the probabilistic causality relationships which exist between the attributes of information storage and retrieval systems. Once these relationships are identified, the physical relationships between the attributes can be established.

Again, these relationships indicate what is currently being practiced; they are not criteria for determining retrieval or economic efficiency. The relationships are useful to the designers of information storage and retrieval systems because they may be considered as an "input" to systems design. During the course of analyzing a problem, some attribute values may be established. When these values are established before the design is formulated, they become conditions of the given system. Then, it is of interest to obtain some knowledge about the probable values which the remaining attributes will assume under these specified conditions.

The results of this study enable the system designer to obtain this knowledge if one or more of the producer attributes contained in Tables 1 and 2 are pre-established and, therefore, are considered as conditions. For example, assume that for a given system design problem, the value for the rate of growth of the collection is pre-established during the analysis phase, and it becomes a condition of the new system. By entering Table 1, it is seen that the attributes of clerical personnel and size of the collection form probabilistic causality relationships with the producer attribute, rate of growth of the collection. The regression equations for these relationships are contained in Table 3, equations (22) and (23). By substituting the specified value for the rate of growth of the collection into equations (22) and (23), values for the number of clerical personnel and size of collection can be estimated. These values constitute a portion of the probable state that the system will assume.

As another example, assume that the value for the degree of mechanization is pre-established, during the analysis of a problem, as manual, and it becomes a condition for the new system. By entering Table 2, it is seen that the product attributes: rate of growth of collection, input processing time, size of collection, rate of terminology authority, depth of indexing, and search time are related to the producer attribute value of manual. These relations are such that the current practices indicate that the product attributes will not possess values within the ranges listed beside each product attribute. Therefore, the condition of manual implies that the new system will not possess attribute values within the ranges listed.

While the latter example does not specify the probable state that a system will assume, it does specify a state that a system will probably not assume. It is believed that the information provided by this example is just as valuable to the designers of information storage and retrieval systems as that information provided by the former example.

Both types of results, while requiring different approaches, do provide the designers with an a priori knowledge about the probable state that a system will assume when certain conditions are specified. Thus, the value of experience is increased because the method developed in this study enables a representation of the knowledge gained by past experience.

APPENDIX A

The following pages contain the data that was extracted from the system descriptions contained in the source.

Table 4. Data Tabulation[8]

System Number	Size of Collection	Rate of Growth	Term. Auth.	Add. to Term. Auth.	Depth of Index.
1.1.1	41,000	6,000	4,000	100	6
1.1.2	400,000	20,000	2,000	5	4
1.2.1	10,500	350	2,700	unk.	11
1.2.2	8,100	500	2,500	75	10
1.2.3	35,000	5,000	5,250	6	8
1.2.4	8,500	600	3,750	unk.	12
1.2.5	65,000	7,500	11,740	200	7
1.2.6	61,260	10,000	7,900	20	5
1.2.7	6,000	650	2,000	200	unk.
1.3.1	7,500	2,225	1,800	unk.	10
1.3.2	3,600	1,450	1,200	9	27
1.3.3	20,000	5,000	1,500	28	12
1.3.4	1,000	3,000	150	10	5
1.3.5	10,000	3,000	1,200	10	8
1.3.6	18,000	8,000	760	unk.	15
1.3.7	2,200	1,300	1,000	unk.	85
1.4.1	50,000	5,000	2,500	30	9
1.4.2	25,000	2,000	1,000	500	10
1.4.3	44,000	3,500	483	unk.	3
1.5.1	12,000	3,500	700	6	10
1.5.2	57,000	6,000	700	unk.	unk.
1.5.3	unk.	unk.	2,000	250	9
1.5.4	9,000	500	unk.	unk.	unk.
1.5.5	20,000	2,700	unk.	unk.	5
1.5.6	8,500	1,500	1,000	50	2
1.5.7	75,000	9,000	1,737	50	unk.
1.5.8	4,000	250	1,600	unk.	22
1.5.9	7,500	2,500	700	1	8
1.5.10	30,000	2,000	60,000	1,500	20
1.5.11	59,000	12,000	unk.	unk.	unk.
1.5.12	14,500	825	600	unk.	50
1.5.13	1,000	200	91	25	10
1.5.14	5,500	1,100	unk.	unk.	unk.
1.6.1	14,000	7,500	3,045	75	40
1.6.2	6,900	720	3,800	10	14
1.6.3	10,000	4,000	6,500	300	20
1.6.4	2,400	350	3,900	unk.	25
1.6.5	10,000	1,350	8,000	275	30
1.6.6	5,000	2,500	14,000	1,400	13.5
1.6.7	6,800	1,100	11,000	45	120
1.6.8	13,400	3,600	1,800	3	13

(Continued)

Table 4. (Continued)

System Number	Size of Collection	Rate of Growth	Term. Auth.	Add. to Term. Auth.	Depth of Index.
1.6.9	35,000	850	2,000	5	12
1.6.10	15,000	3,000	1,600	2	8
1.6.11	140,000	6,000	7,000	unk.	22
1.6.12	40,260	1,000	38,000	300	unk.
1.7.1	2,000	12,000	350	unk.	9
1.7.2	300,000	20,000	40,000	100	5
1.8.1	275,000	25,000	7,000	2	12
1.8.2	18,500	2,225	18,500	800	12
1.8.3	7,500	1,000	28,000	2,000	66
1.8.4	60,000	12,000	7,000	30	18
1.8.5	275,000	50,000	unk.	unk.	unk.
1.8.6	8,000	10,400	6,000	50	12
1.8.7	80,000	10,000	16,500	150	8
1.8.8	2,200	250	unk.	unk.	400
1.8.9	30,000	unk.	unk.	unk.	unk.
1.8.10	80,000	36,000	38,000	180	55
2.1.1	640,000	225,000	200	unk.	unk.
2.1.2	47,000	4,500	275	unk.	4
2.1.3	5,000	2,500	135	unk.	unk.
2.1.4	30,000	unk.	unk.	unk.	unk.
2.1.5	60,000	12,000	800	10	10
2.1.6	49,000	5,000	700	unk.	5
2.1.7	9,000	1,000	unk.	unk.	unk.
2.1.8	5,000	165	30	unk.	30
2.1.9	4,636	1	13	unk.	13
2.1.10	15,000	1,375	1,600	50	unk.
2.1.11	6,500	600	unk.	unk.	unk.
2.2.1	3,000	500	3,500	50	20
2.2.2	235,000	unk.	7,000	unk.	10
2.3.1	60,000	15,000	1,200	6	unk.
2.3.2	4,000	1,600	830	3	75
2.3.3	6,000	1,000	5,000	20	50
2.3.4	40,000	18,000	unk.	unk.	unk.
3.1.1	4,500	1,900	unk.	unk.	6
3.1.2	700	60	1,000	unk.	1
3.1.3	100,000	11,000	7,800	unk.	35
3.1.4	36,000	18,000	8,230	unk.	8
3.1.5	4,600	700	unk.	unk.	5.7
3.1.6	125,000	1,500	5,500	1	2
3.1.7	27,000	11,400	unk.	unk.	14
3.1.8	1,300	100	300	30	3

(Continued)

Table 4. (Continued)

System Number	Size of Collection	Rate of Growth	Term. Auth.	Add. to Term. Auth.	Depth of Index.
3.1.9	26,000	unk.	15,000	unk.	9
3.2.1	1,010	40	unk.	unk.	30
3.2.2	10,000	1,600	600	13	30
3.2.3	450	unk.	208	unk.	55
3.2.4	71,800	20,000	1,556	unk.	45

System Number	Terms per Question	Search Time	Prof. Pers.	Cler. Pers.	Input Proc. Time
1.1.1	unk.	15	12	7	60
1.1.2	unk.	1	6	9	unk.
1.2.1	3.5	30	2	1	15
1.2.2	2.5	10	1	1	25
1.2.3	4.5	45	2	2	12
1.2.4	4	90	1	1	30
1.2.5	3	unk.	2	2	20
1.2.6	2.5	unk.	1	1	unk.
1.2.7	3	unk.	1	1	15
1.3.1	3.5	15	3	1	26
1.3.2	4	35	1	1	30
1.3.3	7.5	3	1	unk.	15
1.3.4	3	3	1	1	5
1.3.5	4	5	1	1	10
1.3.6	4.5	unk.	3.5	unk.	8
1.3.7	10	unk.	3	2	30
1.4.1	5.5	180	18.5	1	300
1.4.2	unk.	60	unk.	unk.	180
1.4.3	unk.	unk.	2	1	unk.
1.5.1	3.5	120	1	1	4
1.5.2	4.5	360	2	1	3
1.5.3	3	15	unk.	unk.	18
1.5.4	unk.	unk.	1.5	2	unk.
1.5.5	unk.	15	unk.	unk.	4
1.5.6	unk.	30	4	4	unk.
1.5.7	unk.	unk.	5	4	unk.
1.5.8	3	unk.	1	unk.	45
1.5.9	3.5	unk.	2	1	10
1.5.10	1.5	30	1	1	22

(Continued)

Table 4. (Continued)

System Number	Terms per Question	Search Time	Prof. Pers.	Cler. Pers.	Input Proc. Time
1.5.11	3	unk.	2	unk.	unk.
1.5.12	7	5	0.5	unk.	60
1.5.13	2	3	1	1	unk.
1.5.14	5	23	2	unk.	22
1.6.1	4	33	4	3	64
1.6.2	1	unk.	1	1	6
1.6.3	4.5	18	9	2	54
1.6.4	unk.	unk.	4	2	unk.
1.6.5	10	90	11	5	45
1.6.6	14	8	2	unk.	120
1.6.7	8	unk.	5	1.5	30
1.6.8	5	unk.	2	1	unk.
1.6.9	3	unk.	3	1	30
1.6.10	5	45	3	2	unk.
1.6.11	15	480	6	2	45
1.6.12	1.5	90	1	3	35
1.7.1	5	5	2	1	20
1.7.2	2	2	5	3	unk.
1.8.1	12.5	18	unk.	unk.	120
1.8.2	4	unk.	6	4	120
1.8.3	5	unk.	3	1	42
1.8.4	2.5	12	2	unk.	22
1.8.5	5	6	12	6	unk.
1.8.6	5	10	6	1	20
1.8.7	4	15	3	2	50
1.8.8	4	4	6	unk.	240
1.8.9	unk.	5	unk.	8	unk.
1.8.10	unk.	unk.	11.1	11.9	25
2.1.1	2	2	1	20	30
2.1.2	unk.	unk.	4	unk.	2
2.1.3	unk.	5	1	1	3
2.1.4	unk.	30	unk.	1	5
2.1.5	4	18	2	3	1.2
2.1.6	unk.	unk.	4	5	60
2.1.7	unk.	unk.	3	3	unk.
2.1.8	2	unk.	3	unk.	5
2.1.9	unk.	unk.	unk.	.25	12
2.1.10	4	40	4	3	5
2.1.11	3.5	18	2	unk.	9
2.2.1	4	unk.	4	unk.	30
2.2.2	unk.	unk.	10	60	unk.

(Continued)

Table 4. (Continued)

System Number	Terms per Questions	Search Time	Prof. Pers.	Cler. Pers.	Input Proc. Time
2.3.1	5	unk.	unk.	unk.	unk.
2.3.2	5	5	1	unk.	15
2.3.3	7	5	1	unk.	unk.
2.3.4	5	unk.	3	4	unk.
3.1.1	unk.	unk.	unk.	unk.	5
3.1.2	2	23	unk.	6	60
3.1.3	unk.	unk.	6	5	35
3.1.4	unk.	unk.	6	6	38
3.1.5	unk.	unk.	unk.	unk.	5
3.1.6	unk.	unk.	20	16	38
3.1.7	unk.	unk.	8	unk.	unk.
3.1.8	unk.	unk.	2	1	120
3.1.9	unk.	unk.	unk.	unk.	unk.
3.2.1	unk.	unk.	.25	unk.	720
3.2.2	unk.	unk.	3	1	90
3.2.3	unk.	37	2	11	60
3.2.4	7	unk.	unk.	unk.	5

System Number	Degree of Mechanization	Contents of the Index File
1.1.1	Manual	References
1.1.2	Manual	References
1.2.1	Uniterm	References
1.2.2	Uniterm	References
1.2.3	Uniterm	References
1.2.4	Uniterm	References
1.2.5	Uniterm	References
1.2.6	Uniterm	References
1.2.7	Uniterm	References
1.3.1	Peek-a-boo	References
1.3.2	Peek-a-boo	References
1.3.3	Peek-a-boo	References
1.3.4	Peek-a-boo	References
1.3.5	Peek-a-boo	References
1.3.6	Peek-a-boo	References

(Continued)

Table 4. (Continued)

System Number	Degree of Mechanization	Contents of the Index File
1.3.7	Peek-a-boo	References
1.4.1	Edge-Notched Card	References
1.4.2	Edge-Notched Card	References
1.4.3	Edge-Notched Card	References
1.5.1	Simple Sorter	References
1.5.2	Simple Sorter	References
1.5.3	Simple Sorter	References
1.5.4	Simple Sorter	References
1.5.5	Simple Sorter	References
1.5.6	Simple Sorter	References
1.5.7	Simple Sorter	References
1.5.8	Simple Sorter	References
1.5.9	Simple Sorter	References
1.5.10	Simple Sorter	References
1.5.11	Simple Sorter	References
1.5.12	Simple Sorter	References
1.5.13	Simple Sorter	References
1.5.14	Simple Sorter	References
1.6.1	Collative	References
1.6.2	Collative	References
1.6.3	Collative	References
1.6.4	Collative	References
1.6.5	Collative	References
1.6.6	Collative	References
1.6.7	Collative	References
1.6.8	Collative	References
1.6.9	Collative	References
1.6.10	Collative	References
1.6.11	Collative	References
1.6.12	Collative	References
1.7.1	Photographic	References
1.7.2	Photographic	References
1.7.2	Photographic	References
1.8.1	Computer	References
1.8.2	Computer	References
1.8.3	Computer	References
1.8.4	Computer	References
1.8.5	Computer	References
1.8.6	Computer	References
1.8.7	Computer	References
1.8.8	Computer	References

(Continued)

Table 4. (Continued)

System Number	Degree of Mechanization	Contents of the Index File
1.8.9	Computer	References
1.8.10	Computer	References
2.1.1	Computer	Data
2.1.2	Simple Sorter	Data
2.1.3	Simple Sorter	Data
2.1.4	Computer	Data
2.1.5	Peek-a-boo	Data
2.1.6	Simple Sorter	Data
2.1.7	Computer	Data
2.1.8	Computer	Data
2.1.9	Computer	Data
2.1.10	Simple Sorter	Data
2.1.11	Simple Sorter	Data
2.2.1	Computer	Data
2.2.2	Computer	Data
2.3.1	Simple Sorter	Data
2.3.2	Peek-a-boo	Data
2.3.3	Peek-a-boo	Data
2.3.4	Computer	Data
3.1.1	Computer	Search Aids
3.1.2	Simple Sorter	Search Aids
3.1.3	Computer	Search Aids
3.1.4	Manual	Search Aids
3.1.5	Computer	Search Aids
3.1.6	Simple Sorter	Search Aids
3.1.7	Computer	Search Aids
3.1.8	Computer	Search Aids
3.1.9	Simple Sorter	Search Aids
3.2.1	Edge-Notched Card	Search Aids
3.2.2	Edge-Notched Card	Search Aids
3.2.3	Edge-Notched Card	Search Aids
3.2.4	Simple Sorter	Search Aids

APPENDIX B

The following pages contain the frequency calculations for the product attributes.

Table 5. Frequency Tabulations

	Class Interval	Frequency	Occurrence Ratio	Members Not Included
A.	Size of collection			
1)	1- 4,500	15	.195	100,000
2)	4,501- 9,000	20	.260	125,000
3)	9,001-13,500	7	.090	140,000
4)	13,501-18,000	5	.065	235,000
5)	18,001-22,500	3	.039	275,000
6)	22,501-27,000	3	.039	275,000
7)	27,001-31,500	3	.039	300,000
8)	31,501-36,000	3	.039	400,000
9)	36,001-40,500	2	.026	640,000
10)	40,501-45,000	2	.026	
11)	45,001-49,500	2	.026	
12)	49,501-54,000	1	.013	
13)	54,001-58,500	1	.013	
14)	58,501-63,000	5	.065	
15)	63,001-67,500	1	.013	
16)	67,501-72,000	1	.013	
17)	72,001-76,500	1	.013	
18)	76,501-81,000	2	.026	
		77	1.000	
B.	Rate of growth of collection			
1)	1- 1,500	32	.415	25,000
2)	1,501- 3,000	14	.182	36,000
3)	3,001- 4,500	5	.065	50,000
4)	4,501- 6,000	7	.091	225,000
5)	6,001- 7,500	2	.026	
6)	7,501- 9,000	2	.026	
7)	9,001-10,500	3	.039	
8)	10,501-12,000	6	.078	
9)	12,001-13,500	0	.000	
10)	13,501-15,000	1	.013	
11)	15,001-16,500	0	.000	
12)	16,501-18,000	2	.026	
13)	18,001-19,500	0	.000	
14)	19,501-21,000	3	.039	
		77	1.000	

(Continued)

Table 5. (Continued)

Class Interval	Frequency	Occurrence Ratio	Members not Included
C. Depth of Indexing			
1) 1- 4	7	.108	66
2) 5- 8	15	.231	75
3) 9-12	18	.277	85
4) 13-16	6	.091	120
5) 17-20	4	.062	400
6) 21-24	2	.031	
7) 25-28	2	.031	
8) 29-32	4	.062	
9) 33-36	1	.015	
10) 37-40	1	.015	
11) 41-44	0	.000	
12) 45-48	1	.015	
13) 49-52	2	.031	
14) 53-56	2	.031	
	65	1.000	
D. Size of Terminology Authority			
1) 1- 500	11	.180	11,000
2) 501-1,000	13	.213	11,740
3) 1,001-1,500	4	.066	14,000
4) 1,501-2,000	11	.180	15,000
5) 2,001-2,500	2	.033	16,500
6) 2,501-3,000	1	.016	18,500
7) 3,001-3,500	2	.033	28,000
8) 3,501-4,000	4	.066	38,000
9) 4,001-4,500	0	.000	38,000
10) 4,501-5,000	1	.016	40,000
11) 5,001-5,500	2	.033	60,000
12) 5,501-6,000	1	.016	
13) 6,001-6,500	1	.016	
14) 6,501-7,000	4	.066	
15) 7,001-7,500	0	.000	
16) 7,501-8,000	3	.049	
17) 8,001-8,500	1	.016	
	61	1.000	

(Continued)

Table 5. (Continued)

	Class Interval	Frequency	Occurrence Ratio	Members Not Included
E.	Rate of Addition to Terminology Authority			
1)	1- 25	20	.476	500
2)	26- 50	10	.239	800
3)	51- 75	2	.047	1,400
4)	76-100	2	.047	1,500
5)	101-125	0	.000	2,000
6)	126 150	1	.024	
7)	151-175	0	.000	
8)	176-200	3	.072	
9)	201-225	0	.000	
10)	226-250	1	.024	
11)	251-275	1	.024	
12)	276-300	2	.047	
		42	1.000	
F.	Professional Personnel			
1)	.1- 1	20	.278	18.5
2)	1.1- 2	17	.236	20
3)	2.1- 3	10	.139	
4)	3.1- 4	8	.111	
5)	4.1- 5	3	.042	
6)	5.1- 6	7	.096	
7)	6.1- 7	0	.000	
8)	7.1- 8	1	.014	
9)	8.1- 9	1	.014	
10)	9.1-10	1	.014	
11)	10.1-11	1	.014	
12)	11.1-12	3	.042	
		72	1.000	
G.	Clerical Personnel			
1)	.1-1	26	.472	11
2)	1.1-2	10	.182	11.9
3)	2.1-3	6	.108	16
4)	3.1-4	4	.073	20
5)	4.1-5	3	.055	60
6)	5.1-6	3	.055	
7)	6.1-7	1	.018	
8)	7.1-8	1	.018	
9)	8.1-9	1	.018	
		55	1.000	

(Continued)

Table 5. (Continued)

	Class Interval	Frequency	Occurrence Ratio	Members Not Included
H. Input Processing Time				
1)	1- 5	13	.228	90
2)	6-10	5	.088	120
3)	11-15	6	.105	120
4)	16-20	4	.070	120
5)	21-25	5	.088	120
6)	26-30	8	.140	180
7)	31-35	2	.035	240
8)	36-40	2	.035	300
9)	41-45	4	.070	720
10)	46-50	1	.018	
11)	51-55	1	.018	
12)	56-60	5	.088	
13)	61-65	1	.018	
		57	1.000	
I. Search Time				
1)	1- 3	6	.150	60
2)	4- 6	9	.225	90
3)	7- 9	1	.025	90
4)	10-12	3	.075	90
5)	13-15	5	.125	120
6)	16-18	4	.100	180
7)	19-21	0	.000	360
8)	22-24	2	.050	480
9)	25-27	0	.000	
10)	28-30	4	.100	
11)	31-33	1	.025	
12)	34-36	1	.025	
13)	37-39	1	.025	
14)	40-42	1	.025	
15)	43-45	2	.050	
		40	1.000	

(Continued)

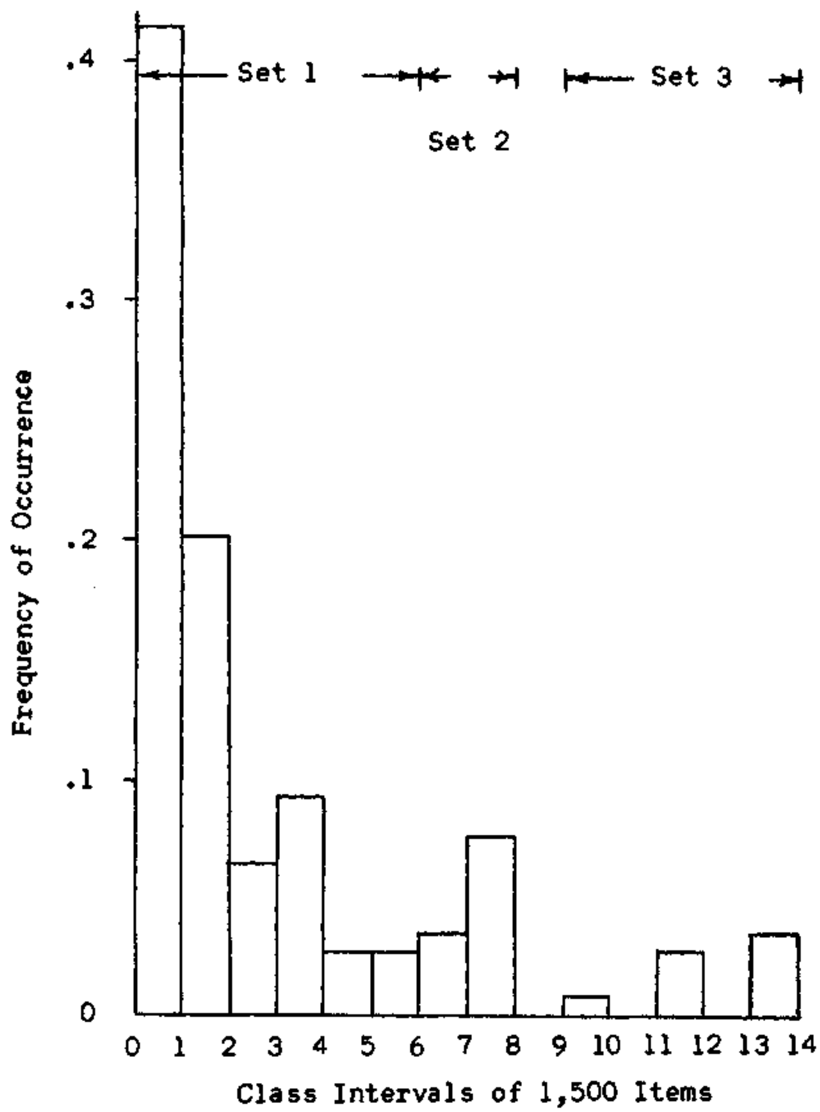
Table 5. (Continued)

	Class Interval	Frequency	Occurrence Ratio	Members Not Included
J.	Terms per Question			
1)	.1- 1	1	.017	
2)	1.1- 2	7	.121	all members
3)	2.1- 3	10	.172	included
4)	3.1- 4	15	.258	
5)	4.1- 5	14	.242	
6)	5.1- 6	1	.017	
7)	6.1- 7	3	.052	
8)	7.1- 8	2	.035	
9)	8.1- 9	0	.000	
10)	9.1-10	2	.035	
11)	10.1-11	0	.000	
12)	11.1-12	0	.000	
13)	12.1-13	1	.017	
14)	13.1-14	1	.017	
15)	14.1-15	1	.017	
		58	1.000	

APPENDIX C

The following pages contain the histograms of the product attributes.

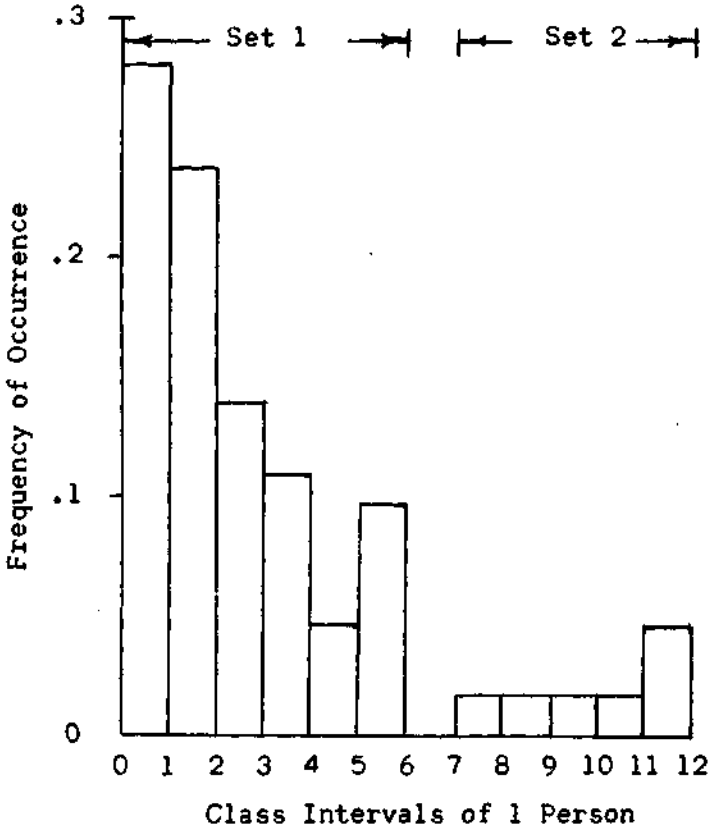
A. Rate of Growth of Collection



			1								1		1	Manual
4			1	1		1								Uniterm
3	4		1		1		1							Peek-a-boo
1	2	1	1											Edge-notched Card
10	4	2	2		1		1		1				1	Simple Sorter
6	2	2	1	1										Collative
							1						1	Photographic
8	2					2	3				1			Computer
20	10	4	6	2	2	3	3						2	References
7	2	1	1				1		1		1			Data
5	2						2				1		1	Search Aids

Figure 5. Histograms.

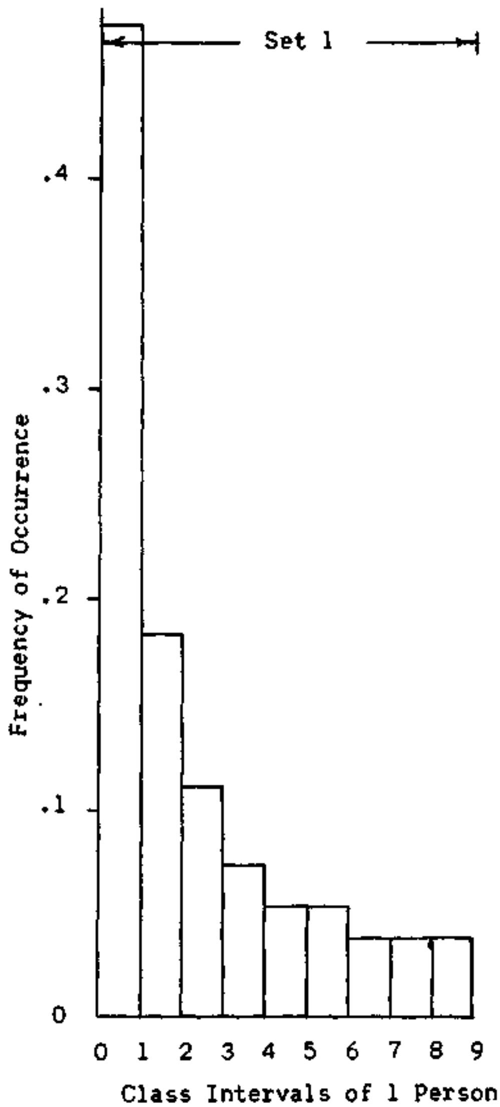
B. Professional Personnel



					2						1	Manual
4	3											Uniterm
6	1	2	1									Peek-a-boo
1	2	1										Edge-notched Card
6	5		4	1								Simple Sorter
2	3	2	2	1	1			1		1		Collative
	1			1								Photographic
1	2	5	1		4		1		1		2	Computer
15	13	6	4	3	5			1		1	3	References
4	2	3	4					1				Data
1	2	1			2		1					Search Aids

Figure 5. (Continued)

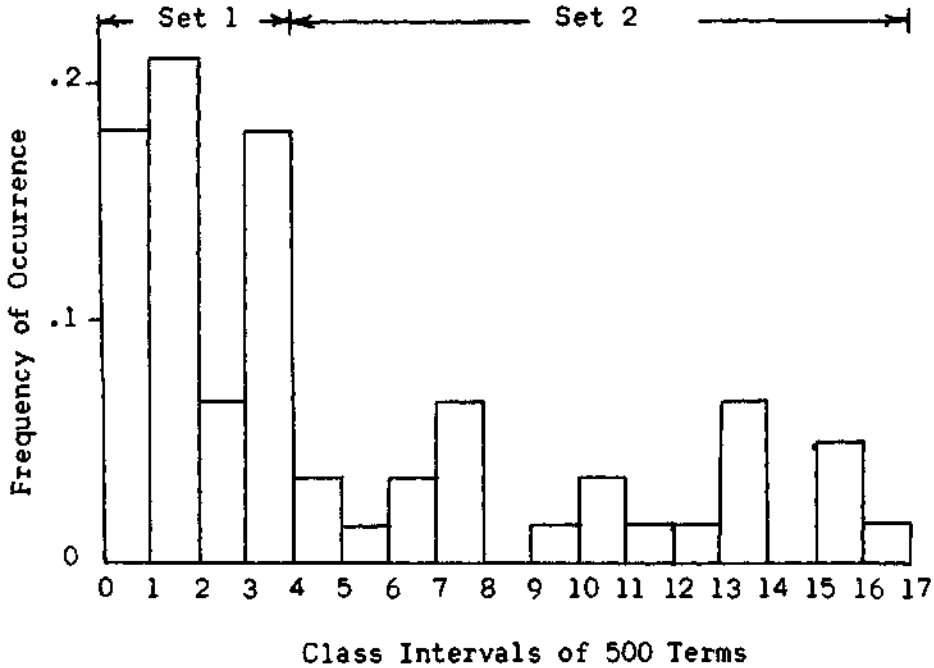
C. Clerical Personnel



					1	1		1	Manual
5	2								Uniterm
4	1	1							Peek-a-boo
4									Edge-notched Card
5	1	1	2	1	1				Simple Sorter
3	5	2		1					Collative
1		1							Photographic
5	1	1	2	1	1		1		Computer
22	10	3	3	1	1	1	1	1	References
3		3	1	1					Data
2				1	2				Search Aids

Figure 5. (Continued)

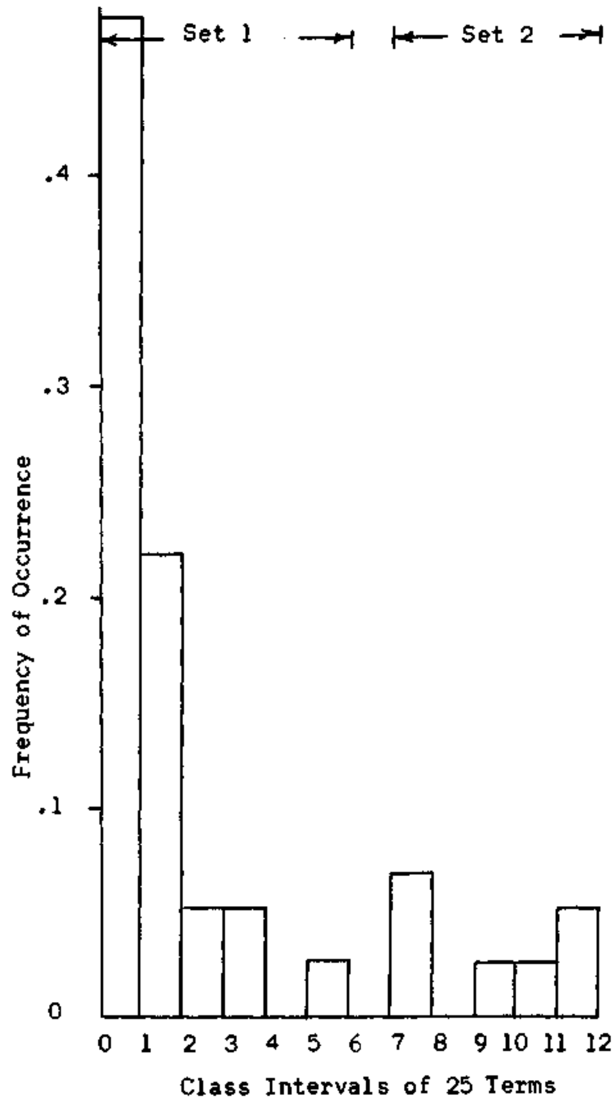
D. Size of Terminology Authority



			1				1									1	Manual
			1	1	1		1			1					1		Uniterm
1	4	3	1						1								Peek-a-boo
2	2			1													Edge-notched Card
3	7	1	5							1							Simple Sorter
			3			1	2					1	1		1		Collative
1																	Photographic
4						1					1		3		1		Computer
4	8	3	9	2	1	1	4			1	1	1	3		2		References
5	3	2				1			1				1				Data
2	2		1							1					1	1	Search Aids

Figure 5. (Continued)

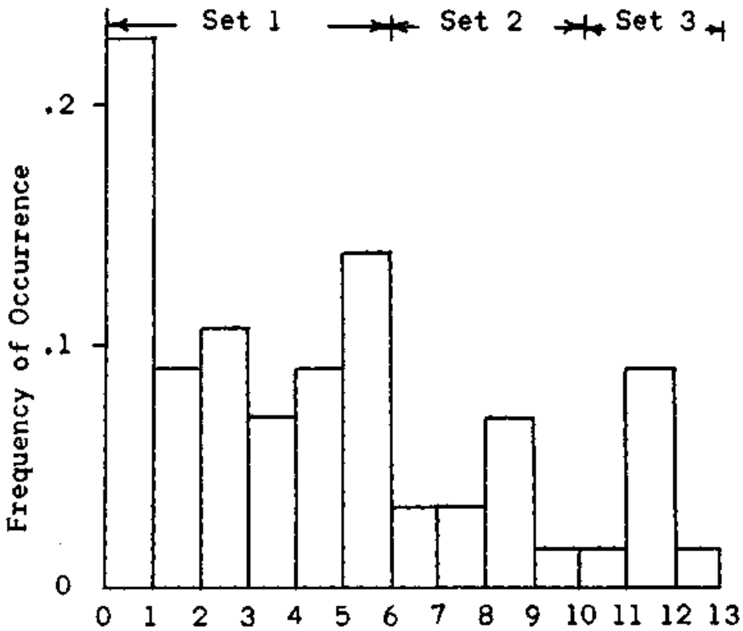
E. Rate of Addition to Terminology Authority



1			1									Manual
1		1					2					Uniterm
6	1											Peek-a-boo
1	1											Edge-notched Card
4	3								1			Simple Sorter
4	1	1								1	2	Collative
			1									Photographic
1	4				1		1					Computer
12	7	2	2		1		3		1	1	2	References
4	2											Data
2	1											Search Aids

Figure 5. (Continued)

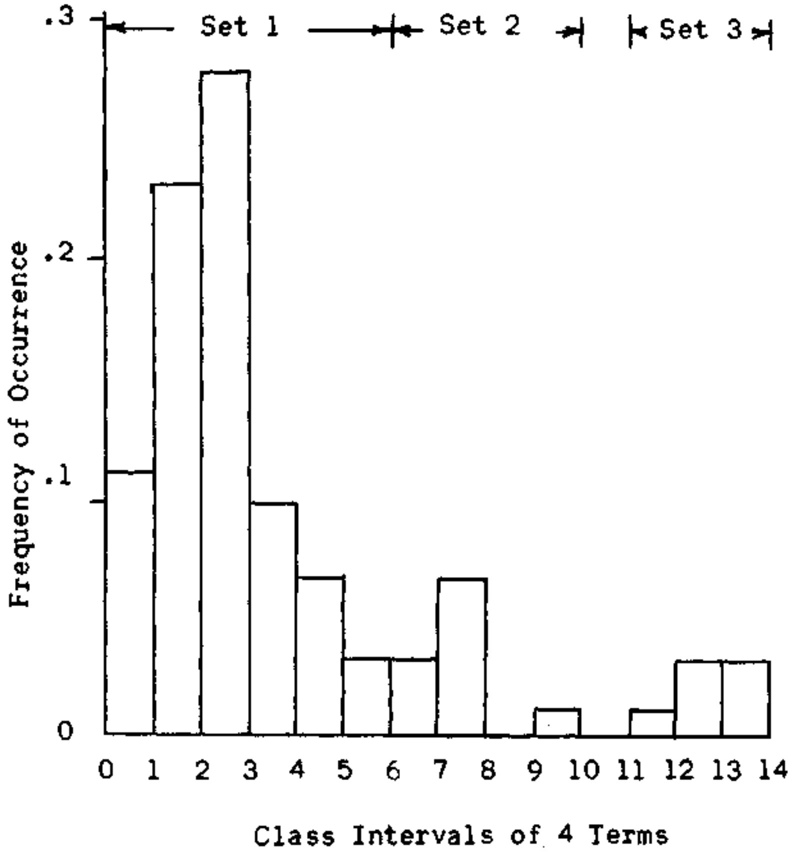
F. Input Processing Time



							1				1		Manual
		3	1	1	1								Uniterm
2	2	2			3								Peek-a-boo
											1		Edge-notched Card
7	2		1	2			1	1			2		Simple Sorter
	1				2	1		2		1		1	Collative
			1										Photographic
4		1	1	2	2	1		1	1				Computer
4	4	4	4	5	6	1		4	1	1	2	1	References
6	1	2			2						1		Data
3						1	2				1		Search Aids

Figure 5. (Continued)

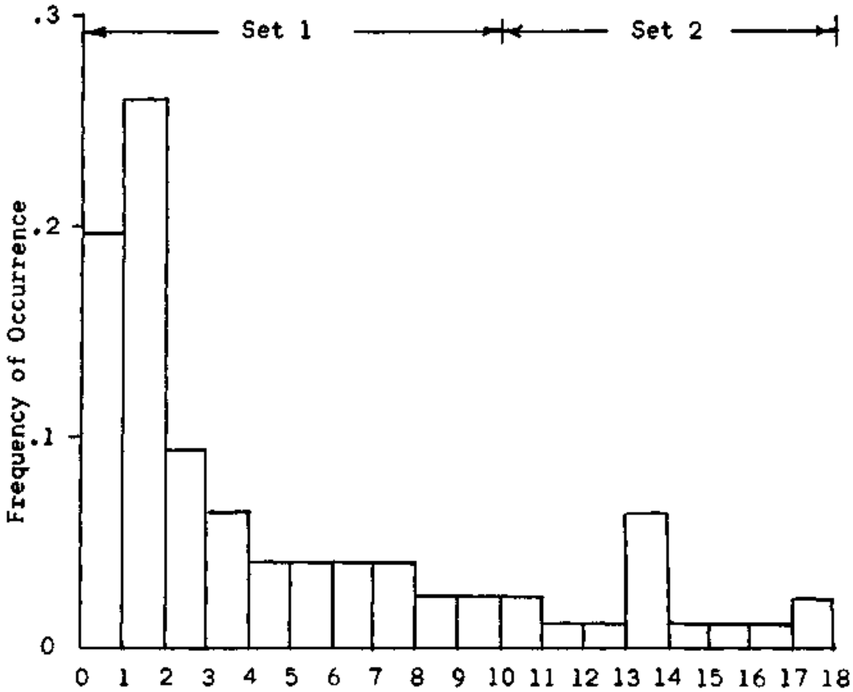
G. Depth of Indexing



1	2													Manual
	3	3												Uniterm
	2	3				1	2					1		Peek-a-boo
1		2											1	Edge-Notched Card
4	3	4		1	1						1	1		Simple Sorter
	1	1	3	1	1	1	1		1					Collative
	1	1												Photographic
1	3	3	2	2			1	1					1	Computer
3	11	15	4	3	2	2	1		1			1	1	References
1	1	2	1	1			1					1		Data
3	3	1	1				2	1			1		1	Search Aids

Figure 5. (Continued)

H. Size of Collection

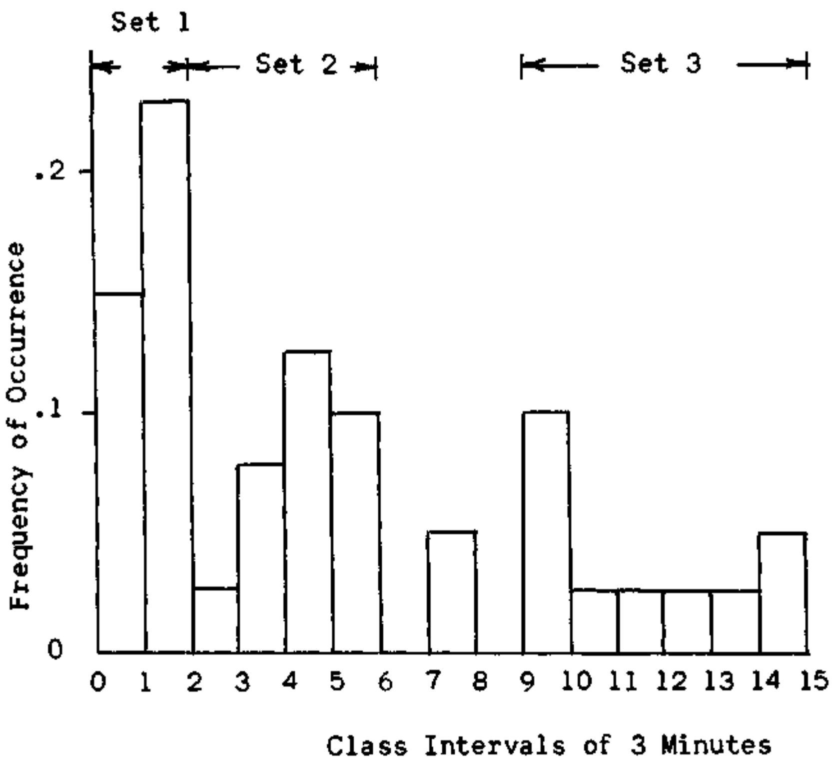


Class Intervals of 4,500 Items

								1		1								Manual
	3	1					1						1	1				Uniterm
4	2	1	1	1									1					Peek-a-boo
2		1			1				1		1							Edge-Notched Card
3	6	1	2	1		1			2			1	2		1	1		Simple Sorter
1	3	3	2				1	1										Collative
1																		Photographic
4	6			1	2	2		1					1				2	Computer
8	13	6	4	3	1	2	2	1	2		1	1	3	1		1	2	References
2	6		1			1		1		2			2					Data
5	1	1			2		1								1			Search Aids

Figure 5. (Continued)

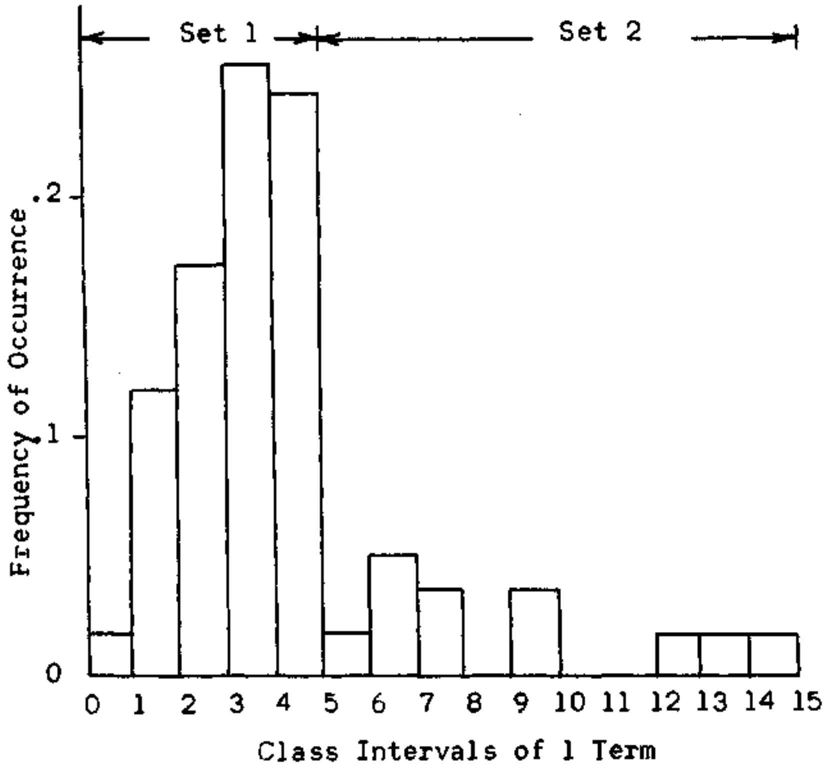
I. Search Time



1			1												Manual
		1							1					1	Uniterm
2	3		1	1							1				Peek-a-boo
															Edge-notched Card
1	2			2	1		1		2				1		Simple Sorter
		1			1									1	Collative
2															Photographic
1	3		2	1	1				1						Computer
6	5	1	3	5	2		1		3		1			2	References
1	3				2				1				1		Data
															Search Aids

Figure 5. (Continued)

J. Terms per Question



															Manual
		4	2	1											Uniterm
		1	4	2		1	1		1						Peek-a-boo
					1										Edge-Notched Card
	2	3	4	3		1									Simple Sorter
1	1	1	1	3			1		1				1	1	Collative
	1			1											Photographic
	2	1	4	4								1			Computer
1	4	10	11	11	1	1	2		2			1	1	1	References
	2		4	3		1									Data
															Search Aids

Figure 5. (Continued)

APPENDIX D

The following pages contain the numerical attribute associations.

Table 6. Numerical Attribute Associations

Producer Attribute	Product Attributes	Range	Members
Rate of growth of collection	Professional Personnel, Set 1	1-6	64
	Set 2	8-12	6
Rate of growth of collection	Clerical Personnel, Set 1	1-9	54
Rate of growth of collection	Input Processing Time, Set 1	1-30	39
	Set 2	31-50	9
	Set 3	51-65	5
Rate of growth of collection	Size of Collection, Set 1	1-45,000	59
	Set 2	45,001-81,000	14
Professional personnel	Clerical Personnel, Set 1	1-9	52
Rate of addition to term. auth.	Size of Term. Auth., Set 1	1-2,000	22
	Set 2	2,001-8,500	15
Input processing time	Depth of Indexing, Set 1	1-24	41
	Set 2	25-40	7
	Set 3	45-56	4
Professional personnel	Depth of Indexing, Set 1	1-24	43
	Set 2	25-40	8
	Set 3	45-56	4

(Continued)

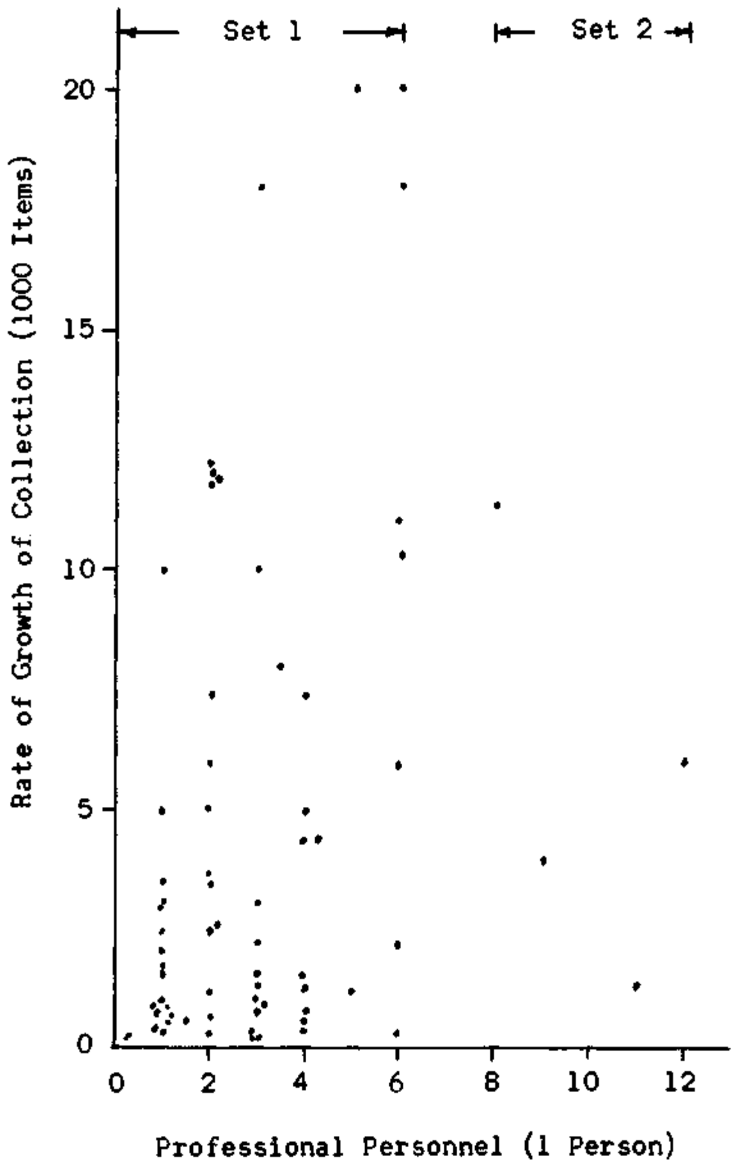
Table 6. (Continued)

Producer Attributes	Product Attributes	Range	Members
Terms per question	Search Time Set 1	1-6	12
	Set 2	7-18	11
	Set 3	30-45	7
Depth of Indexing	Search Time, Set 1	1-6	11
	Set 2	7-18	12
	Set 3	30-45	8
Size of collection	Search Time, Set 1	1-6	15
	Set 2	7-18	12
	Set 3	30-45	9

APPENDIX E

The following pages contain the scatter-grams of the numerical attribute associations.

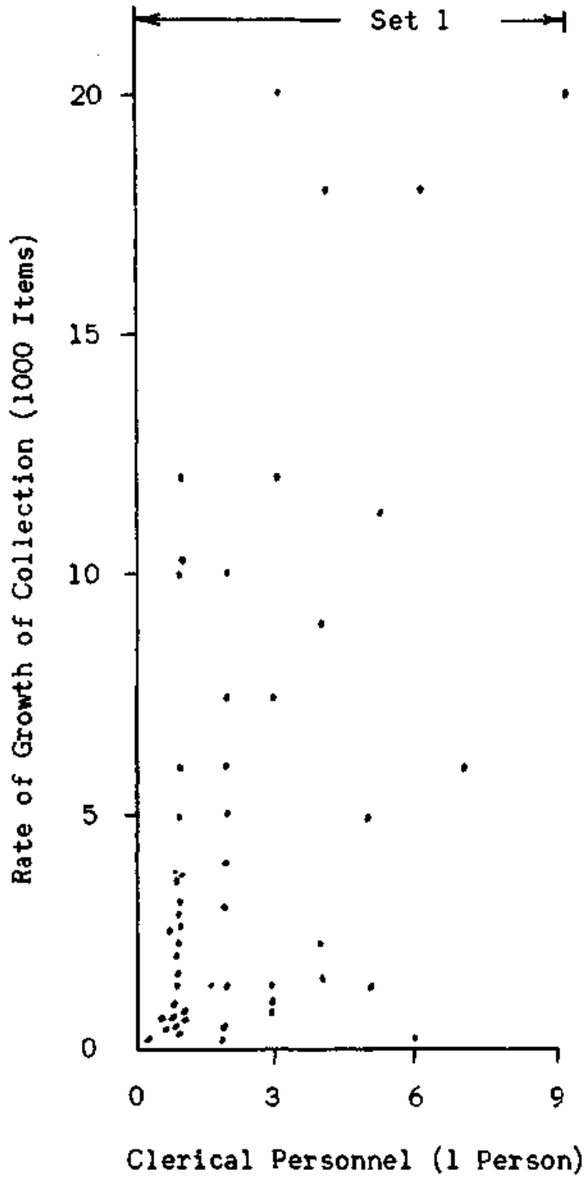
A. Professional Personnel vs Rate of Growth of Collection



Persons	Items
1	22,500
11.1	36,000
12	50,000

Figure 6. Scatter-grams.

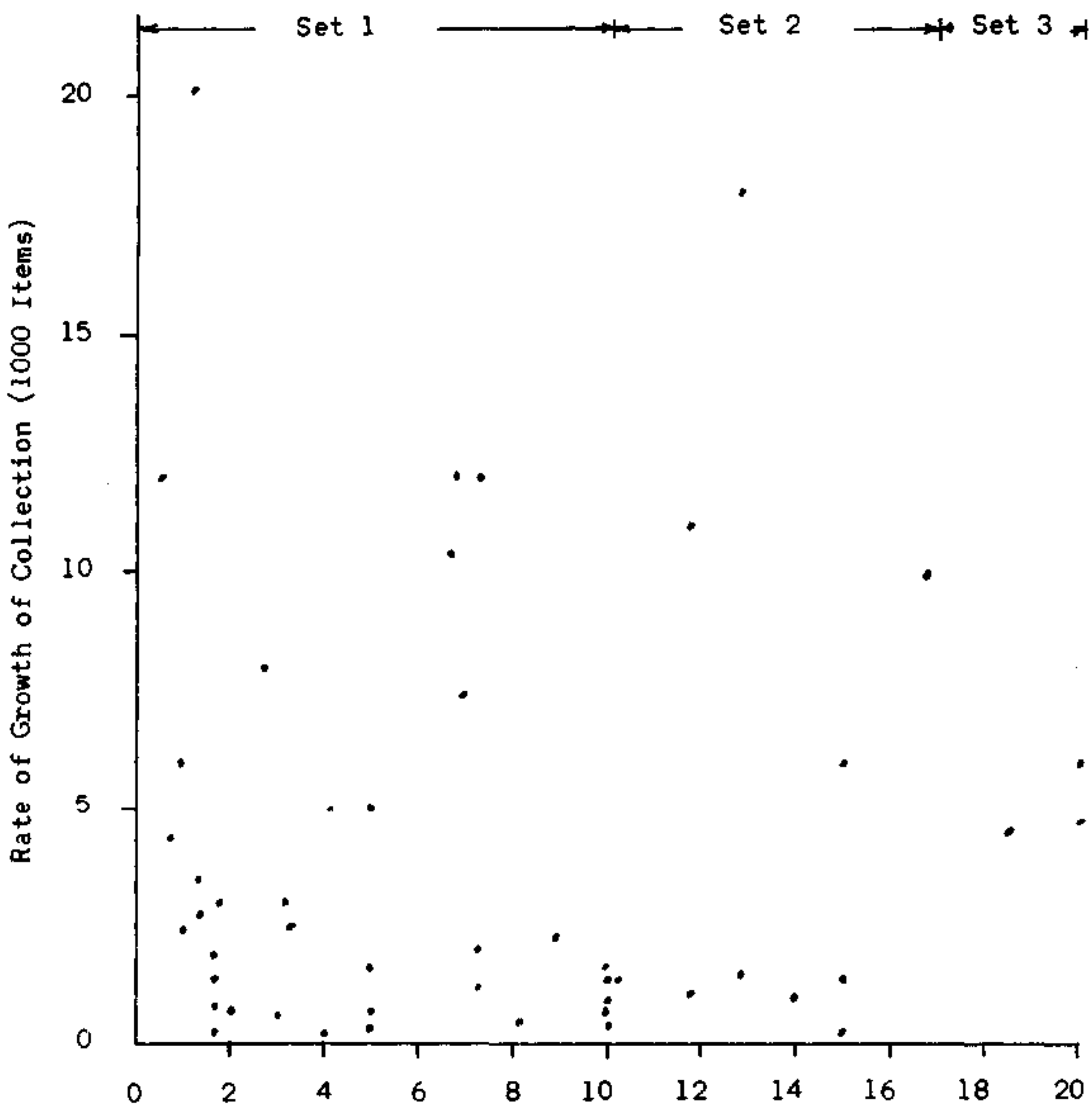
B. Clerical Personnel vs Rate of Growth of Collection



Persons	Items
6	50,000

Figure 6. (Continued)

C. Input Processing Time vs Rate of Growth of Collection



Minutes	Items
10	225,000
25	36,000

Figure 6. (Continued)

D. Size of Collection vs Rate of Growth of Collection

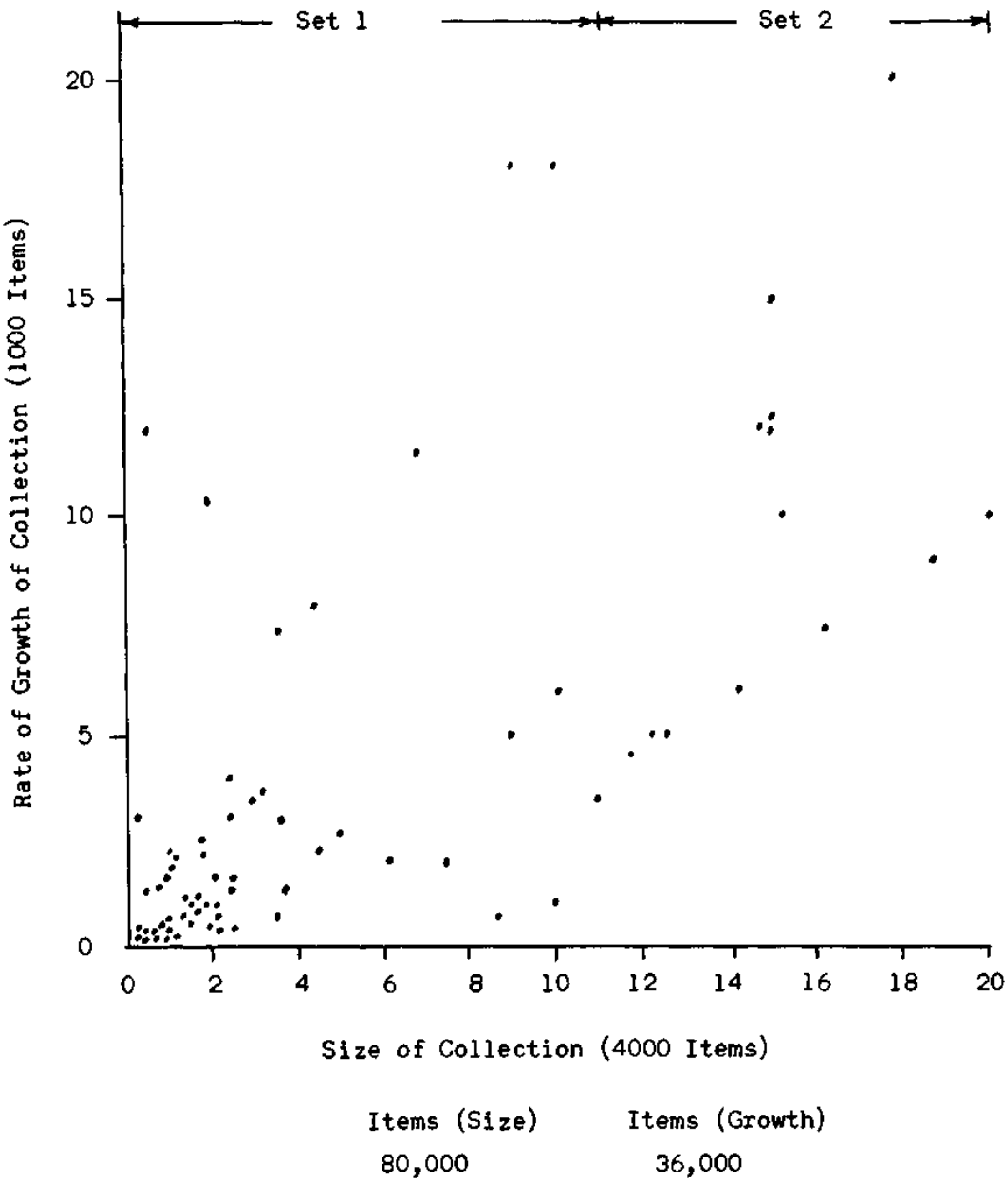


Figure 6. (Continued)

E. Clerical Personnel vs Professional Personnel

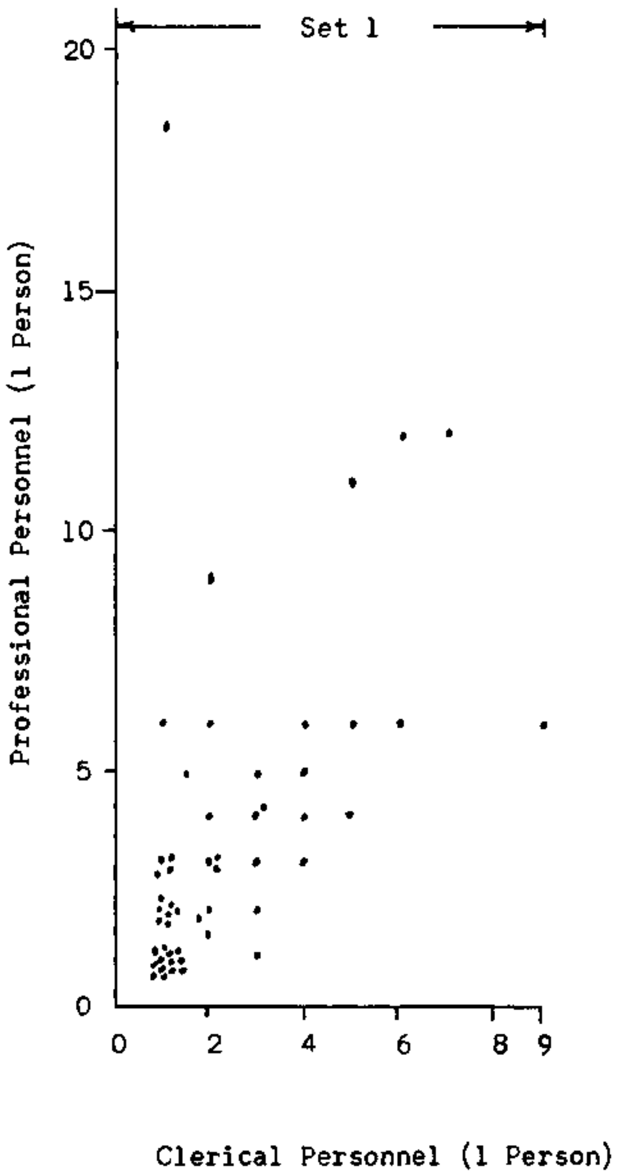
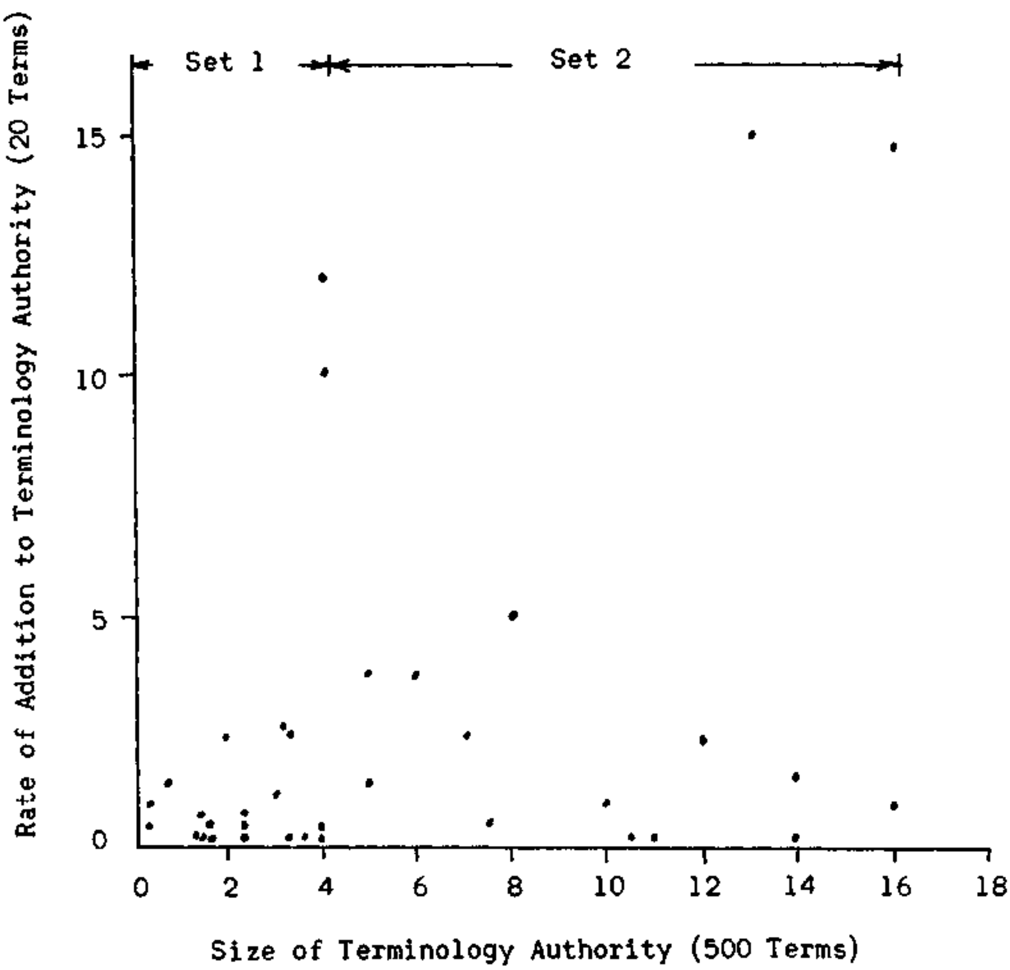


Figure 6. (Continued)

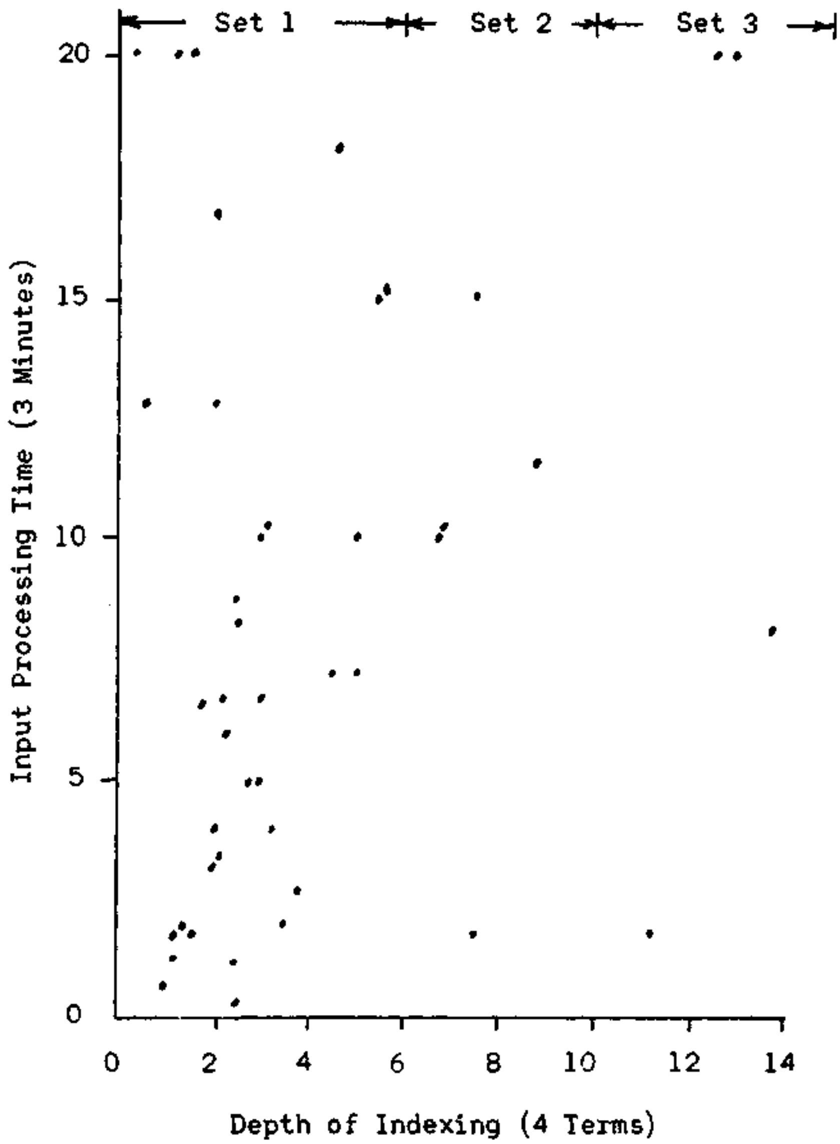
F. Size of Terminology Authority vs Rate of Addition to Terminology Authority



Size (Terms)	Addition (Terms)
1000	500

Figure 6. (Continued)

G. Depth of Indexing vs Input Processing Time



Terms	Minutes	Terms	Minutes
3	120	10	64
9	300	30	720
10	180	30	90
12	120		
12	120		
13.5	120		

Figure 6. (Continued)

H. Depth of Indexing vs Professional Personnel

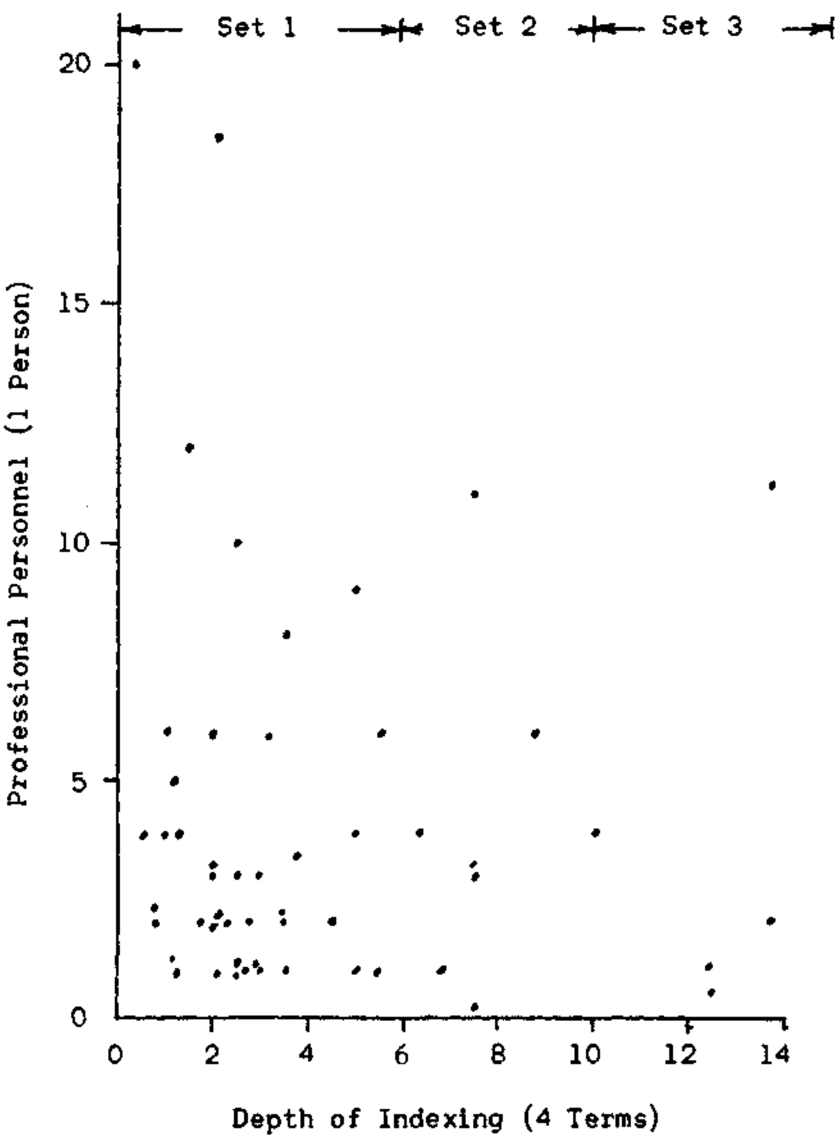


Figure 6. (Continued)

I. Search Time vs Terms per Question

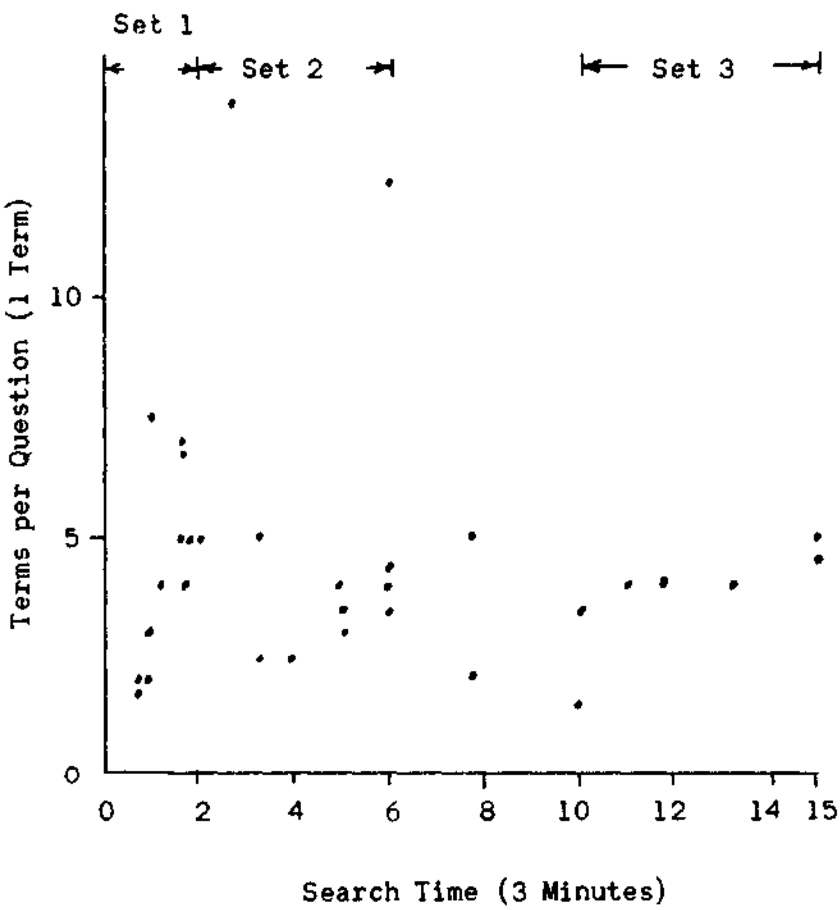
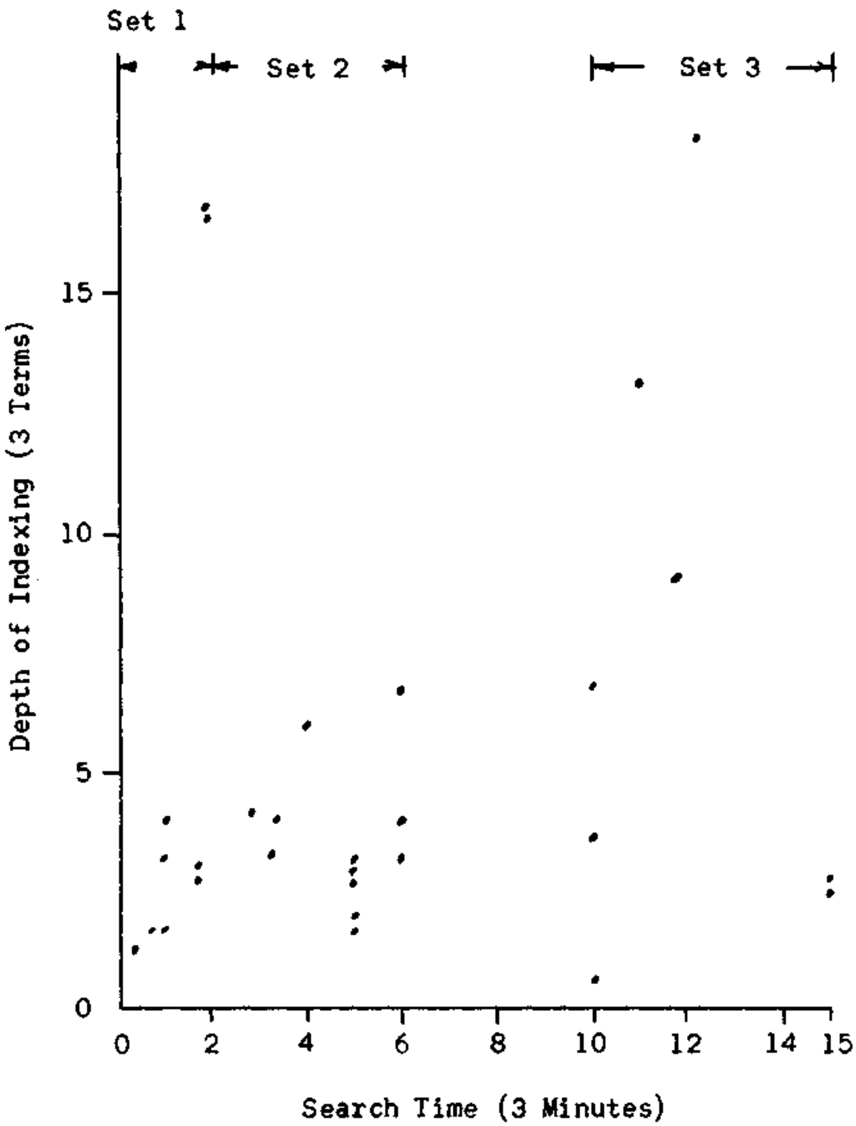


Figure 6. (Continued)

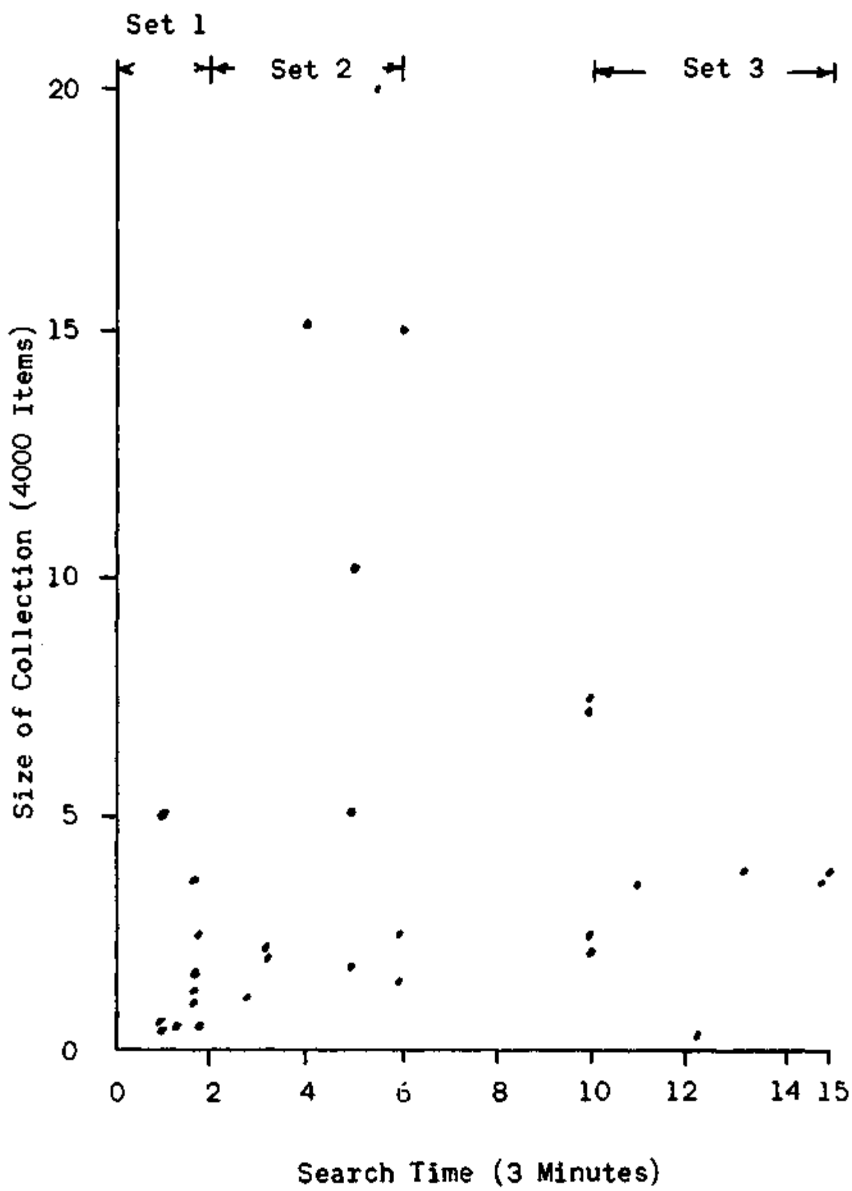
J. Search Time vs Depth of Indexing



Minutes	Terms
4	400
5	75

Figure 6. (Continued)

K. Search Time vs Size of Collection



Minutes	Items
1	400,000
2	300,000
2	275,000
2	640,000
6	275,000

Figure 6. (Continued)

APPENDIX F

The following pages contain the results of this study.

Table 7. Results

	Set 1	Set 2	Set 3
A. Rate of growth of collection			
1. Degree of mechanization:			
Manual	.33	.00	.67
Uniterm	.86	.14	.00
Peek-a-boo	.90	.10	.00
Edge-Notched Card	1.00	.00	.00
Simple Sorter	.86	.05	.09
Collative	1.00	.00	.00
Photographic	.00	.50	.50
Computer	.43	.31	.06
2. Contents of the index file:			
References	.84	.12	.04
Data	.79	.07	.14
Search Aids	.64	.18	.18

(Continued)

Table 7. (Continued)

	Set 1	Set 2
B. Professional Personnel		
1. Rate of growth of collection	-.0452	.4232
2. Degree of mechanization:		
Manual	.67	.33
Uniterm	1.00	.00
Peek-a-boo	1.00	.00
Edge-Notched Card	1.00	.00
Simple Sorter	1.00	.00
Collative	.85	.15
Photographic	1.00	.00
Computer	.76	.24
3. Contents of the index file:		
References	.90	.10
Data	.93	.07
Search Aids	.86	.14

(Continued)

Table 7. (Continued)

	Set 1
C. Clerical Personnel	
1. Rate of growth of collection	.5016
2. Professional personnel	.4955
3. Degree of mechanization:	
Manual	1.00
Uniterm	1.00
Peek-a-boo	1.00
Edge-Notched Card	1.00
Simple Sorter	1.00
Collative	1.00
Photographic	1.00
Computer	1.00
4. Contents of the index file:	
References	1.00
Data	1.00
Search Aids	1.00

(Continued)

Table 7. (Continued)

	Set 1	Set 2	Set 3
D. Input Processing Time			
1. Rate of growth of collection	.2419	-.1209	-.1764
2. Degree of mechanization:			
Manual	.00	.50	.50
Uniterm	1.00	.00	.00
Peek-a-boo	1.00	.00	.00
Edge-Notched Card	.00	.00	1.00
Simple Sorter	.74	.13	.13
Collative	.38	.37	.25
Photographic	1.00	.00	.00
Computer	.77	.23	.00
3. Contents of the index file:			
References	.73	.16	.11
Data	.92	.00	.08
Search Aids	.43	.43	.14

(Continued)

Table 7. (Continued)

	Set 1	Set 2
E. Size of collection		
1. Rate of growth of collection	.4901	.6427
2. Degree of mechanization:		
Manual	1.0	.00
Uniterm	.71	.29
Peek-a-boo	.90	.10
Edge-Notched Card	.83	.17
Simple Sorter	.76	.24
Collative	1.00	.00
Photographic	1.00	.00
Computer	.84	.16
3. Contents of the index file:		
References	.82	.18
Data	.73	.27
Search Aids	.91	.09

(Continued)

Table 7. (Continued)

	Set 1	Set 2
F. Size of Terminology Authority		
1. Rate of addition to the terminology authority	.1686	.2763
2. Degree of mechanization:		
Manual	.33	.67
Uniterm	.17	.83
Peek-a-boo	.90	.10
Edge-Notched Card	.80	.20
Simple Sorter	.94	.06
Collative	.33	.67
Photographic	1.00	.00
Computer	.40	.60
3. Contents of the index file:		
References	.60	.40
Data	.77	.23
Search Aids	.63	.37

(Continued)

Table 7. (Continued)

	Set 1	Set 2
G. Rate of Addition to the Terminology Authority		
1. Degree of mechanization:		
Manual	1.00	.00
Uniterm	.50	.50
Peek-a-boo	1.00	.00
Edge-Notched Card	1.00	.00
Simple Sorter	.87	.13
Collative	.67	.33
Photographic	1.00	.00
Computer	.86	.14
2. Contents of the index file:		
References	.77	.23
Data	1.00	.00
Search Aids	1.00	.00

(Continued)

Table 7. (Continued)

	Set 1	Set 2	Set 3
H. Depth of Indexing			
1. Input processing time	-.0155	-.1450	.5433
2. Professional personnel	-.1454	-.1719	.6690
3. Degree of mechanization:			
Manual	1.00	.00	.00
Uniterm	1.00	.00	.00
Peek-a-boo	.72	.14	.14
Edge-Notched Card	.50	.33	.17
Simple Sorter	.87	.00	.13
Collative	.70	.30	.00
Photographic	1.00	.00	.00
Computer	.79	.14	.07
4. Contents of the index file:			
References	.85	.10	.05
Data	.74	.13	.13
Search Aids	.62	.23	.15

(Continued)

Table 7. (Continued)

	Set 1	Set 2	Set 3
I. Search Time			
1. Terms per question	.5917	-.1794	.7535
2. Depth of indexing	.1902	-.0618	-.1009
3. Size of collecting	-.5780	.3863	.1094
4. Degree of mechanization:			
Manual	.50	.50	.00
Uniterm	.00	.33	.67
Peek-a-boo	.63	.25	.12
Edge-Notched Card	.00	.00	.00
Simple Sorter	.33	.33	.33
Collative	.00	.67	.33
Photographic	1.00	.00	.00
Computer	.44	.44	.12
5. Contents of the index file:			
References	.39	.40	.21
Data	.50	.25	.25
Search Data	.00	.00	.00

(Continued)

Table 7. (Continued)

	Set 1	Set 2
J. Terms per Question		
1. Degree of mechanization:		
Manual	.00	.00
Uniterm	1.00	.00
Peek-a-boo	.70	.30
Edge-Notched Card	.00	1.00
Simple Sorter	.92	.08
Collative	.64	.36
Photographic	1.00	.00
Computer	.92	.08
2. Contents of the index file:		
References	.80	.20
Data	.90	.10
Search Aids	.00	.00

LITERATURE CITED

1. J. Becker and R. M. Hayes, "Information Storage and Retrieval: tools, elements, theories," John Wiley and Sons, Inc., New York, 1963, pp. 292-293.

2. B. C. Vickery. "On Retrieval System Theory," Butterworths, London, 1961, p. 139.

3. S. L. Optner, "Systems Analysis for Business and Industrial Problem Solving," Prentice-Hall Inc., Englewood Cliffs, N. J., 1965, pp. 26 - 27.

4. R. L. Ackoff, "Scientific Method," John Wiley and Sons, Inc., New York, 1962, p. 16.

5. "Nonconventional Technical Information Systems In Current Use," National Science Foundation, NSF-62-34, 1962, pp. xv-xvi.

6. A. H. Bowker and G. J. Lieberman, "Engineering Statistics," Prentice-Hall, Inc., Englewood Cliffs, N. J., 1963, pp. 245-274.

7. Bowker, p. 558.

8. Statements provided for the National Science Foundation report, "Nonconventional Technical Information Systems in Current Use."