Some factors affecting relevance judgements

Janet M. Pope

Thesis submitted for the degree of Doctor of Philosophy

The City University

Centre for Information Science

December 1976

CONTENTS

TABLES

TABLE OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my gratitude to all of the people who provided assistance and encouragement during the preparation of this thesis.

I am especially grateful to Mr. J. [redacted] for suggesting this area of study; to Dr. S. [redacted] for making signal detection theory comprehensible; to Dr. S. [redacted] for many stimulating discussions and for running several computer programs and to my supervisor Dr. R. T. Bottle.

I would also like to thank the directors of INSPEC and UKCIS for providing the SDI profiles used in this study.

Finally I wish to express my gratitude to all of the users who co-operated in this study, for giving up so much of their valuable time and for their patience and forebearance in answering my questions. Without their willing assistance this thesis could not have been written.

ABSTRACT

Relevance judgements were obtained from 31 research workers in Universities and Government Institutions, on titles and abstracts. The titles were obtained using SDI profiles for each user. A questionnaire was given to each user to discover their information gathering habits, attitudes towards information and various demographic details.

Signal detection theory was used to obtain a single measure of a user's sensitivity to relevance and the strictness of his criterion of relevance. Four different detection models were examined. It was found that an exponential model was most suitable for analysing relevance data.

Several variables were tested to see whether they affected relevance judgements. Several statistical tests were used to test the effects of the variables on the conditional probabilities of hits and false alarms and also on the sensitivity and criterion measures. It was found that 15 of the 37 variables examined significantly influenced relevance judgements. These variables related to the user's work and to his use of different information sources.

The distributions of the hit and false alarm probabilities were lognormal in all except one case.

The variables examined were highly inter-related. Factor analysis and principal component analysis were used to reduce the complexity of these variables and to discover the underlying factors. Ten factors were extracted from a total of 36 variables.

The research students in the sample formed a distinct group in terms of their information gathering habits and attitudes towards information.

An attempt was made to understand the types of relevance judgements made by users by placing each judgement in one of five categories. This arbitrary classification was helpful in examining the reasons why various relevance decisions were made.

A model of information seeking and acquisition was proposed which takes into account the storage of information in human memory and the way in which information is organized. This model was used to describe the relationships between information need, question asking and the subsequent relevance judgements.

SYMBOLS USED

T_1	Original set of titles
T_2	Repeat set of titles
A	Abstracts
$T_1 - T_2$	Comparison of relevance judgements on T_1 and T_2
$T_1 - A$	Comparison of relevance judgements on T_1 and A
$T_2 - A$	Comparison of relevance judgements on T_2 and A
P(S/n)	Conditional probability of a false alarm
P(S/s)	Conditional probability of a hit
d'e	Sensitivity (unequal variance signal detection model)
β	Response bias (unequal variance signal detection model)
α	Sensitivity (Luce)
ν	Response bias (Luce)
Δm	Sensitivity (unequal variance signal detection model)
d'	Sensitivity (equal variance signal detection model)
κ	Sensitivity (exponential signal detection model)
a_r	Response bias (Robertson)
δ_r	Sensitivity (Robertson)

ABBREVIATIONS USED

IR	Information Retrieval
SDT	Signal Detection Theory

PCA	Principal Component Analysis
ABNO	All but not only
OBNA	Only but not all
ILL	Inter Library Loan
STM	Short Term Memory
LTM	Long Term Memory
CAC	Chemical Abstracts Condensates

Lao Tzu

1. INTRODUCTION

Information science is concerned with the communication of information between people, and the underlying processes involved. Communications between people occur on a stimulus-response basis; where a statement by one person is received by a second person and a response (of whatever nature) is made by the receiver to the initial stimulus. In cases where the receiver does not comprehend the stimulus no response is possible and communication is not achieved.

Information is relevant to the receiver if it is within his knowledge or 'terms of reference'. The concept of relevance within the framework of a discussion has been examined by Weiler [1962]. He looked at cases where certain information is either relevant, non-relevant or partially relevant to a subject under discussion, depending on the degree of definition of the framework of the discussion. It can be concluded that relevance is a fundamental aspect of human communication.

The concepts of information and relevance within the field of communication have been discussed for many years, with ideas ranging from the mathematical theories of information first expounded by Shannon and Weaver to the psychology of how information is comprehended and stored in human memory. In information science the concepts of relevance and information are considered in a restricted

framework, such as can be handled as working concepts. A definition of a communication process that has been adopted by Saracevic [1970a] is that:

> "A communication process is ..... a sequence of events in which something called information is transmitted from one object (source) to another object (destination), often in a series of reiterative or feedback sequences."

The objects, sources and destinations may be people, records, machines or systems in general.

The relevance of material in an interchange between people is usually obvious, the difficulty arises when the communication process is between a human and a machine, as occurs in information retrieval (I.R.) systems. In the case of man-machine communications the flow of information is severely restricted because the two sides of the interaction operate using entirely different 'terms of reference'. In this situation the human has to adapt to the machine's communication pattern, as the machine cannot (at least at the present) adapt to the human process.

Relevance is of fundamental importance in communications in general and to information science in particular, or as Saracevic [1975] puts it:

"In the most fundamental sense, relevance has to do with effectiveness of communication. Underlying all information systems is some interpretation of the notion of relevance."

The basic purpose of I.R. systems is to provide relevant information to their users. Despite the recognition by information scientists of the fundamental nature of relevance there is still a great deal of argument over definitions of relevance and its suitability as a criterion for assessing I.R. systems. It has been said [Doyle 1963, Rees 1965] that relevance is a very inadequate measure, but there is at present no other alternative.

This feeling arose from the fact that very little was known about relevance, the relationship between relevance per se and human relevance judgements and the factors which may affect relevance judgements. Although much more information about relevance judgements and the factors which affect them is known now, the problem of the suitability of relevance as a measure of the performance of information systems is still with us.

1.1 Theories of relevance

The beginnings of the concept of relevance as used by information scientists were in the early 1950's, when the problems of the retrieval of non relevant material by I.R. systems were under discussion. The views expressed at that time were affected by Shannon's treatment of noise in information theory. Relevance was regarded as solely the property of the system and the documents within the system.

At the International Conference for Scientific Information (ICSI) in 1958 [National Academy of Sciences 1959], there was the first recognition that relevance may not be a simple phenomenon related to the system alone. The consensus of the ICSI debate was that relevance is:

a) More than the operation of relating, performed internally within systems
b) Not exclusively the property of document content or document relatedness
c) Not a dichotomous decision
d) There is such a thing as 'user relevance' which can be judged, thus a notion of relevance as considered by a destination was born

There was no explicit discussion of the subjectivity of relevance. The views in 1958 were based on logical and intuitive inferences, as no experimental evidence was available.

After this debate there arose two different approaches to relevance, firstly there were attempts to build a theory of relevance based on logical, mathematical and statistical methods. The second approach accepted the human nature of relevance, and concentrated on the environment and the variables which affect relevance decisions. Essentially the first group were looking at system relevance and the second group were looking at human relevance.

The two aspects of relevance became very mixed up, this confusion was noted by Taube [1965] who indicated

> "...the illegal shift from subjective relevance as a reaction of a user, to mathematical relevance as a property of systems..."

The confusion arose mainly because of the lack of distinction between

> "relevance as a relation between propositions and the recognition of relevance or its judgement by a user, which resembles a utility or significance judgement." [Cooper 1971]

The confusion between system relevance and human relevance continues, and is perpetuated in the definition of relevance given by Saracevic in 1975. In a very general definition he regarded relevance as

"a measure of the effectiveness of contact between a source and a destination in a communication process."

1.11 System relevance

The earliest attempt to find a mathematical theory of relevance was by Maron and Kuhns [1960] who treated relevance as a primitive concept. They regarded relevance as a property of the document alone. Goffman [1964] defined relevance as

"a measure of information conveyed by a document to a query."

This measure was obtained with respect to a set of documents which contain the relevant document, rather than only the document itself. In a later paper Goffman [1968] stated that

"the relevance of a document is not absolute but may depend upon what information is conveyed by other documents in the file which is being interrogated ...consequently the relation between a query and a document is not sufficient to establish relevance."

The idea of using the conceptual relatedness of documents and queries was proposed by Hillman [1964]. Similarity-judgements made by people were used to provide empirical data for the formal definitions of concepts and conceptual relatedness. These similarity judgements depend on human relevance judgements and are difficult to accomodate in a formal theory.

A model of relevance based on the matching of documents and queries was suggested by Jackson [1970]. The relevance of documents obtained from the model was used to construct pseudo-classifications which could then be applied to predict the relevance of subsequent documents and for the evaluation of I.R. systems.

Konigova [1971] distinguished three types of relevance

i Formal relevance (correspondence between terms in the document and the query)

ii Content relevance (correspondence between the content of the document and query)

iii Subjective or human relevance

She used formal and content relevance to develop ideas about 'noise' in I.R. systems, but recognised the influence of subjective relevance.

> "It is impossible to exclude the influence of subjective comprehension of the content of the document in question."

Cooper [1971] proposed a definition of relevance in terms of logical implication. For simple YES-NO answering

systems his definition was mathematically precise, but for systems where the information is expressed in natural language (i.e. the majority of I.R. systems) this definition is not precise. More recently Cooper [1973] has attempted to give a measure of relevance using peoples' subjective evaluation of the utility of a system's output. He concluded that this method is impractical because of the assumptions and compromises that have to be made, and any measurement obtained would have a high error factor.

In an extension of Cooper's ideas of a logical basis for a definition of relevance, Wilson [1973] suggested a measure which he called 'situational relevance.' This is the relation between an item of information and an individual's situation as he sees it, rather than as others see it. Situational relevance is sensitive to the dynamic nature of an individual's particular interests.

The inconsistencies in relevance judgements made by various judges was taken into account in a simple probabilistic model developed by Gebhardt [1975]. In this model the relevance assigned by a judge is considered a random variable. One of the conclusions reached by Gebhardt is that the worst possible method of assessing relevance is a mere bisection into relevant and irrelevant.

Robertson [1975] has suggested an interesting variable which relates to the closeness in subject matter

between document and query, he calls this 'synthema'. Relevance is regarded as an aspect of synthema, relevant documents are those which are highly synthematic to the question. Synthema is a more generalized concept than relevance since it deals with documents that would always be non relevant; such documents can be ranked or partitioned according to their 'distance' from the question. Like relevance, synthema is a continuous variable, and can be thought of as a kind of semantic distance between document and query.

The various theories and arguments about relevance were regarded by Weiler [1973] as being typically cross-paradigm arguments, which have the effect of re-drawing the boundaries of paradigms. He felt that there are basic philosophical uncertainties that any adequate theory of relevance must take into account.

1.12 Human relevance

Human relevance is the subjective human decision regarding the relation of a document to a particular query or information need. There have been several attempts to define relevance from the human point of view. Rees and Schultz [1967a] summarized these definitions using lists of what relevance is and what relevant materials and information are.

Saracevic [1970b] endeavoured to make sense of the mass of definitions using an algorithm of the form:

Relevance is the A of a B existing between a C and a D as judged by an E where

A is the gauge of measure

B is the aspect of relevance

C is the object judged

D is the context within which relevance is measured

E is the assessor

This algorithm is expanded in Figure 1

Doyle [1963] regarded relevance as a subjective and elusive property which cannot be used as a criterion for measure in I.R. systems. Cuadra [1964] on the other hand conceded that relevance is subjective, but called for rigorous experimentation in order to obtain greater understanding of the problems involved in relevance judgements. Rees and Saracevic [1966] also accepted that relevance judgements contain a subjective element, they believed that under conditions where the variables of time, people, subject area and responses are controlled, relevance could be an objective and useful measuring instrument. Relevance is subjective and is an expression of the user, not the system, on the degree of matching of a document

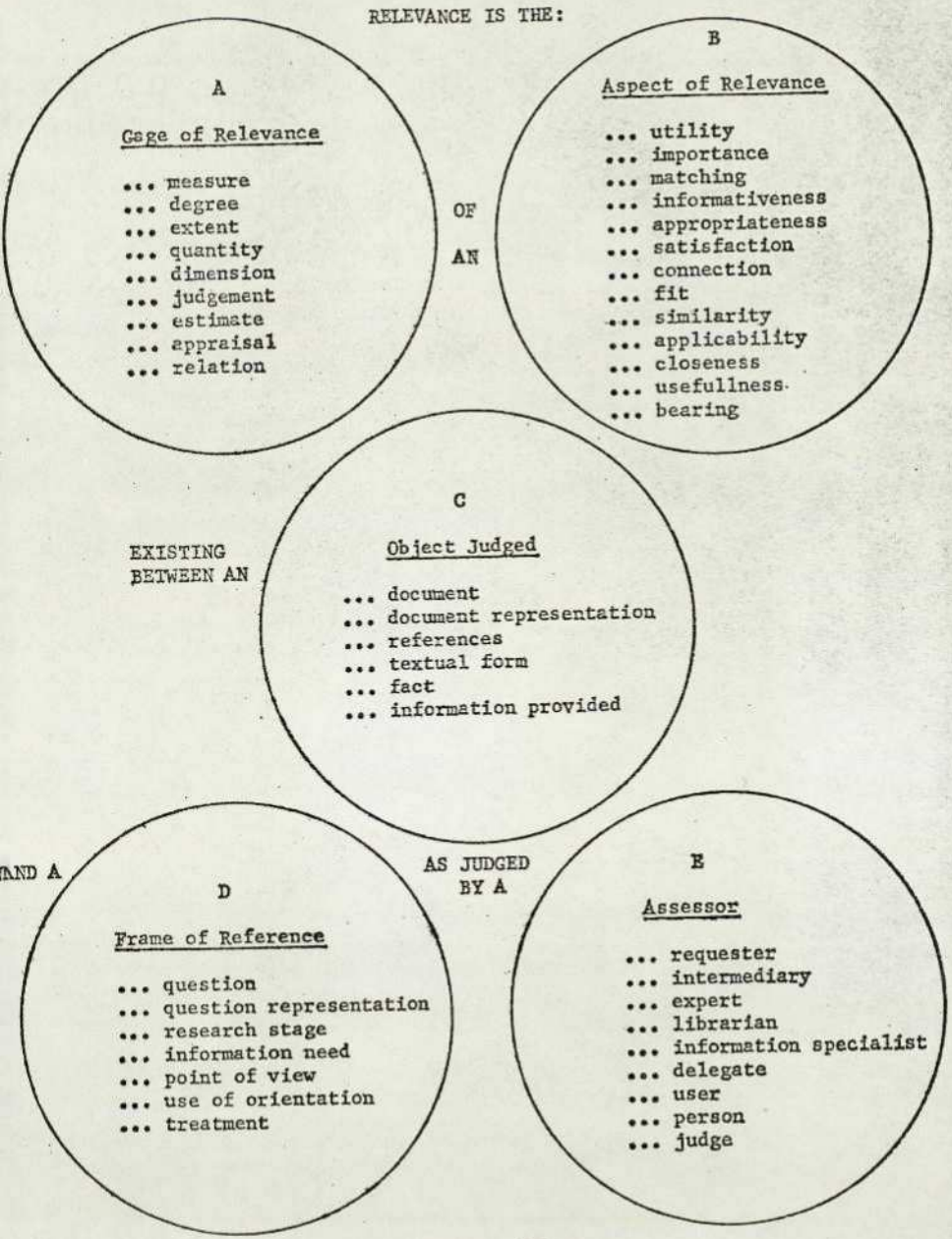


FIG. 1. Definitions of Relevance.

with a query. [Rees 1966]. In a similar way Kent et al [1967] regarded relevance;

> "not as an inherent characteristic of a document but rather as a function of certain idiosyncratic variables of the user as he evaluates the relevance of a particular document in relation to his own information need at that time."

The fact that relevance judgements are basic to the operation of I.R. systems was noted by Snyder [1966], who felt that there is no absolute or final criterion of what is relevant. Relatively little is known about how people judge relevance, and we need to know what guidelines are used by people in deciding whether a document is relevant. Snyder suggested that psychonomic techniques could provide a basis for investigating the lack of reliability in relevance judgements.

The destinction between human (user) relevance and system relevance was clearly made by King and Bryant [1971]. They felt it highly unlikely that system relevance would be the same as user relevance, and stated that

> "System relevance responses are independent of user relevance judgements in the sense that the relevance numbers are assigned by different entities."

According to Lancaster and Fayen [1973] relevance is the subjective assessment of user satisfaction and is open to

a great number of influences. It is related to users information needs and is a value judgement based on these needs. In a model of relevance decisions, Cook [1975] assumed that there is a relevance threshold for every individual which is a fundamental characteristic. The relevance judgement is a matching process between a document and the user's information need. Cook suggested a method for determining the user's relevance threshold based on his relevance judgements on a set of documents from his subject area.

Relevance is now generally regarded as being a continuous variable, which for convenience may be dichotomised into relevant and irrelevant, or given several arbitrary categories (for example highly relevant, partially relevant, possibly relevant, irrelevant) [Saracevic 1969, Leggate 1971, Robertson 1975a].

1.13 Relevance and pertinence

Several people have distinguished between the relevance and pertinence of documents [Rees 1963, Salton 1965, Foskett 1972]. The consensus of their views is that relevance is the relation of a document to a particular request, and can be agreed upon by several judges. The pertinence of a document relates to an information need and can only be decided by the user. Thus there is a clear distinction between relevance to a question and pertinence to the information need underlying the question.

Pertinence is judged subjectively by a user in relation to his information need. In the evaluation of the retrieval performance of I.R. systems, relevance can and should be used as an objective measure [Kemp 1974]. The relationship between user need, user request, relevance and pertinence is given in Figure 2.

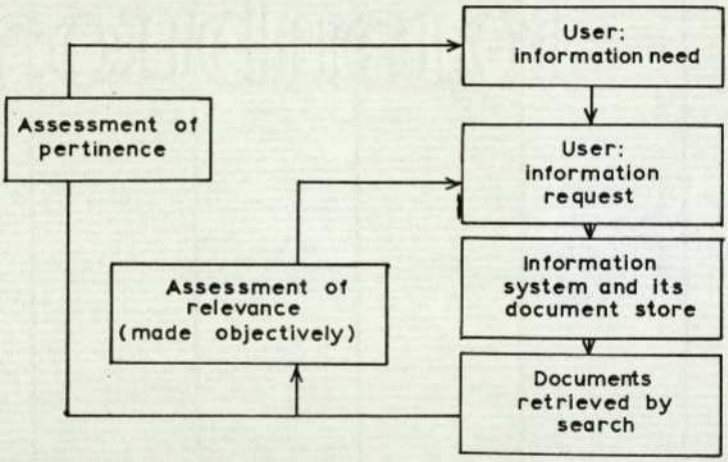


Figure 2

Relationship between relevance & pertinence

1.2 Experiments on relevance

The call by Cuadra [1964] for an experimental study of relevance was recognized by the National Science Foundation (NSF) conference in 1964. One of the main conclusions of this conference was that:

> "a major obstacle to progress in the area of (I.R. systems) evaluation methodology is the lack of sufficient knowledge of the character and variability of human assessments of the relevance of retrieved documents ..... Such relevance assessments are fundamental to the development of reliable methods and techniques for measuring the effectiveness of document searching systems."

Two large scale experimental studies of relevance were carried out in the late 1960's, by Cuadra and Katter at the Systems Development Corporation (SDC) and by Rees, Saracevic, Schultz and others at Case Western Reserve University (CWRU).

Cuadra and Katter [1967] suggested 38 variables which might affect relevance judgements (see Figure 3). These variables fall into five broad categories; the document, the judgemental conditions, the judge, the information request statement and finally the available mode of expression. Of these variables 19 were examined in the SDC experiments. The aim of the SDC study was to

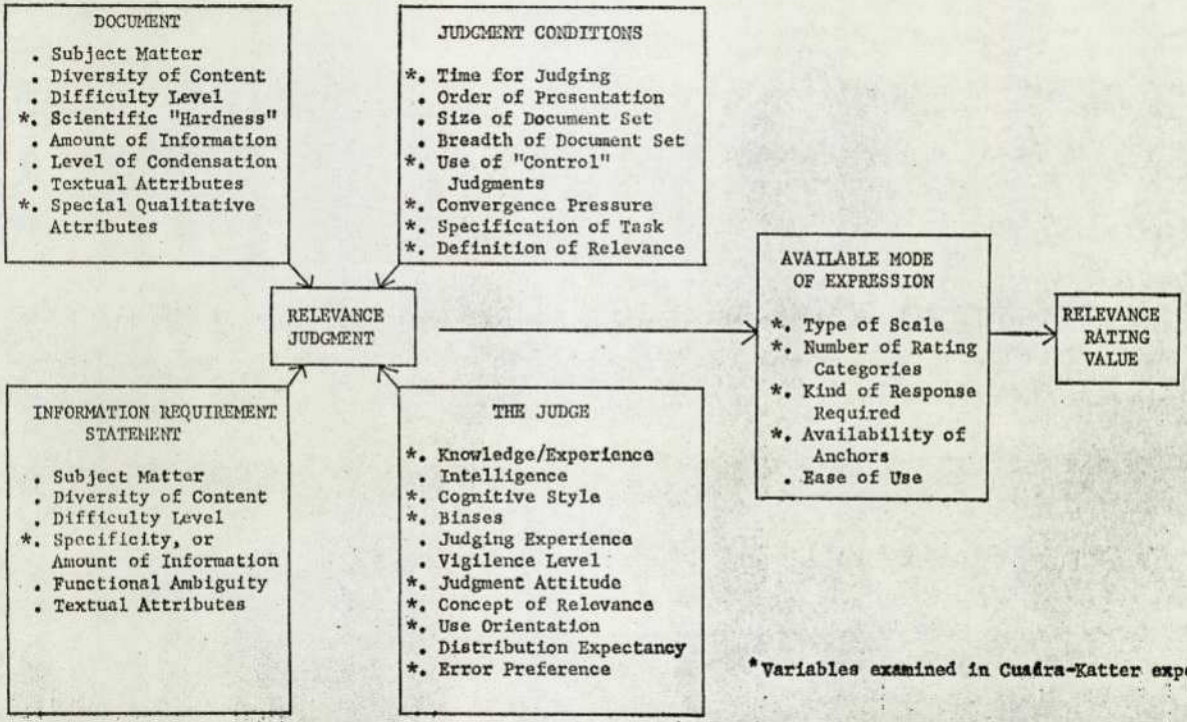


FIG. 3 Variables Related and Potentially Related to Relevance Judgments as Hypothesized by Cuadra and Katter (1967).

identify the variables thought to affect relevance judgements and to conduct a series of laboratory studies to determine the affects of these variables on relevance judgements.

The relevance experiments at CWRU were part of a much larger study of the testing and evaluation of I.R. systems, performed by the Comparative Systems Laboratory (CSL). The CSL was initiated with the very broad aim to observe and compare systematically the behaviour of various experimentally constructed I.R. systems. One of the specific objectives was to gain further understanding of, and to design methodologies to test, the variables and processes operating within retrieval systems. This objective included an examination of the variables affecting relevance judgements [Saracevic 1968].

The study at CWRU was on relevance judgements performed in the context of the simulation of a completed research project. The objective of the experiment was to observe systematically some of the psychological and environmental factors which were hypothesised to influence the relevance judgement process [Rees and Saracevic 1966].

These major experiments and several smaller ones carried out in the 1960's have been comprehensively reviewed by Saracevic [1970b]. Brief mention will be made of these studies together with later experiments.

1.21 Documents and document surrogates

There have been many studies that examined various document surrogates as predictors of the relevance of the full text, to a given query. These have shown that titles are poor indicators of the eventual relevance of documents, abstracts perform quite well as relevance predictors [Rath, Resnick and Savage 1961, Shirey and Kurfeerst 1967]. Kent et al [1967] noted that relevance is difficult to assess on titles as these are often ambiguous. Resnick and Savage [1964] found however that judgements were made consistently and were independent of the surrogate or text used, except for an inconsistency with respect to abstracts which they could not explain.

Rees and Schultz [1967a] and Hannah [1971] found that judges usually overestimate the relevance of a document when judged on titles alone. Research workers find it easy to distinguish non-relevance and high relevance from titles alone [Elwen 1972].

The probability of making a successful relevance prediction using titles is only increased by about 10% when the titles are augmented with other portions of the document (e.g. keywords or abstracts) [Persson 1974]. In the INSPEC evaluation it was found that relevance predictability of the full document is worst with titles alone, and that abstracts perform as well as augmented titles [Clague 1971].

1.22 Questions

Correlations between the occurrence of terms in queries and in relevant documents have been found by Tague [1965] and Gifford and Baumanis [1969]. A similar approach was adopted by Sastri [1968], who looked at the co-occurrence of concepts rather than words or word combinations.

Ivankin [1975] examined the semantic relationships between the texts of queries and relevant documents. He was particularly interested in the specificity of keywords and the frequency of occurrence of various relationships between keywords in queries and in documents. The specificity of questions affects the consistency of relevance judgements made by various judges. The more specific queries produced significantly higher inter-judged agreement [Cuadra and Katter 1967].

The studies by Sastri and Gifford and Baumanis were part of the CSL project. They were based on the assumption that a user is guided in his relevance judgement primarily by the presence of words or linguistic features in a document that are present in the question. Hence the relevance judgement is objective in its matching of question and answer concepts.

In an analysis of the questions answered by a research association library, Hibbert [1972] found that the best

results were from enquiries which stated why the information was needed. Enquiries by letter required more clarification than those received by telephone, as questions can be discussed over the telephone. Written enquiries received by three information departments were examined to see why interaction between users and information staff was required [Friend 1970]. In most cases it was because of incomplete or vague initial requests. There was a significant difference in the number of faulty requests dealt with by academic and industrial information departments.

O'Connor [1969] suggested that disagreements between judges, on the relevance of documents to a query are due either to unclear requests and documents, or due to disagreements about the meaning of a passage. In an experiment he found that disagreements in relevance judgements between judges were resolved upon discussion.

1.23 Judgement conditions

In the experiments by Cuadra and Katter [1967], and Rees and Schultz [1967a], the exact conditions and situation had to be defined for the judges, as non user judges were employed. Most of the investigations of the effects of varying these conditions are irrelevant to the present research as user relevance is being investigated.

The most important variables found in these studies were, the definition of relevance and the stage of research at which the query was made.

Garvey, Tomita and Woolf [1974] investigated the various types of information required at different stages of research. At the beginning and end of a research project researchers require a great deal of both theoretical and experimental information. During the middle stages much less information is required, and it is mainly of an experimental nature. Users have lenient criteria of relevance at the beginning and end of a project, and more stringent criteria in the middle of a project.

Cuadra and Katter [1967] and Katter [1968] looked at various scales of relevance, to see how they affect relevance judgements. They examined ranking scales, rating scales and scales with various numbers of categories. Judges using six or eight category scales were more certain in their relevance judgements than judges using fewer scales. Rees and Schultz [1967a] found that a two point scale (relevant or irrelevant) was not sufficiently sensitive for relevance judgements, and advocated the use of scales with several categories. A six point scale was used by Carroll and Tague [1973]. Here users could distinguish amongst the various levels but the discrimination requested was too fine, so the authors suggested that a five point scale or less should be used.

In the evaluation of operational I.R. systems, Leggate [1971] says

"there is no point in burdening users with more than three categories for judging relevance."

He pointed out that each user has a highly personal interpretation of the definition of each category. The level of discrimination applied by users is affected by the information available in making the judgement (title, abstract etc.) and by the way in which a relevant reference is noted (marking a copy or writing out the citation)

1.24 People

The effects of the academic training of judges, in a given subject area, on relevance judgements were investigated by Cuadra and Katter [1967]. They found that the group with the lowest academic training showed the least consistency in relevance judgements. Those with intermediate training showed the greatest consistency and rather surprisingly, the most highly trained group had a fairly low consistency in judgements.

Carrington [1973] found that the variations in relevance judgements made by research workers in similar subject areas were due to differences in experience. Inexperienced researchers tended to overestimate the relevance of abstracts, compared with more experienced researchers.

Relevance judgements made by medical researches, students and librarians on titles and full texts, were compared by Rees and Schultz [1967a]. The researchers judged more liberally on titles than on full texts, whereas the librarians tended to find titles less relevant than the full text.

A few studies comparing user and non-user relevance judgements have been carried out. Barhydt [1964, 1967] compared these two types of relevance judgements and concluded that subject experts and system experts show a high consistency in assessing the relevance of a document to a query, but that neither match at all well the user's relevance evaluation.

Similar results were obtained by Dym [1967] using 'motivated' and 'quasi-motivated' users. Again it was found that judges and users did not give consistent relevance judgements. These inconsistencies occur because the user judges the pertinence of a document to his information need and this task cannot be delegated to a judge [Foskett 1972]. Whereas the judge assesses the relevance of a document to the query.

1.25 Criticisms of relevance experiments

There have been several criticisms of, and doubts about the value of the experiments on relevance

judgements. Several of these criticisms are concerned with the fact that the majority of studies used non-users as relevance judges, which made it necessary to provide elaborate definitions of relevance, use orientation and many other factors. This emphasis on external environmental variables has provided a great deal of information concerning the problems of assessing relevance; however very little attempt has been made to examine the factors which affect relevance judgements made by 'real-life' users of operating I.R. systems. Cole [1966] made this criticism of the experiments at CWRU, and stated that they lacked realism. He suggested that terms indicative of the utility of documents are better than using relevance in the evaluation of I.R. systems.

The use of non-user relevance judgements in testing I.R. systems has been questioned. Rees and Saracevic [1966] criticised the use of such relevance judgements:

> "In a number of experiments evaluating the performance of a retrieval system, relevance judgements are performed by judges or juries of judges on the basis of short questions of which they are not the author. Such a relevance judgement, made in reference only to the subject content of an impersonal question without taking anything else into account, is a very crude simulation and approximation of a judgement made in relation to a real information need. These artificial judgements have such serious limitations as instruments that they have little or no value."

Leggate [1971] also doubted the value of non-user judgements as they have little bearing on user relevance judgements, which are required for the evaluation of operational I.R. systems. He states that relevance decisions should not be based on the user's written statement of interests, but on his actual information need. In a review of relevance experiments King [1968], said:

> "The reports (of relevance experiments) reveal the hazards of using relevance assessments blindly. The prescriptions are given for avoiding such pitfalls, however, and one must concede that, since relevance as a concept cannot be avoided, it must be used. However, it is comforting to know that one can now properly feel nervous about using it."

1.3 Evaluation of I.R. systems with motivated users

It is only within the last eight years that evaluations of I.R. systems have been carried out under operating conditions, where the actual user of the system makes the relevance (or rather pertinence) judgements. The need for study of both system and user relevance in I.R. systems was recognized by Goffman and Newill as early as 1966, they stated:

> "The system can be optimized only in terms of measures that indicate the degree of agreement between system assignment of relevance and user assignment of relevance.....Human factors and variability must be

recognized, carefully studied and either accounted for or controlled in the study of I.R. systems."

Evaluations of three major SDI services have been carried out under, as nearly as possible, normal operating conditions, these services are MEDLARS [Lancaster 1971], UKCIS [Barker, Kent and Veal 1970] and INSPEC [Clague 1971]. In addition the Experimental Information Unit at Oxford carried out evaluations of three data bases as potential SDI systems [Leggate 1973 a,b]. In all of these evaluations users were selected from a variety of locations, and relevance judgements were obtained on the SDI output received over a period of time. The performance of these systems was measured in terms of recall and precision. A great deal of information on users information searching habits, attitudes towards information and ability to keep up to date was collected as part of these studies.

An interesting and valuable technique used in these evaluations was that of failure analysis. This involves examining the reasons for recall and precision failures. Precision failures were looked at on the basis of the users relevance judgements. Lancaster [1971] found that in the MEDLARS evaluation, the major cause of both recall and precision failures was imperfect request statements. He suggested that these failures could be reduced by improving user-system interaction. Leggate,

Rossiter and Rowland [1973b], in the evaluation of an SDI service using the Index Chemicus Registry System [ICRS] data base, also observed that a breakdown in the communication process was responsible for most of the recall and precision failures. This breakdown arose mainly from inadequate requests which did not reflect the users' actual information need. In the UKCIS evaluation [Barker, Kent and Veal 1970] it was again found that requests did not express fully or accurately the users' information needs.

Clague [1971] found in the INSPEC evaluation that 40% of recall failures were due to items being called relevant which were considered outside the users information request statement (IRS). Precision failures were attributed to the general nature of the IRS, and the fact that relevance assessments were made by users without direct reference to their IRS. A small study was carried out by INSPEC to see whether certain characteristics of poor IRS's had any affect on the performance of subsequent profiles; the characteristics studied were lack of detail and lack of precision. No clear results were obtained, but this may have been due to the very small number of profiles examined and the fact that the SDI service was completely new to most of the users.

The findings from failure analysis indicate very clearly the problems and dangers of using judges rather than

users to evaluate the relevance of references retrieved in response to a given query or IRS. A judge can make relevance decisions on the SDI output, based on the IRS which bear little relationship to the relevance of the output to the users actual information need. Failure analysis shows that the IRS does not necessarily reflect the users information need. It is obviously very important to reduce the precision failures and prevent communication failures by greater interaction between users and the people operating SDI systems.

In an experiment using B.A. Previews as an SDI service, Leggate et al [1973a] surveyed users reactions to the service. This survey found that 81% of users had difficulty at some time in deciding the relevance of an article to their work, using titles only. Measurements of relevance predictability showed an 18% probability that a user will not be able to predict correctly the relevance of a document using enriched title alone. Most of the comments made by users on the output received during the B.A. Previews and ICRS experiments were concerned with the amount of irrelevant material retrieved. In very few cases were suggestions made for profile amendment and improvement.

1.4 Information needs

There has been much discussion about the information needs of scientists, but there is no accepted meaning of

this expression. Rees [1963] noted that there was confusion between the information wanted by users and their real needs. The expressed want was often an imperfect reflection of what was actually needed, if indeed this was known. Rees also outlined some of the problems involved in defining information need. Some definitions of information needs, wants, and demands have been suggested by Line [1974].

Need:- what an individual ought to have for his work, research, recreation etc.

Want:- what an individual would like to have

Demand:- what an individual asks for

Requirement is a useful bridging term that can refer to information that is needed, wanted or demanded. Line pointed out that many studies of information need were in fact studies of information requirement.

The problems of what is actually meant by information need were highlighted by O'Connor [1968]. He thought that there were three different interpretations of 'need' in the literature. Two of these correspond to Line's definitions of need and want, the third interpretation relates to the negotiation of initial requests to find the information need behind the verbalized question.

One of the difficult problems that I.R. services have to cope with is the difficulty that some users have in expressing their information needs. In some cases this is just an inability to verbalize a clearly felt need [King and Bryant 1971]; in other cases the problem is more profound as the user does not really know what he wants. [Fairthorne 1963]. The inability to verbalize a need can be readily overcome by negotiation of the user's query. The user who does not know what he wants is much more difficult to deal with; this problem may not be resolved despite lengthy negotiations between the user and a system liaison scientist.

According to King and Palmour [1974], users information needs are affected by their behaviour pattern of seeking and exchanging information, their way of working and the subject area. Needs change with time due to results of previous research and/or changing interests. They are also affected by the accessibility and ease of location of information.

In a study of chemists, Arnett [1970] found that

> "...the information storage and retrieval needs of chemists depend to a considerable extent upon their particular fields of work and upon their particular research styles and personal habits..."

Different people have different information needs. The various categories and groups of users that have been suggested were reviewed by Marchlewska [1970]. The bases used for the classification of users included; professional education, professional occupation (e.g. research, management, technicians) and level of research expertise. The needs of people in different occupations and at different levels of expertise vary considerably.

There are also different types of information required by people. The most obvious distinction is that between current awareness and retrospective information needs. There are many other types of information that users need, six categories of information have been suggested by Herner [1962]. These were conceptual, empirical, procedural, stimulatory, policy and directive information. The type of information required by a user depends on whether it is well within his subject field or on the fringe of it, and also on his stage of research [Bernal 1960]. In a review of information needs, Nanlin and Garvey [1972] gave several factors which affect these needs, they included; type of work (research, teaching etc), discipline and basic or applied work.

The literature searching habits of scientists have been extensively studied. They are beyond the scope of the present research project and will not be mentioned

except in passing. Reviews of the literature searching and reading habits of scientists can be found in the Annual Review of Information Science and Technology, Parker and Paisley [1966] and in Meadows [1974].

In a study of the reading habits of chemists, Gushee [1968] found that they varied from being highly literature orientated, to having a virtual literature phobia. Chemists specialize in their reading and pick out articles to read from the table of contents in journals. General titles get more attention than specific titles, however if they are too general they become vague and this causes a loss of interest. The longer an article is the less likely it is to be read, although very short articles are also unlikely to be read.

The personal literature searches of scientists are governed by subjective influences. In order to satisfy his entire information need, the user cannot dispense with the browsing serendipity of personal searches. These factors must be recognized when a literature search is delegated (e.g. to another person or to an I.R. service). Recognition of the inherently subjective nature of information seeking, by information scientists will prevent futile attempts at making retrieval operations more objective and formal, and will advance scientific communication. [Fugmann 1973]

1.5 Suggestions for future work and the present study

There have been two main suggestions for further work in the area of relevance. The first suggestion was that the methodology used should be improved, and that a firm theoretical basis for research into relevance should be established. The second suggestion was that psychology and psychological methods should be applied to relevance studies.

Rees and Schultz [1967b] advocated the use of psychological techniques to investigate the variables affecting relevance judgements. Parker [1966] thought that social psychology, communication theory and verbal behaviour studies could be used to investigate users needs, the communication of information and the classification of information. He noted that;

> "...information seeking and information processing behaviour of scientists should be the subject of considerable psychological research."

At the present time there have been very few studies of relevance using psychological techniques and, as far as can be gathered, none have looked at the psychology of information needs and information processing in real life situations. Semantic differentials have been used by Meister and Sullivan [1967] and by Katzer [1972] to obtain the attitudes of users towards information and

sources of information. The current state of psychological research into information acquisition and processing is of little help to information science, as the work is only at the stage of examining lists of words and simple concepts.

This present research is concerned with investigating the variables which may affect relevance judgements, made by actual users of an SDI service. The aim was to discover the variables that affect relevance judgements and also to obtain information on the ways in which research workers make relevance decisions. With an understanding of the processes involved in making relevance judgements it should be possible to explain the effects of certain variables on the relevance decision process. As part of this investigation the psychophysical technique of signal detection theory has been applied, for the first time, to the relevance judgements made by users on titles and abstracts. Using this technique a single measure of a users sensitivity to relevance was obtained. In addition to obtaining relevance judgements, a large amount of information regarding the users backgrounds and information gathering activities was obtained.

In order to examine the actual processes involved in making relevance judgements a variety personality and other tests would have to be given to each user. This is

outside the scope of the present study, and would require a trained psychologist to carry it out successfully.

Research into relevance has been neglected since the late 1960's, despite the fact that little is known about this elusive and controversial subject. The need for further work on relevance has been expressed by Saracevic [1975], who said that:

> "The better we understand relevance, the better chance we have of avoiding failures and of restricting the variety of aberrations committed in the name of effective communication. In that lies the importance of advancing the thinking on relevance."

2. THEORETICAL CONSIDERATIONS OF INFORMATION SEEKING AND ACQUISITION

2.1 Information storage in human memory

Psychological studies have indicated that human memory is composed of three different types of memory store, these are the sensory register, short term memory (STM) and long term memory (LTM). A distinction is usually made between the memory storage network where information is recorded, and the control processes that govern the flow and sequencing of information. The relations between the memory stores are given in Figure 4.

Stimuli impinge on the system via the receptors and are transmitted to the sensory register which analyzes and transforms the input. The information is briefly retained in the sensory register while it is selectively read into one of the memory stores. The STM is a working memory of limited capacity from which information decays fairly rapidly unless maintained by rehearsal. The contents of LTM are permanent, and it is essentially a large memory bank. The LTM can be thought of as the current state of consciousness of the subject. Information once recorded does not decay, but its availability depends on the effectiveness of the memory retrieval processes. The two memories STM and LTM, although usually depicted as separate, are not necessarily neurologically separate.

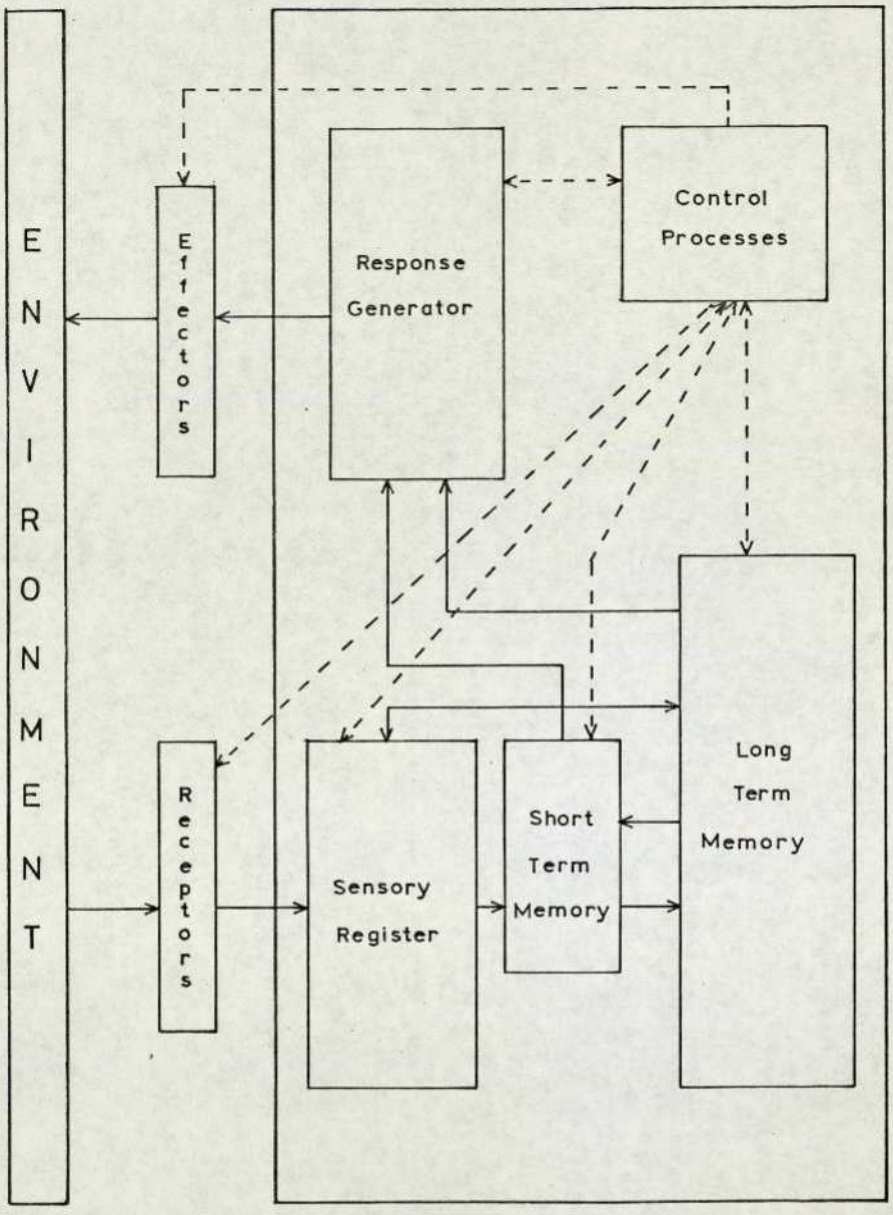


Figure 4 Flowchart of the memory system

—— Path of information transfer

----- Connections that permit comparison

(from Atkinson and Juola , 1974)

The control processes regulate the transfer of information from one store to another, and also the sequencing of operations within each memory store. These processes are strategies adopted by the subject in response to environmental and task conditions, they include selective attention, selection of retrieval cues and all types of decision strategies. [Atkinson and Juola 1974].

When information is presented to a person it is initially classified and stored in STM, then checked against LTM to see if it is new information. This is the decision process. If the information is new it is processed and transferred to LTM, the learning process. The outcome of learning is that information is stored in LTM which was not there previously, or an existing structure is modified. [Greeno and Bjork 1973]. Both the decision and learning processes are heavily influenced by the way information is stored and retrieved from LTM, and on how well it is stored.

In the context of making relevance judgements, when an item is presented to a person he first has to check his LTM to see if the item relates to the concepts present and on the basis of this search, make a relevance decision. If the item is judged relevant, it will be transmitted into LTM by a learning process which will modify (probably only slightly) the concepts stored in LTM.

2.2 Model of information seeking and acquisition

A model of information requirements and relevance judgements has been suggested by Harmon [1970]. It is based on the amount of information known by the person and on the way in which it has been stored in the memory. This model of information need consists of cognitive sets which are built up into a complete mental 'picture' using cognitive organizers. The addition of successive pieces of information to the person's knowledge store enables the completion of these cognitive sets. Harmon suggested that there are four stages of development within the model, these stages correspond to the various stages of ordering of the sets and organizers. These stages are

i Insufficient and unordered information

ii Insufficient but ordered information

iii Sufficient but unordered information

iv Sufficient and ordered information

Once ordered, the sets can be expanded or remodelled by the individual. Experts in a subject have highly ordered sets with many cognitive organizers, whereas people starting work in a new field have very few organizers and hence very unordered, as well as incomplete, sets of knowledge.

Following on from the ideas of Harmon, a model of users information needs and relevance decisions is proposed.

This model consists of a representation of information storage in the users memory (his knowledge structure) together with various sources of information, and the processes involved in recognizing information needs and making relevance judgements. The model is represented in Figure 5.

The new information that is received by the user can come from a wide variety of sources; these include reading the literature, informal discussions, chance finds and inspiration. The knowledge structure that a person has relating to a particular topic is stored in LTM and is just one of a large number of structures concerned with different subjects and situations. Information within a particular knowledge structure is divided into various classes with relationships between the classes. There is also within the structure means whereby new information can be assigned to the appropriate classes, and concept organizers which group the separate classes into a structured knowledge of the particular subject. Briefly the knowledge structure consists of various categories of information and organizers which provide connections between the categories and which also assign new information to the correct class. These two distinct features of the knowledge structure correspond approximately to Harmon's sets and ordering of sets. The organization of knowledge within the memory is affected by personal biases and characteristics, learning strategy and background knowledge (that is knowledge already in LTM) amongst other factors.

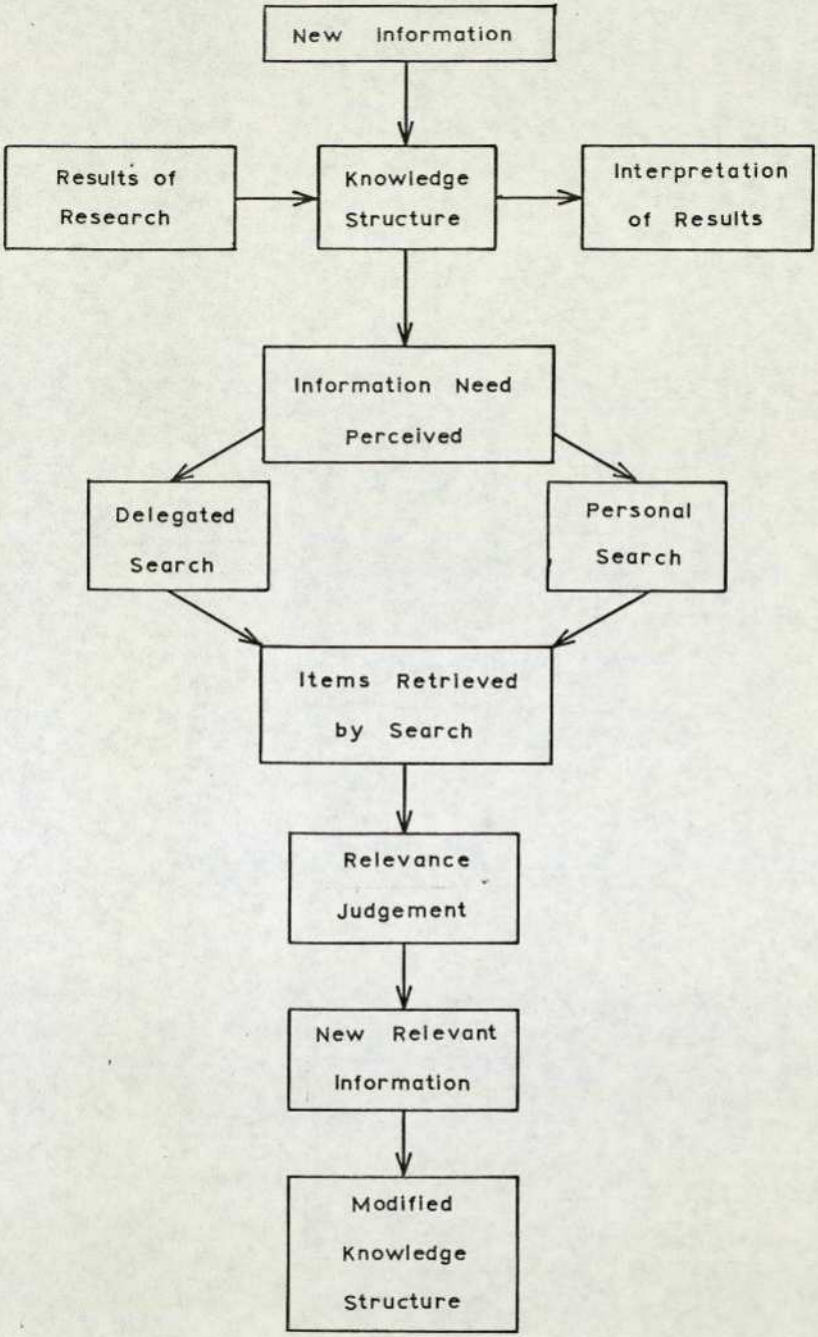


Figure 5

Model of information seeking and acquisition

The users' ability to assess the relevance of documents to his information requirements depends very largely on the degree of order in his knowledge structure. When he is first starting in a field of study the knowledge structure is almost totally unordered and contains very little information. Most of the order and content derives from other structures within his long term memory. At this point the person finds it very difficult to assess what is relevant and what is not. Almost any information that is received leads to an increase in information content of the knowledge structure and browsing is important. As his knowledge of the subject increases, the person gains the ability to classify new information and to build up an organization of the classes, so the whole knowledge structure becomes increasingly ordered. This is an intermediate stage of development and it is at this point that relevance can be recognized in most cases fairly easily by the user, and relevance judgements have a degree of objectivity. People who have a great deal of knowledge and experience of a field have highly ordered knowledge structures. They are very capable at assessing the relevance of documents to their information need. This does not necessarily mean that several experts in the same field will agree in their relevance judgements, as each expert has a unique and highly personal knowledge structure. The knowledge structures of experts in a field remain fairly stable with only minor modifications, unless there is a major revolution in their subject. (For example the introduction of quantum theory in physics). The process of acquiring and sifting information,

and ordering of the knowledge structure in continuous.

When making relevance judgements, particularly on titles, users often respond very rapidly. These judgements are made on the basis of associations between words in the title and words in LTM. In the cases where the title does not trigger an association response, the relevance judgement takes longer to make. In this latter case the words or meaning of the title is transferred from the sensory register to STM while a search is made of the LTM (i.e. the particular knowledge structure). The searching of LTM continues until a decision is made on relevance or that further searching is not practicable. In a few instances a rapid relevance judgement is made which is then modified by further information received. For example, an item is judged relevant on the title, but the decision is reversed on discovering that it is written in Russian. This interaction between the knowledge structure and the relevance decision process is continuous. It in fact is a feed back system, where changes in the knowledge structure affect relevance judgements and judgements made affect the knowledge structure and produce a slight modification in it.

A user's specific information need depends to a large extent on the amount of organization in his knowledge structure. If this structure is highly ordered he will have a quite specific or precise information need. He may have difficulty verbalizing this need, but he knows what he wants. Users

with relatively unordered structures have much wider information needs, mainly because they are not sure what they really want. Users with a fairly well ordered knowledge structure generally require specific types of information. These include

i Information to fill gaps in knowledge

ii Latest information in the field (current awareness)

iii Information in a related field which may be applicable

iv Information of general interest

There are also differences in information need when retrospective and current awareness information is sought.

Once the user has perceived a specific information need he has to formalize it either as a question for a delegated search, or as a strategy for a personal literature search. The process involved in asking a question is an evolutionary one in which several steps can be identified. Taylor [1962] has suggested four levels of question formation.

i The actual, but unexpressed need for information. (visceral need)

ii The conscious within brain description of the need. (conscious need)

iii The formal statement of the question (formalized need)

iv The question as presented to the information system (compromised need)

These stages represent a transition from the psychological to the logical interpretation of information need. The visceral need, which is subconscious, represents gaps and uncertainties within the knowledge structure. Once these omissions have been recognized by the user he has a conscious need, this corresponds to the perceived information need of the model. The actual question asked by the user relates to both his perceived information need and his knowledge structure; it also depends on his ability to verbalize the need. In the case of delegated searches (either to another person or to a retrieval system), there is usually some negotiation of the question. The user has to compromise his information need in order to achieve good communications with the information source.

Belzer [1974] suggested that a user is never sure that his query represents his information need. When two people put the same query to an I.R. system they do not necessarily want the same response as their information needs and knowledge structures are different. This must be realized when trying to assess relevance in the evaluation of I.R. systems, hence one should only employ user evaluations.

2.3 Applications to I.R. systems

The proposed model incorporates activities which can be

observed and activities which take place within the mind of the user. The manifest activities, question asking or search delegation and relevance judgements, have been extensively studied particularly in relation to the evaluation of I.R. systems. Almost all of these evaluations have ignored the mental processes behind these activities; although some consideration has been made of information needs.

The model should be of assistance in the construction of SDI profiles, as it indicates some of the processes involved in arriving at a query or information request statement. The important points are the information need which prompts the query, the user's ability to translate this need into a formal question and, central to all of these activities, how the relevant information is stored and ordered in the memory. It is important to realize that the knowledge structure is constantly being modified. These modifications are usually only very small and do not affect the overall information need. However at certain times the knowledge structure is changed to the extent that information need is markedly different than previously. In these circumstances the SDI profile no longer reflects the user's information needs and should be amended. This dynamic situation is rarely appreciated either by the user or by information scientists, and SDI profiles are not amended frequently enough to correspond to changes in information need.

3 METHODOLOGY

There are three main aspects to this research, firstly the investigation of relevance judgements made by actual users of a current awareness information retrieval service. The second aspect is the postulation of variables or factors which may affect the ways in which relevance judgements are made and the collection of data on these variables from the users. The third aspect is an attempt to find out if there are any correlations between the relevance judgements made and the postulated variables. This latter aspect requires the calculation of a continuous measure for relevance so that a single figure can be given to the relevance judgements made by each user.

Two different measures of relevance have been used. These are the conditional probabilities of hits and false alarms and the sensitivity and bias towards relevance measures obtained by applying signal detection theory (SDT). Conditional probabilities have been used by Kent et al [1967] Saracevic [1968] and by King and Bryant [1971] in assessing relevance. The signal detection methods have not been applied to relevance judgements before, although Swets [1963] has applied SDT to the evaluation of information retrieval systems.

3.1 Users

The sample of users in this study was drawn from universities, polytechnics and government institutions

in London. It was necessary to keep within the London area to reduce the travelling time involved as all contacts with the users were by interview. It had been hoped to include users from industry in the sample, however this was not possible as all of the companies approached were unwilling to co-operate.

Most of the sample were lecturers and research students in the chemistry and physics departments at City University. These people were easily accessible and, it was thought, constituted a representative cross section of research chemists at British universities. The physicists in this sample were working in areas which overlapped with chemistry. The size of the sample was increased by including lecturers from two polytechnics and chemists from two government institutions. The sample size was very small (only 31), this was mainly due to the length of time (five to six months) that was required to study each user. This meant that finding volunteers who would be available over this period of time was difficult and that the number of users who could be dealt with by one researcher in the period of time available was limited.

The prospective users were initially interviewed to explain the project and to see if they were prepared to co-operate. A detailed description of the volunteers' research work and other interests was obtained, in order to build up SDI search profiles. It was possible to

combine the interests of several users into one SDI profile, in the cases where several people were working on similar projects. A brief outline of the users research interests are given in Table 1. Fortunately none of the users dropped out during the experiment.

3.2 Material for relevance judgements

Small amounts of SDI output were obtained for each profile from either UKCIS or INSPEC. The output consisted of three or four consecutive searches against Chemical Abstracts Condensates (CAC) or physics abstracts data bases. In both cases the output given to the users consisted of titles, bibliographic citation and keywords. There were three profiles from INSPEC and six from UKCIS. For four profiles UKCIS Macroprofiles were used. These have only titles and citation, no keywords are added to the output although the profiles are searched against CAC. Three successive issues of Macroprofiles were given to the users for relevance judgements.

Prifiles for the INSPEC searches were drawn up by the INSPEC staff using information from the user regarding his research interests together with a few highly relevant papers. UKCIS profiles were constructed by the author based on the users stated information needs, then sent to UKCIS for amendment where necessary and searching against CAC. In the cases where Macroprofiles were used, the appropriate issues were requested from the list of

TABLE 1

Users research interests and position

USER	INTERESTS
AB lecturer	Adsorption of low energy alpha particles in liquids. Factors affecting the stopping power of liquids, solids and gases.
AC lecturer	Adsorption of low energy alpha particles by gases, specifically unsaturated molecules.
AD student	Surface forces and interactions in crystal growth and epitaxy. Theoretical predicition of crystal orientation.
AE student	The conformation of molecules which act on the nervous system, using molecular orbital calculations.
AF student	Structure activity relations of small organic molecules especially GABA in the nervous system. Neurotransmitters and their analogues.

AG lecturer	Synthesis of better GABA antagonists than bicuculline. Neurotransmitters and antagonists.
BC student	Analysis of diesel fuels and diesel exhaust fumes. Techniques such as GC and GC/MS for this purpose.
BD lecturer	Precise pH measurement. Ion-sensitive electrodes. Relation of structure formation constants in complexes.
BE government institution	Automation of analytical techniques and applications of computers in chemical analysis
BF government institution	Automation of analytical techniques and application of computers in chemical analysis
BG government institution	Automation of analytical techniques
BH government institution	Automation of analytical techniques, gas chromatography

CD research fellow	Iron-air and other metal-air secondary batteries. Porous electrodes.
CE student	Reduction of oxygen using graphite electrodes. Decomposition of hydrogen peroxide using cobalt ferrate.
CF student	Catalysts especially Ni and W for use in electrochemical oxidation.
CG student	Electrochemical oxidation of carbon monoxide to carbon dioxide. Current sensing devices.
CH research fellow	Regeneration of zinc. Metal-air and Ni-Fe batteries. Industrial electrochemical processes.
DE research fellow	C^{13} spectra and lanthanide shift reagents in the NMR spectra of heteroaromatic and some aromatic systems. The chemistry of quinolines.
DF research fellow	Kinetics and mechanisms of aromatic sulphonation.

DG student	Nitration of aromatic compounds in acid media.
DH student	Kinetics and mechanisms of aromatic nitration using N_2O_4 and N_2O_5 Also nitrosation.
DJ lecturer	Kinetics and mechanisms of aromatic nitration and sulphonation.
DK government institution	Tea, coffee and cocoa from harvesting onwards. Volatile constituents of foods, mainly tropical. Changes in flavour on storage or processing.
EF student	The effect of metal compounds on combustion and degradation of polyolefines. Thermal analysis.
EG student	Kinetics of gas phase combustion of hydrocarbons especially styrene. Analysis of reaction mixtures by G.C.
EH research fellow	Burning and burning mechanisms in plastics. Flame retardants and fire proofing of polymers and textiles, expecially polyurethane foams.

EJ research fellow	Kinetics and mechanisms of polypropylene oxidation and related polymers. Carbonization of polypropylene.
FG student	Postcuring mechanisms of epoxy adhesives. Physical properties of high temperature cured epoxy resins.
FH lecturer	Properties and structure of epoxy resins cured at high temperatures. Creep of epoxy resins.
FJ lecturer	Very wide interests in epoxy resins
FK lecturer	Adsorption and desorption of volatile fluids by epoxy resins. Structure and performance of epoxy resins.

subjects available. The profiles for each user and their sources are given in Table 2.

Relevance judgements were made on titles and keywords (or just titles) and on abstracts. Abstracts corresponding to the titles retrieved by each profile were photocopied and given to the users for relevance assessment. In cases where a large number of items were retrieved, a selection of abstracts was given. This selection contained all of those items judged relevant on the basis of titles, together with a random selection of irrelevant items. A total number of abstracts not exceeding about 70 were given, usually in two separate sessions.

The consistency of relevance judgements based on titles (T_1) was checked by giving the users the titles again (T_2) Titles corresponding to the abstracts seen were given a second time about three months after relevance judgements were made on the last set of abstracts.

3.3 Relevance judgements

Relevance judgements on titles and abstracts were obtained from each user at an interview. The only exception was the judgements on abstracts given by users at a government institution, where the abstracts were returned by post. The total number of interviews with each user was five or six depending on whether there were three or four SDI outputs.

TABLE 2

Profiles used and their source

PROFILE	SOURCE	USERS
Alpha particles	INSPEC	AB, AC
Crystal growth	INSPEC	AD
Molecular orbitals	INSPEC	AE
Neurotransmitters	UKCIS	AF, AG
Automated analysis	UKCIS	BE, BF, BG, BH
Nitration and sulphonation	UKCIS	DF, DG, DH, DJ
Combustion	UKCIS	EF, EG, EH, EJ
Flavours	UKCIS	DK
Epoxy resins	UKCIS	FG, FH, FJ, FK
Gas Chromatography	Macroprofile	BC
Analytical electrochemistry	Macroprofile	BD
Electrochemical reactions	Macroprofile	CD, CE, CF, CG, CH
NMR-Chemical aspects	Macroprofile	DE

The users were asked to make their relevance decisions using the three point scale adopted by Leggate et al [1973a]. These were defined as follows:

i	MAJOR value	A reference directly related to your research interests, which you will read as soon as possible.
ii	MINOR value	A reference which, although not vital is of some interest or relevance i.e. You are glad to be notified of it.
iii	IRRELEVANT	A reference of no interest at all.

These definitions were shown to each user before he examined the first set of titles, and again at subsequent interviews if a reminder was required.

As well as giving relevance judgements, the users were asked to indicate why they regarded particular references as relevant, and also to indicate if they had seen any of the references before. Most of the users quoted a word or words in the title or keywords as the reason for choosing a particular reference. Precision values were calculated for the titles and abstracts given to each user.

By looking at the reasons given for selecting a reference as relevant and examining the information needs of each

user, each positive relevance judgement on titles was placed in one of five categories. These categories were:

i Obvious from the information request statement

ii A vaguely worded title (hence a hopeful decision)

iii Decision using prior knowledge of the subject

iv No apparent reason, or interests of a colleague

v A hopeful decision (it might contain something of interest)

The assignment of relevance judgements to these categories was purely subjective. On the whole it was not difficult to decide in which category to place the judgements. Changes in relevance judgements between titles and abstracts and between the two groups of titles (T_1 and T_2) were examined with respect to the decision categories assigned to the relevance judgements on the first set of titles.

Conditional probabilities of hits and false alarms were used as measures of relevance. They were obtained from the relevance judgements on the first set of titles and abstracts using the following formulae;

$$\text{Hits} = \frac{\text{No. relevant on titles and abstracts}}{\text{Total no. relevant abstracts}}$$

$$\text{False alarms} = \frac{\text{No. relevant on titles but irrelevant on abstracts}}{\text{Total no. irrelevant abstracts}}$$

Hit and false alarm probabilities were also calculated from relevance judgements on the two groups of titles (T_1 and T_2). The equations are

$$\text{Hits} = \frac{\text{No. relevant on } T_1 \text{ and } T_2}{\text{Total no. relevant on } T_2}$$

$$\text{False alarms} = \frac{\text{No. relevant on } T_1 \text{ but irrelevant on } T_2}{\text{Total no. irrelevant on } T_2}$$

In calculating the conditional probabilities, no distinction was made between MAJOR and MINOR relevance, references were regarded as being either relevant or irrelevant.

The use of signal detection theory to calculate the measures of sensitivity to relevance and bias is described in section 5.

The use of keywords to retrieve items in the SDI output was examined. A distinction was made between their use to retrieve all items and to retrieve relevant items. It was not possible to do this for the Macroprofiles as keywords are not printed on the output, although it was obvious in some cases that items had been retrieved using the keywords in the CAC data base.

The frequency of occurrence of profile terms was noted for relevant items and for all of the items retrieved. The

most frequently occurring terms in the relevant references were examined to see if they had been specifically mentioned by the user or if they had been added during profile construction.

3.4 Variables and statistics

3.41 Variables

The variables being studied are described in section 6. Detailed information on the variables being examined in this project was obtained from a questionnaire given to each user. All of these questionnaires were eventually returned and the data from them coded and put on computer. Further details that were required and clarification of unclear replies was carried out at a subsequent interview.

It was thought that the stage of research that a person is at has an effect on the relevance judgements made. At each interview the user was asked at what stage he was in his research work. Five stages were identified, these were:

i Preliminary work

ii Experimental work and data collection

iii Analysis and interpretation of results

iv Writing up of research

v Supervisor or section leader

It was necessary to add the final category of supervisor or section leader as these people are often involved in several research projects which may be at different stages. Some supervisors are not actually engaged in experimental work and therefore do not fit into any of the other categories.

3.42 Statistics

Most of the statistical tests being used in this study are non-parametric. This type of test does not make any assumptions about the distributions of the variables being examined. It was felt necessary to use these tests as the sample size is very small and the distributions of the variables are unknown, also most of the variables are discrete rather than continuous. Most parametric tests require a high level of measurement (usually at least interval level), the variables being studied are mainly of only ordinal level of measurement. [Siegel 1956]. The only parametric test used was the Pearson or product moment correlation co-efficient and it was used only for the continuous variables.

Distributions of the measures of relevance (conditional probabilities and sensitivities) were studied using rankit plots [Colquhoun 1971]. The probabilities were ranked in ascending order and the rankit corresponding to each rank was looked up in tables. Each probability and its logarithm was then plotted against its rankit.

Data from the questionnaire was analysed using a computer package of statistical programs called 'Statistical package for the social sciences' (SPSS) [Nie et. al. 1975]. Several programs in this package were used to see if there were any correlations between the variables and the measures of relevance.

Means, standard deviations and various other distributional characteristics of the continuous variables were obtained using the SPSS subprogram CODE BOOK. Relationships between the variables were tested using the CROSSTABS subprogram, which performs cross-tabulations and calculates chi-squared values. The subprogram NONPAR CORR was used to calculate Spearman rank correlation coefficients between the variables. An attempt was made to reduce the number of variables by clustering the related variables using factor analysis. This is fully described in section 7.

Correlation coefficients between the variables and the relevance measures were calculated using the subprograms PEARSON CORR and NONPAR CORR. PEARSON CORR was used to calculate Pearson correlation coefficients for only the continuous variables, NONPAR CORR was used to calculate Spearman ranked correlation coefficients for all of the variables.

There was only one non-parametric test in SPSS, so the remaining statistical tests had to be calculated

manually. The Mann-Whitney U test was used to see if there were any statistically significant differences between the two categories of users within each dichotomous variable (for example between users doing pure or applied research). The Mann-Whitney test is one of the most powerful non-parametric tests and provides a useful alternative to the t-test, when the assumption of normal distributions cannot be made [Siegel 1956].

For variables with more than two categories, the Kruskal-Wallis one way analysis of variance was used to see if there were any significant differences between the categories [Siegel 1956]. For example to see if there were any significant differences between the measures of relevance for users with BSc, MSc or PhD. Both the Mann-Whitney and the Kruskal-Wallis tests use the rankings of relevance measures rather than the actual values calculated.

4 QUESTIONNAIRE

4.1 Questionnaire design and distribution

The most widely used survey techniques in user studies have been interviews, questionnaires and diaries. Comparisons of the results obtained using these various methods have indicated that they all produce similar findings. The final choice between the techniques depends on the reasons and objectives of the study and on the population and sample size [Jahoda 1965].

There have been problems and disadvantages with all three methods. Diaries require a great deal of time to fill in and are limited in the type of information collected. Interviews are fraught with problems of bias, but they are very flexible and one can ensure that the respondent understands the questions being asked. Questionnaires do not have such obvious problems of bias, but care must be taken in the wording of questions. The main disadvantages of questionnaires are lack of response and lack of flexibility [Oppenheim 1968].

All survey techniques suffer from the 'spotlight effect'. This refers to the change in behaviour of a subject when he knows that he is being observed. The distortion is generally less marked in prolonged surveys [Martin and Ackoff 1963]. This also affects the truthfullness of the answers given. A similar problem which one must be aware

of is the natural desire of respondents to present themselves in a favourable light. This may give rise to biased responses which are very difficult to identify or check.

The interpretation of questions, particularly in questionnaires, has according to Herner and Herner [1967] been a regrettable feature of surveys into information needs and uses. Even the word 'information' itself has had several interpretations by different scientists.

The purpose of the present questionnaire was solely to obtain information from each user relating to the variables under investigation. It was decided to use a questionnaire rather than an interview, as the information could be collected in standardised form which would be easier to code and analyse.

The topic of questionnaire design was not pursued in great depth as the main purpose of the present questionnaire was to collect information from a small and easily accessible sample. However the design of questionnaires is very important in large scale postal surveys. A detailed description of questionnaire design has been given by Oppenheim [1968].

The present questionnaire was designed with the following points in mind:

i To use mainly precoded questions which facilitate the coding of responses

ii To avoid biased questions, and to avoid influencing the responses of the user. /This is a greater problem in opinion than in factual surveys/

iii To make the questions straightforward and easily understandable

iv To put the questions in a logical sequence

v To keep the number of questions to a minimum and to avoid asking unnecessary ones

There were two parts to the questionnaire. The first was designed to collect factual and biographical information, the second part related to opinions and attitudes towards scientific information and information sources. The layout of the questionnaire is given in Appendix A1.1

The questionnaire was given to each of the 31 users personally, usually after the interview at which relevance judgements on the second SDI output were obtained. Some users filled in the questionnaire immediately, others completed it later and returned it either by post or at an interview. If there was a long delay in returning the

questionnaire the user was reminded about it. By, in some cases, repeated requests for the return of the questionnaires a 100% response was obtained.

Any questions which were not completed and any queries, were clarified on a subsequent visit to the user. This eliminated many of the problems associated with the interpretation of questions. Because of the high degree of personal contact involved the present survey falls somewhere between a structured interview and a questionnaire.

4.2 Results and discussion

This survey cannot be regarded in the same light as a postal survey because of the high degree of personal contact with the sample population, the limited aims of the survey and the small sample size. However it is of interest to compare the results obtained from this questionnaire with those reported in the literature. There is such diversity in the types of survey, the samples used and the way questions are worded, that comparing results from different surveys is difficult. Providing that one is aware of the problems it is quite possible to compare surveys where the samples were similar and where similar questions were asked.

The sample size was very small (only 31) and consisted of 12 PhD students, 8 university or polytechnic lecturers, 6 research fellows and 5 researchers in government

institutions. The average age was 33.2 years with a range of 21 to 63 years, only one of the sample was female. The length of time spent in the field varied between 6 months and 25 years with a mean of 5.8 years.

The complete results from the questionnaire are tabulated in Appendix A1.2. In the following sections some of the results obtained will be described and compared with results from other surveys.

4.21 Current awareness

The time spent per week on information gathering varied between 0.5 hours and 20 hours, with a mean of 3.5 hours. This average figure was lower than that obtained in most surveys of total literature usage. Bernal [1948] found an average of 5.3 hours per week, Hanson [1964] in a review of 6 surveys quoted figures of 4½, 5, 5, 4½, 4 and 9 hours per week. This last value was unrepresentative as it was for Scandinavian scientists who were reading English which was for them a foreign language.

The figure of 3.5 hours obtained was somewhat nearer that found for the time spent actually reading. Kean [1973] found that chemists at City University spend, on average, 2½ hours reading scientific literature, which was similar to the value of 2.7 hours for chemists in the study by Martin and Ackoff [1963]. Physicists also spend about 2½ hours per week reading scientific literature [Hall, Clague and Aitchison 1972].

Of the 12 PhD students in the present survey half spent less than 2½ hours a week on literature activities. There were 6 users (19.4%) who spent more than 5 hours a week information gathering, two were section leaders in government institutions, one was writing a review, one was a student doing a literature review for his thesis and the remaining two were university lecturers. The majority of scientists in this sample who spent 2½ to 5 hours a week were at the experimental stage of their research.

There was a slight tendency for older scientists to spend longer on literature activities. However this may be related also to seniority or position. There was a definite relationship between the time spent information gathering and the amount written in the persons' field. Only 2 of the 13 people who spent less than 2½ hours a week felt there was a lot of literature in their field. Those who thought there was a great deal written in their field (12, 38.7%) spent an average of 5.4 hours a week compared to the average of 2.3 hours for the remaining 19 people.

No correlation was found between the amount of time spent on literature activities and speed of reading or subject area. This latter result is in agreement with the findings of Hanson [1964].

No great significance should be attached to the actual figures obtained as scientists often give estimates of time spent based on guesswork or on how long they would like to spend reading the literature.

In the present sample the number of journals scanned regularly varied between none and 10, with a mean of 4.7. This again was a rather low figure compared with other surveys of scientists. Most surveys have suggested that an average of 10 journals were scanned regularly [Meadows 1974, Thomas 1968]. Slightly lower figures were obtained by Bernal [1948] (7.7 journals) and Hall, Clague and Aitchison [1972] (8 journals).

Scientists in government institutions scanned more journals (mean 6.4) than academics (mean 4.3). This was in agreement with the view of Meadows [1974]. Students tended to scan fewest journals (11 of the 12 students scanned less than 5 journals) but there was no real correlation with age. No correlation was found between the number of journals scanned and speed of reading or the amount written in the field.

Not all of the 28 users who scanned journals regularly listed the ones they looked at. Despite this a total of 70 journals were mentioned. Chemical Abstracts was scanned regularly by the largest number of people (6). The Analyst and the Journal of the American Chemical Society were each

scanned by 5 people, 4 journals were each scanned by 4 people. From the total of 70 journals, 40 were listed by only one person. A list of journals mentioned by at least 2 people is given in Appendix A1.2.

4.22 Foreign languages

The majority of the present sample (23, 74.2%) could understand papers written in French, but only 14 (45.2%) could understand German. This was similar to the figures given by Bernal [1948]. A further 5 people had a slight knowledge of German and one had a slight knowledge of French. (i.e. they could manage to read the title and possibly the abstract of a paper). If these were included, the overall figures for French 15 (77.5%) and German 19 (61.3%) were similar to those given by Wood [1967]. He found that 90% could handle French and 66% could handle German. In the present sample 11 people (35.5%) could read both French and German. Only 2 people (6.4%) were unable to cope with any foreign language, this is a considerable improvement on the high figure found by Bernal [1948].

Despite the fact that Russian is the second most commonly used language in the chemical literature [Wood 1967], under 10% of the sample had even a slight knowledge of it. The situation was even worse for Japanese,

where only one person (3.2%) had any knowledge of it and he was Chinese. The results of the present survey correspond well with those of Wood [1967], who had a much larger sample drawn from several scientific disciplines.

The limited language ability of most scientists has been well documented [Bernal 1948, Wood 1967]. The aim of this question on accessibility of language ability was to see if scientists felt that, although they could not cope with many languages themselves, there was somebody accessible who could help them with any potential language difficulties.

Nearly a third of the sample (10, 32.3%) had no access to language ability, this included the 4 members of the group studying nitration and sulphonation and also included 6 students. The general competence in French and to a lesser extent German was reflected in the low figure of 8 people (25.8%) stating they had access to these languages. A similar number of people (9, 29%) had access to Russian. This figure was quite high when one considers that less than 10% of scientists had any knowledge of Russian. It was probably explained by the presence of a Czech chemist at City University who could read Russian well. Among the other languages mentioned as being accessible were Czech, Chinese, Japanese, Polish and Spanish.

There was a correlation between the access to language expertise and the size of work group. Three of

the 5 people who worked alone had no access to languages, whereas 3 of the 4 people working in a large group had access to more than one language. This indicated both the greater range of linguistic ability within the group and also the greater number of external contacts that the combined group had. It was also found that people with no access to language ability tended not to use libraries outside their own institution.

4.23 Formal and informal sources

Written sources of information alone were used by 54.8% (17) of the sample, verbal sources alone by only one person (3.2%) and both written and verbal sources by 13 people (41.9%). These figures were the opposite way round to those of Kean [1973], who found in a sample of scientists at City University that 12% preferred written sources and that 40% used both written and verbal sources.

Several surveys have shown that pure scientists use written sources more than applied scientists [Wood 1971, Meadows 1974, McAlpine et al 1972]. In the present survey no correlation was found between sources used and pure or applied research or with experimental or theoretical research. However all the scientists were doing research of some kind, so it was likely that there were no truly applied scientists in the sample.

There was no correlation between type of source and age as was found by Meadows [1974], but 9 of the 12 PhD students used written sources only. This indicated that they had not been incorporated into an informal network of communication.

Most of the present sample had contacts outside of their own institution (21, 67.7%). A larger number of the sample (25, 80.6%) were used as an informal information source by other people, both internally and externally. Of the 10 people with no external contacts 7 were students, and 5 of the 6 people not used as an information source were students. People with extensive experience of information gathering (8, 25.8%) all had external contacts and were all used as information sources. No correlation was found between external contacts or use as an information source and size of the work group.

The people without external contacts were generally not used as sources of information; they tended not to use external libraries, to use only written sources, to be at the experimental stage of research and to have spent less than two years in their field of research. They scanned fewer journals, spent less time on literature searching and to use abstracts journals rarely.

4.24 Literature searching habits

Interlibrary loan (ILL) was used quite frequently

by the scientists in the present survey; 18 (58.1%) used this service monthly or more, whilst only 9 (29%) used it rarely or never. There was a correlation between the frequency of use of ILL and the amount of material written in the field. Of the 9 users who felt that there was a great deal written in their field, 7 used ILL at least monthly. In addition 5 of the 6 people who had delegated searches used this service more than monthly.

All of the sample made photocopies in connection with their work; 11 (35.5%) made copies more than once a week. These results were similar to those of Bell [1973], who found that 98% of his sample made photocopies and the most frequently used mode was more than weekly. Of the 14 people (45.2%) who photocopied weekly or more only 2 were students. This reflected the stricter control that university departments generally exercise on the amount of photocopying done by research students.

Most of the sample had used an abstracts journal within the previous 3 months, although a rather large number (9, 29%) had not. This figure was somewhat higher than the 10-20% suggested by Meadows [1974] as the rate of non-usage of abstracts journals. Twelve people (38.7%) had used abstracts occasionally while only 5 (16.1%) had used them frequently. Of these 5, 2 had been doing large scale literature searches.

The journal scanned regularly by most scientists in the present sample was Chemical Abstracts. This may have been because some scientists used the abstracts journal as a substitute for the primary literature, as has been suggested by Meadows [1974].

There was a tendency for those who used ILL and an abstracts journal frequently to make photocopies frequently. The reverse of this was true for most of the research students in the sample.

A large proportion of the sample (22, 71%) used libraries outside of their own institution. This reflected the easy access to large and specialized libraries within London. Kean [1973] in a study of lecturers at City University found that only 34% used any libraries outside the university. There was a tendency for the users of external libraries to spend more time on literature activities, than non-users.

The most frequently used libraries were the Science Reference Library (SRL) and the Chemical Society library. The SRL was used by 16 people and the Chemical Society library by 10 people. Kean [1973] found that 42% of her sample who used external libraries used the SRL.

Many surveys of the literature searching habits of scientists have found that they do not delegate searches

or consult librarians [Meadows 1974, Kean 1973]. This was also found in the present survey, only 6 people (19.4%) had ever delegated searches. Only one of these was a student, which possibly bears out the unfavourable attitude of students towards librarians found by Slater and Fisher [1969]. Scientists do not delegate searches either because they do not feel that a third party is competent to do the search or because they are unable to explain sufficiently clearly what they require. Few people (5, 16.1%) in the survey had used an SDI service before.

Almost all of the scientists in the present sample used Chemical Abstracts (25, 80.6%). These people were asked which indexes in Chemical Abstracts they had used. The subject and author indexes only were used by 14 people. Three of the organic chemists used the formula index and 7 people had used the index guide. The remaining indexes were very rarely used if at all, and then only by a very small number of scientists.

4.25 Attitudes towards information

There was a fairly even distribution of experience in information gathering in the present sample. Only 9 people (29%) felt they had slight experience, 14 (45.2%) felt they had moderate experience and 8 (25.5%) had extensive experience. All of those who had spent less than one year in their field had slight experience, whereas 5 of the 8 with

extensive experience had spent 6 years or more in the field. None of the people who had only slight experience delegated searches.

The scientists approach to information seeking was investigated by asking if they looked at all references of possible interest (ABNO), only looked at references of obvious interest (OBNA) or compromised between these two alternatives. Most of the sample (12, 38.7%) looked only at obviously interesting items. Half of these people were involved in both teaching and research, so they had to be more selective in their reading because of pressure of work. Only 2 of those with the OBNA approach were students. The compromise approach was adopted by 10 people (32.3%) 7 of whom were at the experimental stage of their research at some time during the investigation. The members of this group were probably just doing routine scanning to keep up to date. The section leaders in government institutions looked at everything of interest. This is not surprising as they have wider interests than most of the academics and they are responsible for the literature searching in their section. The other scientists who looked at everything of possible interest did so for a variety of reasons. Two were about to start writing their PhD theses and two found very little written in their field of research.

The sample was fairly evenly divided on whether they spent enough time on information gathering activities. Slightly less than half (13, 41.9%) felt they spent enough time, this was very low compared to the 75% found in a survey of physicists by Hall, Clague and Aitchison [1972]. However Kean [1973] found that 62% of her sample would read more if they had the time available. This latter figure corresponds quite well to the 58.1% (18) of the present sample who felt that they did not spend enough time reading the scientific literature.

4.26 Research field

The present sample consisted of chemists and physicists working in several distinct subfields. These subfields were combustion, adhesion, electrochemistry, analytical chemistry and inorganic and nuclear chemistry. The majority (26, 83.9%) felt that their work was experimental rather than theoretical, and most (21, 67.7%) regarded it as applied rather than pure. All of the analytical chemists regarded their work as being experimental and applied. The organic chemists were all experimentalists and all the adhesion group were doing applied research. There weresome slight disagreements amongst members of a particular sub-field regarding the nature of their work, but on the whole most people agreed with each other.

Relevance judgements from each user were collected over a period of about five to six months. During this time

19 people (61.3%) had not changed their stage of research, the remaining 12 (38.7%) had changed the emphasis of their work. The category of section leaders and supervisors obviously remained the same over the period of study. At any one time there were more people doing experimental work than involved in data analysis or report/thesis writing. The stage of research is important as it affects the type of information required and the need for informal discussion [Meadows 1974].

The scientists in the present survey were mostly working in established fields which produced an average amount of written material. Only 3 people (9.7%) felt that there was a great deal written in their field and 3 (9.7%) felt that very little was written. In 5 cases the person felt that little was written on his specific topic, but that there was a great deal of written material in the general field. This was particularly noted by those working in the combustion field.

Most of the scientists in this survey (22, 71%) worked in a small group (with up to four other people), 5 (16.1%) worked alone and 4 (12.9%) were part of a large group (more than 5 people). Of the 5 who worked alone, 4 had been in their research field for at least 5 years.

5. SIGNAL DETECTION THEORY

5.1 Theoretical models

The process whereby a person decides whether a title or abstract is relevant to his work can be likened to the situation in experimental psychology, where a person has to decide whether a stimulus has or has not been presented. The problem of detecting stimuli (signals), especially in a noisy environment, has been studied by psychologists for many years. It is the intention in this section to apply the methods developed by psychologists to the problem of relevance, and the detection of relevance based on titles as the initial stimulus.

The applicability of signal detection theory (SDT) to relevance judgements lies in the fact that it is concerned with the decision making process involved in such judgements. SDT is particularly useful as it applies in situations where decisions have to be made with some degree of uncertainty. The most immediate benefit of SDT is that it can provide a single measure of a person's performance in detecting relevance, and also provide a measure of a person's criterion for deciding whether a title is relevant to his information requirements.

5.11 Signal detection models

The original SDT model was proposed by Tanner and Swets in 1954. It was an adaption of the theory relating to the detection of radar signals, that provided a psychophysical model of the detection of signals by human observers. The main idea behind SDT according to Luce [1963] was that:

> "pertinent information available to the observer as a result of stimulation by a signal, can be summarized by a number; however repeated presentations of the same stimulus produce not the same number but a distribution of them. The observer is assumed to behave as though he knew these distributions."

The observer calculates the likelihood of a signal occurring, if it is high he responds 'yes', if it is low he responds 'no'. Thus the observer must establish a cut-off point or criterion and apply a decision rule. This criterion depends on factors such as the probability of a signal occurring and the consequences of the decision.

There are two distinct aspects to the signal detection task

i The detection process - the actual detection of the signal

ii The decision process - deciding whether to

reply 'yes' or 'no' when a stimulus is presented.

The decision process (the choice of criterion adopted) is at the disposal of the observer and can be manipulated by asking him to apply a natural, lax or strict criterion. SDT enables these two processes to be studied either separately or co-jointly. It is assumed that a continuum of detectability, rather than a specific threshold, applies to human observers.

During a simple YES-NO signal detection task, an observer is asked to decide whether a signal (frequently a visual or auditory signal) is present when transmitted with a noisy background. A large number of trials are performed, in some instances a signal is present and in others only noise. A two by two table of signal and response is drawn up and the conditional probabilities associated with each cell of the table calculated [McNicol 1972].

Response

	YES	NO
SIGNAL	HIT P(S/s)	MISS P(N/s)
NOISE	FALSE ALARM P(S/n)	CORRECT REJECTION P(N/n)

The hit P(S/s) and false alarm P(S/n) probabilities are plotted for several different criteria, and a smooth curve is usually obtained. This curve is called the operating characteristic (O.C) for that particular observer. The area under this curve gives a measure of the detectability of the signal. If the O.C. is plotted on probability axes (double probability plot) a straight line is obtained [Green and Swets 1966].

An operating characteristic is defined as being a relation between the operating probabilities [P(S/s) and P(S/n)] of those decision rules that might be chosen, with respect to a given discrimination, by a co-operative subject [Laming 1973]. A co-operative subject is one who is 'doing his best', and not deliberately adopting a distorted criterion. The O.C. is a strict monotone between P(S/s) and P(S/n). The likelihood ratio (λ_x) is a monotonically increasing function of x if, and only if, the O.C. is convex.

The response by an observer depends on whether the likelihood ratio of a signal occurring exceeds the criterion value (X_c). Hence if

$\lambda_x > X_c$ the response will be YES

$\lambda_x < X_c$ the response will be NO

If P(S/s) equals P(S/n) the detection process is random.

The detection rate is almost always better than random so P(S/s) is greater than P(S/n).

The nature of the signal and noise distributions are generally unknown and assumptions have to be made regarding them. In the SDT model of Tanner and Swets [1954] these distributions were assumed to be normal and to have equal variances. The variance of the noise distribution was defined as being unity.

The signal and noise distributions for each person overlap, and the distance between the two means is taken as being a measure of an individual's sensitivity to a signal (d'). (see Figure 6). An individual's response bias or criterion (β) is measured by the ratio of the heights of the signal and noise distributions at the criterion point (X_c). Hence (from Figure 6)

$$\beta = y_s / y_n$$

The values of sensitivity (d') and bias (β) can be calculated directly from the hit and false alarm probabilities, after conversion to Z-units (standard deviation units for a standard normal distribution). The equations involved are; [McNicol 1972]

$$d' = z(S/n) - z(S/s)$$

$$\beta = e^{-\frac{1}{2}(x - d')^2} / e^{-\frac{1}{2}x^2} \quad \text{where} \quad x = z(S/n)$$

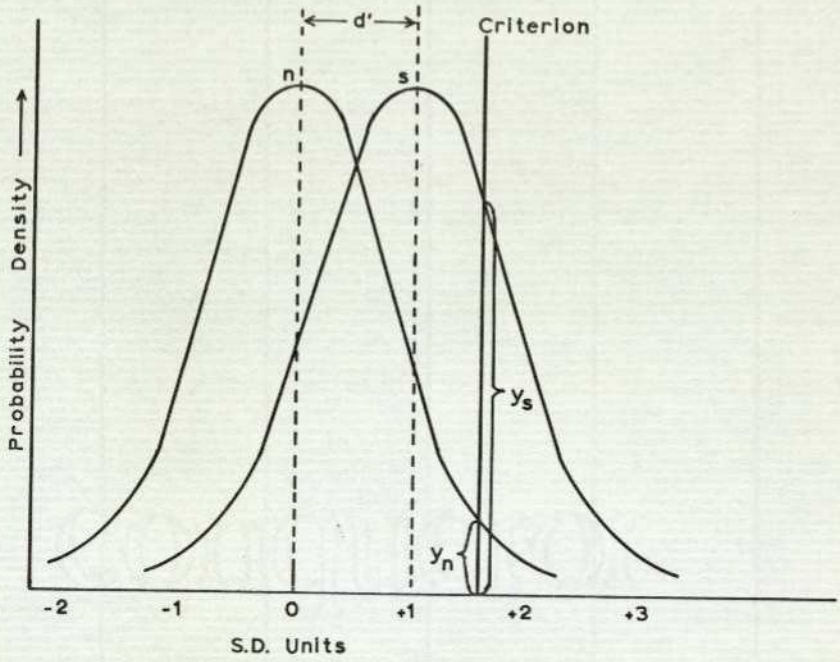


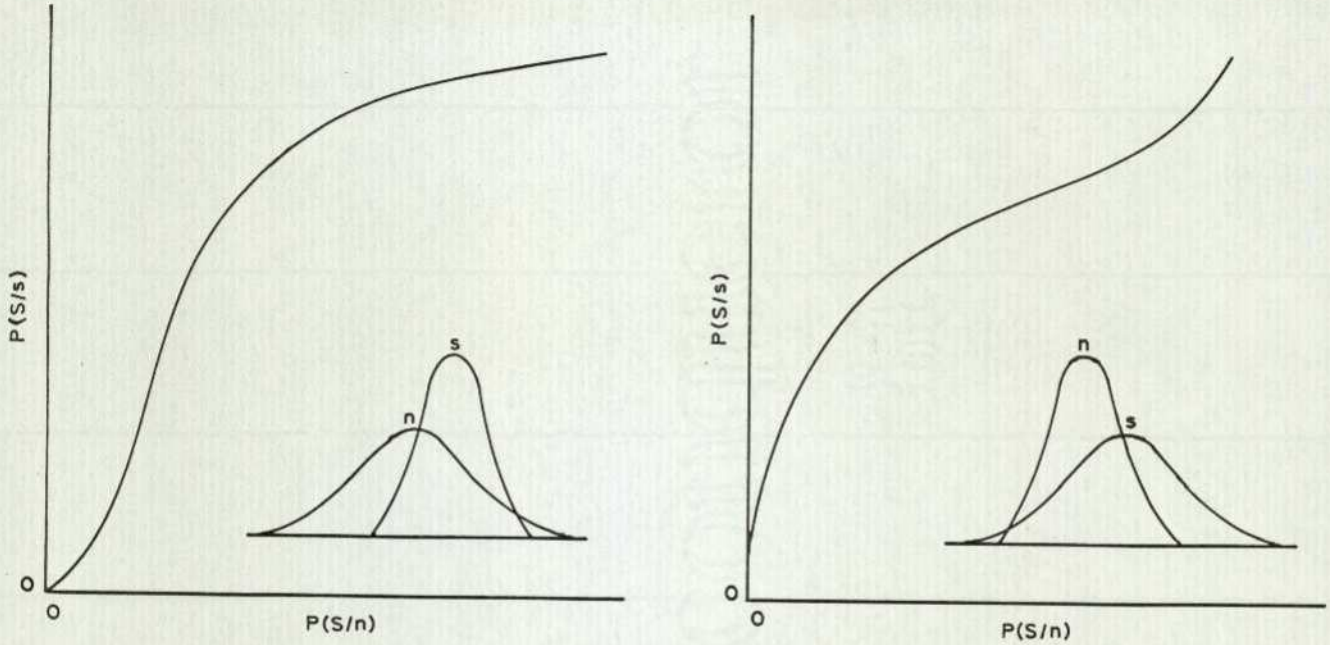
Figure 6 Normal distributions of signal and noise.

The operating characteristic can be plotted on double probability scales (i.e. probability scales on both axes). This produces a straight line graph with a slope of one, and which is symmetrical about the anti-diagonal.

Sensitivity (d') is relatively invariant, in that it has a high consistency for the same person in different types of signal detection tasks and at different periods of time. [Green and Swets 1966]. If d' is calculated assuming equal variance of signal and noise distributions and these variances are in fact not equal, the value of d' will be dependent on the value of β [Broadbent 1971].

5.12 Unequal variance SDT model

There are many cases where the O.C. is not symmetrical about the anti-diagonal. In these circumstances SDT is applied but without the assumption of equal variances of the signal and noise distributions. The distributions are still assumed to be normal. The shape of the O.C. depends on whether the signal variance (σ_s^2) or the noise variance (σ_n^2) is larger. (See Figure 7). The fact that the O.C. is not symmetrical may be due to several factors including different signal and noise distributions. A linear O.C. when plotted on double probability scales can be obtained when the signal and/or noise distributions are exponential, poisson, Rayleigh or various gamma functions [Green and Swets 1966].



i $\sigma_s < \sigma_n$ ii $\sigma_s > \sigma_n$

Figure 7 O.C. Curves for unequal variance cases.

The use of this unequal variance model is based on practical rather than theoretical considerations. If the variances are not equal, the likelihood ratio is not monotonic with x. For this reason the use of this model cannot be justified theoretically, despite its widespread application to experimental data.

The slope (s) of the O.C. plotted using probability scales is not unity, but is equal to the ratio of the signal and noise variances.

$$s = \sigma_n^2/\sigma_s^2$$

Measures of sensitivity and bias have to be obtained from the linear O.C. Two measures of sensitivity for the unequal variance model have been proposed.

i The difference between the signal and noise distribution means, (Δm) [Green and Swets 1966]

ii Twice the ordinate in Z-units of the point at which the O.C. crosses the negative diagonal (d'e) [Egan and Clarke 1966]

Both of these measures are obtained directly from the linear O.C. Δm is the value of P(S/n) in Z-units at the point on the O.C. where P(S/s) = 0.5 [i.e. Z(S/s) = o]. This measure is directly analogous to d' of the equal variance SDT

model. The other measure (d'e) is twice the value of P(S/s) or P(S/n) in Z-units where the O.C. crosses the negative diagonal, the values of P(S/s) and P(S/n) at this point are equal.

There appears to be little to choose between these two measures of sensitivity. Δm gives equal weight to the signal and noise distributions; whereas d'e has the useful property of being invariant over a series of trials, despite minor changes in slope of the O.C. Markowitz and Swets [1967] regarded d'e as being superior to Δm on purely practical grounds. These two measures have been related to each other and to the slope of the O.C. by Green and Swets [1966] using the empirical formula

$$d'_e = 2\Delta m\left[\frac{s}{1+s}\right]$$

As the likelihood ratio is not monotonic with X there will be a point on the linear O.C. where the bias values (β) reverse. This usually occurs at the ends of the O.C. where the estimation of points is unreliable, as observers do not naturally adopt criteria of such extreme strictness or laxity.

Calculation of β is very similar to the method used in the equal variance SDT model. The criterion value (β) is the ratio of the heights of the signal and noise distributions, but

with the height of the signal distribution (y_s) modified to take into account the signal variance. The modified formula for y_s is: [McNicol 1972]

$$y_s = \frac{1}{\sigma_s(2\pi)^{1/2}} \exp\left[\frac{(x-\Delta m)^2}{2\sigma_s^2}\right]$$

$$= s\left[\frac{1}{(2\pi)^{1/2}} \exp(-1/2\,[z(S/s)]^2)\right]$$

The formula for the height of the noise distribution (y_n) is the same as in the equal variance SDT model

$$y_n = \frac{1}{(2\pi)^{1/2}} \exp(-1/2\,[z(S/n)]^2)$$

$$\beta = y_s/y_n$$

$$\text{Hence } \beta = \frac{s \cdot \exp(-1/2\,[z(S/s)]^2)}{\exp(-1/2\,[z(S/n)]^2)}$$

The theory and methods of SDT have also been applied to rating scale tasks, where the observer is allowed several response categories. For example, an observer might be asked to respond using one of four categories labelled

Definitely a signal

Possibly a signal

Possibly noise

Definitely noise

The use of rating scale methods could be extended to relevance judgements, where several categories of relevance are used.

5.13 Exponential model

Green and Swets [1966] proposed a signal detection model which assumed exponential rather than normal distributions of signal and noise. These distributions were represented by the density functions

$$fN(x) = e^{x}$$

$$fS(x) = ke^{-kx}$$

The operating characteristic had the equation

$$P(S/s) = P(S/n)^{k}$$

or $$\log P(S/s) = k \log P(S/n)$$

Laming [1976] regards this exponential distribution as being a special case of a general gamma function, with the following density functions

$$fN(x) = x^{v} e^{-x}$$

$$fS(x) = k^{v+1} x^{v} e^{-kx} \qquad \text{where } 0 < k \leq 1$$

The exponential model of Green and Swets is the case when

$$v = 0$$

The exponential model produces a smooth asymmetric O.C. which has a monotonic likelihood ratio. [Laming 1973]. A measure of an individual's ability to detect a signal is given by k the scaling factor [Broadbent 1971].

This model is an improvement on the unequal variance SDT model in four important respects

i The likelihood ratio is monotonic with x

ii It uses a one parameter distribution, whereas the SDT unequal variance model requires two parameters to give a complete description

iii The exponential model arises naturally in many processes, especially in counting

iv The exponential model does not require the drawing of graphs to calculate sensitivity, as this can be done from the hit and false alarm probabilities.

In the YES-NO tasks involving the recognition of words in a noisy background, the O.C. for the receiver is based on response-response rather than stimulus-response contingencies. During this type of experiment the observer attempts to discriminate between correct and incorrect responses. This is rather different from the usual signal detection task, where the observer simply has to decide whether or not a signal was presented.

The relevance judgement situation would appear to rely on response-response contingencies rather than stimulus-response contingencies. This is another reason for preferring the exponential rather than the unequal variance SDT model in the present study.

5.14 Luce's choice theory

Luce [1963] proposed a model for signal detection based on the conceptual framework of his choice theory. This theory assumed a logistic distribution of variables. Choice theories and signal detection theories, although developed from different viewpoints, resemble each other closely in many cases [McNicol 1972].

The application of this choice model to a simple YES-NO experiment can be represented by a two by two table of stimulus and response. The conditional probabilities associated with each cell are expressed in terms of sensitivity (a) and response bias (v)

		Response: YES	Response: NO
Stimulus	SIGNAL	HIT $P(S/s) = \frac{a}{a+v}$	MISS $P(N/s) = \frac{v}{a+v}$
	NOISE	FALSE ALARM $P(S/n) = \frac{1}{1+av}$	CORRECT REJECTION $P(N/n) = \frac{av}{1+av}$

The expressions for hits and false alarms were obtained from the product of stimulus matrix and a response matrix. A full derivation of these expressions can be found in Luce [1963] and Laming [1973].

The operating characteristic is obtained by plotting hit and false alarm probabilities for a series of trials, using various criteria of strictness. The O.C. however is representable only if the response probabilities are independent of the stimulus presented [Marley 1971]. The O.C. curves are symmetric about the antidiagonal, and resemble O.C. curves obtained using the equal variance SDT model. This resemblance reflects the basic similarity between the logistic and normal distributions.

The distance between the signal and noise distributions in Luce's model is given by 2 Log α. This is a measure of the observer's sensitivity to the signal and is equivalent to d' of the SDT model. If Log α is plotted against d', for values obtained from the same set of data, the graph is linear. Hence there is a high correlation between these two measures of sensitivity. [Luce 1963].

Sensitivity (α) and response bias (v) values for each observer can be calculated directly from the conditional probabilities associated with each cell of the two by two table.

$$a^2 = \left[\frac{a}{a+v} \cdot \frac{av}{1+av} \right] \Big/ \left[\frac{1}{1+av} \cdot \frac{v}{a+v} \right]$$

$$a^2 = P(S/s) \times P(N/n) \;/\; P(S/n) \times P(N/s)$$

$$v^2 = \left[\frac{v}{a+v} \cdot \frac{av}{1+av} \right] \Big/ \left[\frac{a}{a+v} \cdot \frac{1}{1+av} \right]$$

$$v^2 = P(N/s) \times P(N/n) \;/\; P(S/s) \times P(S/n)$$

where $P(N/s) = 1 - P(S/s)$

$P(N/n) = 1 - P(S/n)$

There are two important practical advantages in using Luce's choice model, rather than the SDT models, to calculate sensitivity in signal detection tasks.

i It is not necessary to assume normal distributions of signal and noise

ii a and v can be calculated directly from the hit and false alarm probabilities, hence it is not necessary to plot O.C. curves.

The main disadvantage of this model is that it is incompatible with studies showing a fair proportion of detections coupled with a low proportion of false alarms [Laming 1973].

5.15 Robertson's model

Robertson [1975a] has proposed a theoretical model of the retrieval characteristics of information retrieval systems, which bears a superficial resemblance to signal detection models for the YES-NO type of situation. This model has greatest similarity to Luce's model, as they both employ logistic distributions.

A two by two table of retrieval and relevance was used by Robertson, which is similar to the stimulus-response tables of SDT. The hit probabilities (θ_{j1}) and false alarm probabilities (θ_{j2}) were subjected to a logistic transformation and then used in a linear model. The equations involved were;

$$\text{logit}(\theta_{j1}) = a_j + \delta_j = \lambda_{j1}$$

$$\text{logit}(\theta_{j2}) = a_j - \delta_j = \lambda_{j2}$$

where $\text{logit}\,\theta = \log\left[\frac{\theta}{1-\theta}\right]$

In the original model a_j was a measure of the specificity of question formulation and δ_j a measure of the separation

between relevant and non-relevant documents. These measures can be interpreted as being response bias (a_r) and sensitivity (δ_r). [Robertson 1975b].

Various assumptions were made regarding the distributions of a and δ

i They are either a. separately normally distributed or
b. have a bivariate normal distribution

ii The variance of the δ_j distribution is zero

iii a_j and δ_j are independent.

In practice the bivariate normal distribution gave a better representation of the data obtained experimentally, as no relationship was assumed between λ_{j1} and λ_{j2} This means that the probability plots of hits and false alarms are not constrained to fall on a straight line with a slope of one, by the assumptions made in the model.

The values of a_r and δ_r for each individual case were obtained from the pairs of hit and false alarm probabilities, using an estimating procedure. The effect of this estimating process was to pull the individual probabilities towards a common mean. Tests of this model using information retrieval data showed that a logistic plot of λ_{j1} against λ_{j2} gave a straight line with a slope less than one. Theoretically the slope should equal one. Robertson [1975a] explained this as being due to a negative correlation between a and δ

The estimating (smoothing) process pulls all of the pairs of a_r and δ_r points onto a straight line. This effectively reduces the number of degrees of freedom from two to one.

5.16 Comparisons between the models

The equal variance SDT model is the most widely used, although it is not applicable to the many detection experiments where the O.C. is asymmetric. This model can be easily generalized from the normal distribution case by substituting other pairs of signal and noise distributions. One such generalization is the use of exponential distributions, in this case the O.C. does not have to be symmetrical. Another model based on SDT, which allows for an asymmetric O.C. is the unequal variance model. This model has been used quite often for the analysis of experimental data, despite its theoretical shortcomings (the likelihood ratio is not monotonic with x)

Luce's choice model is based on logistic distributions and gives rise to symmetric O.C. curves. Unlike the SDT models, it cannot be generalized to allow for asymmetric O.C. curves, although this should theoretically be possible. [Ingleby 1969].

One difficulty that arises when analysing signal detection data is the appearance of asymmetric O.C. curves. This can seriously affect the choice and application of any

signal detection model. This non-symmetry of the O.C. may be due to several factors [Ingleby 1969] including:

i In YES-NO experiments the outer points of an O.C. may really belong to a different O.C.

ii In the rating experiments, the extreme points of the rating scale are rarely used. This affects the sampling process underlying the O.C. construction.

iii Variability in placing outer criteria, as there is no agreed definition of 'certain' and 'doubtful'.

iv The signal and noise distributions may be other than normal.

This last factor is probably the most important, especially in experiments where the signal and noise distributions cannot be manipulated directly. Many distributions will give a linear normalised O.C. but with a slope not equal to one. This corresponds to an asymmetric O.C. when plotted on linear scales. Examples of this type of distribution are the various gamma functions, including the exponential distribution.

Because of the differences between the various signal detection models, both in theoretical framework and in

practical application, it is not possible to say, a priori, which model should be used in a particular set of circumstances. This is particularly true in situations where the distributions of the data to be analysed are not known, or where the technique is being used for the first time. In the analysis of data from relevance judgements, both of these situations arise, so it is not possible to state which model should be used. Hence several models were used to analyse the data.

The important points regarding the five models described above are briefly summarized in Table 3.

TABLE 3

Comparison of signal detection models

	SDT eq. var.	SDT uneq. var.	Luce	exponential	Robertson
Assumed distributions	normal	normal	logistic	exponential	logistic
Shape of O.C.	symmetric	asymmetric	symmetric	asymmetric	symmetric
Likelihood ratio monotonic	YES	NO	YES	YES	-
Sensitivity measure	d'	d'_e or Δm	a	k	δ_r
Response bias measure	β	β	V	none	a_r

5.2 Experimental application of SDT to relevance judgements

5.21 Signal and noise

In a signal detection task an observer is requested to indicate whether he thinks a signal has been given, and the observers responses are compared with the presence or absence of a signal using a two by two table (see page 96). This method can be applied to a person's relevance judgements, by comparing the judgements on titles and on the corresponding abstracts or full documents. The relevance judgement on the abstract or document is taken as the definitive judgement (ie. the presence or absence of a stimulus or relevant document). The relevance judgement on the title is equivalent to the users response in the SDT experiment.

Comparisons of judgements on titles and abstracts can be made using a two by two table.

		Title (response)	
		Relevant	Irrelevant
Abstract (stimulus)	Relevant	HIT P(S/s)	MISS P(N/s)
	Irrelevant	False Alarm P(S/n)	Correct Rejection P(N/n)

Thus a reference that is relevant on title and abstract is a hit and a reference that is judged relevant on title but irrelevant on abstract is a false alarm.

The conditional probabilities of hits and false alarms for relevance judgement data are calculated using the following equations:

$$\text{Hit } P(S/s) = \frac{\text{no. hits}}{\text{total no. relevant abstracts}}$$

$$\text{False alarm } P(S/n) = \frac{\text{no. false alarms}}{\text{total no. irrelevant abstracts}}$$

One potential difficulty that arises in the application of SDT to relevance data is the different amount of information available in titles, abstracts and full texts. In effect, the person making relevance judgements on titles has to make a decision based on incomplete information, rather than simply detect a signal in a noisy background. This problem of the incomplete information in titles affected all of the people involved, and any effects due to this would probably have been averaged out, providing a large enough sample and large number of trials were used.

Four different models were used in the analysis of the relevance judgement data. These were the unequal variance SDT, Luce, exponential SDT and Robertson's models. The equal variance SDT model was not used as the assumptions regarding the signal and noise distributions could not be met by the data.

5.22 Experimental methodology

To obtain sensitivity and bias measures using the unequal variance SDT model it was necessary to construct O.C. curves for each user in the present sample.

Points on the O.C. for each individual were obtained by plotting P(S/s) against P(S/n). Several pairs of probabilities were obtained for each person by using

i Two criteria of relevance (major and minor)

ii Several trials, which corresponded approximately to the individual SDI outputs received.

The latter method is similar to the use of trials where the signal probability is varied, but in this case the variation is uncontrolled.

The probabilities were plotted on linear axes and probability axes. The probability plots were done on 'home made' graph paper as such paper is not available commercially. The graph paper used was made up from two sheets of linear-probability paper [Chartwell paper reference 5571].

Sensitivity and bias values for each individual were calculated using each of the four models, and using the methods and formulae given in section 5.1 For all of the

models except the unequal variance SDT model, these measures were calculated directly from the hit and false alarm probabilities for each individual. The overall hit and false alarm figures were used. The equations used to calculate sensitivity and bias were:

i Exponential model

Sensitivity (k)

Log P(S/s) = k Log P(S/n)

No. bias was calculated.

ii Luce's model

Sensitivity (a)

$$a^2 = P(S/s) \cdot \left[1 - P(S/n)\right] \Big/ P(S/n) \cdot \left[1 - P(S/s)\right]$$

response bias (v)

$$v^2 = \left[1 - P(S/n)\right] \cdot \left[1 - P(S/s)\right] \Big/ P(S/s) \cdot P(S/n)$$

iii Robertson's model

The sensitivity and bias measures together with the smoothed values of P(S/s) and P(S/n) were calculated using a computer program developed by Robertson. This program involved an iterative estimating procedure. The bivariate normal distribution model was used. [The programs were very kindly run by Dr. Robertson].

iv Unequal variance SDT model

Sensitivity (d'_e) and response bias (β) were calculated from the O.C. curve for each individual, as described in section 5.12.

The sensitivity and bias values obtained for each user from the four models used were compared in three different ways.

i The relative rankings for each user on each model were compared.

ii Results from the three signal detection models were plotted

iii Pearson correlation coefficients were calculated for various pairs of sensitivity measures.

The pairs of sensitivity measures compared in ii and iii, were in the form that they appear in expressions of the likelihood ratio. The pairs that were compared were

d'_e and $\log a$

d'_e and $-\log k$

a and $-k$

$\log a$ and $-\log k$

The distributions of signal and noise were examined by graphs of rankit against signal (%) and rankit against log signal (%). A graph of rankit against noise (%) was not plotted as noise (%) = 100 - signal (%). [Colquhoun 1971].

5.3 Results

The graph of rankit against log signal (%) had two linear sections connected by a 'step'. This step was caused by four almost identical values of signal (%) occurring together. This large group of tied scores probably caused an interruption of what might otherwise have been a good straight line. Alternatively it could indicate that the signal (%) distribution was bimodal, with both parts having a lognormal distribution. It is assumed that in this case the signal (%) distribution is a unimodal lognormal distribution. For the distributions of signal (%) and conditional probabilities see Appendix A5.3.

A large number of hit and false alarm probabilities were either zero or one, and hence could not be used to calculate sensitivity and bias values. This did not occur with calculations using Robertson's model, as the probabilities were 'smoothed' prior to calculation. The presence of one or zero probabilities also makes it impossible to plot O.C. curves. Because of this problem,

sensitivity and bias values could not be calculated for all of the users on the SDT and Luce's models.

Operating characteristic curves could be plotted for only eighteen users, as there were not enough valid hit and false alarm probabilities to plot curves for the remaining thirteen users. In all cases the O.C. curves were not symmetric about the anti-diagonal. When plotted on double probability scales, the O.C. curves were, on the whole, good straight lines with slopes not equal to one. There were two O.C. curves with only three points and two curves with only four points. This arose because of the number of invalid points (probabilities of zero or one), for these four users. [O.C. curves for the eighteen users are given in Appendix A5.1]

The values of sensitivity and response bias obtained for each user on each of the four models are given in Table 4. Values were calculated for eighteen users using the unequal variance SDT model, for twenty four users using Luce's and the exponential models and for all of the users using Robertson's model. Response bias values were not obtained using the exponential model.

The consistency of the sensitivity values obtained from the four models was examined by comparing the ranking of sensitivity for each user on each model. These rankings

TABLE 4

Sensitivity and Bias values

	SENSITIVITY				BIAS		
User	d'_e	a	-k	δ_r	β	v	a_r
AB	-	6.35	0.128	1.507	-	3.08	-0.712
AC	-	-	-	1.332	-	-	0.114
AD	-	5.86	0.053	1.420	-	0.368	-0.304
BC	-	3.08	0.232	1.257	-	1.54	0.472
CD	0.826	2.14	0.356	1.380	0.451	1.28	-0.114
CE	-	-	-	1.448	-	-	-0.435
CF	0.770	1.48	0.618	1.393	2.438	1.90	-0.174
AE	1.908	5.42	0.076	1.326	1.719	0.677	0.141
CG	1.080	2.04	0.350	1.300	1.386	0.891	0.267
BD	0.936	2.04	0.336	1.372	1.229	0.735	-0.075
DE	-	10.2	0.044	1.466	-	1.28	-0.520
CH	2.116	4.73	0.122	1.408	2.159	1.27	-0.245
AF	1.380	2.20	0.274	1.185	0.677	0.479	0.810
EF	2.032	3.96	0.201	1.494	4.295	2.38	-0.653
EG	2.994	9.50	0.051	1.506	3.291	1.36	-0.711
EH	1.544	5.68	0.071	1.215	0.730	0.674	0.668
EJ	1.444	2.30	0.346	1.453	0.557	1.66	-0.458
AG	1.498	2.88	0.238	1.394	0.748	1.30	-0.180
DF	-	-	-	1.552	-	-	-0.925
DG	-	-	-	1.551	-	-	-0.925
DH	-	-	-	1.382	-	-	-0.121
DJ	-	-	-	1.366	-	-	-0.046
FG	-	3.38	0.213	1.490	-	1.69	-0.634

	SENSITIVITY				BIAS		
User	d'_e	a	-k	δ_r	β	v	a_r
FH	1.974	6.65	0.099	1.426	1.096	2.00	-0.333
FJ	-	4.05	0.282	1.587	-	6.08	-1.092
FK	2.152	7.73	0.068	1.407	3.64	1.43	-0.240
BE	0.992	2.64	0.252	1.350	1.399	1.06	0.030
BF	1.718	4.36	0.124	1.348	1.117	1.02	0.040
BG	-	4.61	0.198	1.566	-	3.46	-0.992
BH	1.166	3.12	0.146	1.186	0.613	0.370	0.080
DK	1.763	-	-	0.962	0.014	-	1.869

are given in Table 5. The ranks for each user were very similar on the SDT, Luce and exponential models, but those from Robertson's model were very different.

Direct comparisons of ranks were made for the seventeen users for whom sensitivity values had been calculated on all three signal detection models. Four users (23.5%) had the same rank on all three models, six users (35.3%) had the same rank on two models, of the remaining seven users, six (35.5%) had adjacent ranks on two models [see Table 6].

Comparisons of sensitivity rankings were possible for twenty four users on Luce's and exponential models. Five users (20.8%) had the same rank, six (25%) were within two ranks and eight (33.3%) were within three ranks. The remaining five users (20.8%) had widely separated ranks. There was no correlation between the sensitivity ranking and the rankings of hit and false alarm probabilities.

The graphs of d'e against Log a and d'e against $-\text{Log } k$ showed a fair scatter of points but they had a general linear trend. [see Appendix A5.2]. This suggested a linear correlation between d'e and a, and between d'e and k. The graph of a against $-k$ showed less scattering of points with a curved trend. The graph of log a against $-\log k$ had slightly scattered points and a linear trend.

Pearson correlation coefficients between the various sensitivity measures were high, and they were all significant

TABLE 5

Rankings of sensitivity and bias values

	SENSITIVITY				BIAS		
Rank	d'_e	a	-k	δ_r	β	v	a_r
1	CF	CF	CF	DK	DK	AD	FJ
2	CD	CG	CD	AF	CD	BH	BG
3	BD	BD	CG	BH	EJ	AF	DH
4	BE	CD	EJ	EH	BH	EH	DF
5	CG	AF	BD	AD	AF	AE	AB
6	BH	EJ	FJ	CG	EH	BD	EG
7	AF	BE	AF	AE	AG	CG	EF
8	EJ	AG	BE	AC	FG	BF	FG
9	AG	BC	AG	BF	BF	BE	DE
10	EH	BH	BC	BE	BD	CH	EJ
11	BF	FH	FH	DJ	CG	DE	CE
12	DK	EF	EF	BD	BE	CD	FH
13	AE	FJ	BG	CD	AE	AG	BC
14	FG	BF	BH	DG	CH	EG	CH
15	EF	BG	AB	CF	CF	FK	FK
16	CH	CH	BF	AG	EG	BC	AG
17	FK	AE	CH	FK	FK	EJ	CF
18	EG	EH	FG	CH	EF	FH	DG
19	-	AD	AE	BC	_	CF	CD
20	-	AB	EH	FH	-	FG	BD
21	-	FG	FK	CE	-	EF	DJ
22	-	FK	AD	EJ	-	AB	BE
23	-	EG	EG	DE	-	BG	BF

	SENSITIVITY				BIAS		
Rank	d'_e	a	-k	δ_r	β	v	a_r
24	–	DE	DE	FG	–	FJ	AC
25	–	–	–	EF	–	–	AE
26	–	–	–	EG	–	–	CG
27	–	–	–	AB	–	–	AD
28	–	–	–	DH	–	–	EH
29	–	–	–	DF	–	–	BH
30	–	–	–	BG	–	–	AF
31	–	–	–	FJ	–	–	DK

TABLE 6

Rankings on four models of sensitivity values for 17 users

Rank	d'e(SDT)	α (Luce)	κ (exponential)	δ (Robertson)
1	CF	CF	CF	AF
2	CD	GG	CD	BH
3	BD	BD	CG	EH
4	BE	CD	EJ	CG
5	CG	AF	BD	AE
6	BH	EJ	AF	BF
7	AF	BE	BE	BE
8	EJ	AG	AG	BD
9	AG	BH	EF	CD
10	EH	EF	BH	CF
11	BF	BF	BF	AG
12	AE	CH	CH	FK
13	FG	AE	FG	CH
14	EF	EH	AE	EJ
15	CH	FG	EH	FG
16	FK	FK	FK	EF
17	EG	EG	EG	EG

at the 0.001 level [see Table 7].

The lower correlation between **a** and **-k** and the very high correlation between Log **a** and -Log **k** reflects the slightly curved nature of this relationship. Luce [1963] found that the relationship between Log **a** and d' was approximately linear, and the most difference was observed when the probabilities approached zero or one.

5.4 Discussion

Luce's model requires logistic distributions of signal and noise and the SDT unequal variance model requires a normal distribution. The distribution of signal (%) (and therefore noise (%)) suggests that the requirements of these two models are not grossly violated. This is presupposing that lognormal distributions of signal and noise can be treated as normal distributions with unequal variances. The exponential model is derived from a gamma function which requires less stringent assumptions regarding the distributions. [Laming 1973]. Robertson's model derives sensitivity and bias values from estimates of hit and false alarm probabilities after an iterative procedure. In this model the signal and noise distributions are not important.

The O.C. curves obtained using the SDT unequal variance model were of a quality comparable to that obtained from various detection tasks in experimental psychology, except

TABLE 7

Correlation between sensitivity measures

Pair	N	Correlation coefficient	t	Significance (2 tailed)
d'e, Log a	17	0.960	13.27	0.001
d'e, -Log k	17	0.817	5.48	0.001
a, -k	24	0.676	4.30	0.001
Loga, -Log k	24	0.931	11.98	0.001

that in a few cases very few points were plotted. The lines of the double probability O.C. plots were drawn freehand. More accurate lines might have been obtained by using linear regression methods. However the small number of points on some of the curves makes these techniques inapplicable. The O.C. curves on linear scales were not symmetric about the anti-diagonal, so the most appropriate model to use is one which does not require symmetry.

There was a high degree of consistency between the rankings of sensitivity for each user on the three signal detection models. Robertson's model gave different rankings. This was not surprising since this latter model was based on very different theories and assumptions.

The correlations between the actual sensitivity values calculated from the signal detection models was very high (significant at the 0.001 level). This indicated a high level of agreement in the measurement of users sensitivity to relevance.

6. VARIABLES

6.1 Choice of variables

One of the aims of this project was to identify some of the variables that affect the relevance judgements made by users of an SDI service. Initially it was necessary to hypothesize the variables which might affect relevance judgements; or that might help in providing a better understanding of how such judgements are made.

Previous studies, that have involved relevance judgements made by actual users, have usually included variables relating to demographic factors, information gathering habits and to the type of work and subject of the users. [Kean 1973, Orr 1970, Wersig 1970]. Experiments at CSL and SDC which used judges rather than the actual users of the SDI service concentrated on variables which related to the judgemental conditions. They examined the effects of relevance definition, scales of relevance, academic training and subject knowledge of the judges. [Cuadra and Katter 1967, Rees and Schultz 1967a].

In the present study actual users gave the relevance judgements, and it was possible to hold constant some of the variables examined by SDC and the CSL. The variables held constant were:

Document representation (titles and abstracts)

Definitions of relevance

Scale for relevance judgements

Judgemental conditions

The variables that were tested in the present study are listed in Table 8. They fall into six different groups.

i Relevance judgement

ii Demographic variables

iii Information gathering habits

iv Attitudes towards information

v Type and nature of work

vi Experience and abilities

The variables within these groups are listed in Table 9.

Demographic variables have been included in the majority of studies involving both judge and user relevance judgements. In the present study several such variables were included (age, education, position) however sex was not included as a variable since the sample contained only one woman. The length of time that a person had been working in their particular subject area was included as a variable. Although this is closely related to age, there are instances where people either change their field of interest mid-way through their career, or they begin work in an area rather late in life. This variable was particularly useful in assessing

TABLE 8

Variables examined

SPSS/Code no.	Continuous variables
VAR 002	Decision category 1
VAR 003	Decision category 2
VAR 004	Decision category 3
VAR 005	Decision category 4
VAR 006	Decision category 5
VAR 007	Number journals scanned
VAR 008	Age
VAR 009	Length of time in field
VAR 010	Hr/wk information gathering
	Discrete variables
VAR 011	Educational level
VAR 012	Position and place of work
*VAR 013	Experience of Industrial R and D
*VAR 014	Experience of teaching
VAR 015	Subject area
VAR 016	Stage of research
VAR 017	Size of work group
*VAR 018	Use of external libraries
*VAR 019	Time limit on research
*VAR 020	Use of delegated searches
VAR 021	Language ability
VAR 022	Access to language ability

SPSS/Code no.	Discrete variables
VAR 023	Frequency of use of Interlibrary Loan
VAR 024	Frequency of photocopying
VAR 025	Speed of reading
VAR 026	Experience of information gathering
VAR 027	ABNO or OBNA
VAR 028	Frequency of use of abstracts journal in past three months
VAR 029	Use of written or verbal sources
*VAR 030	External contacts
*VAR 031	User as an information source
*VAR 032	If enough time spent on literature work
*VAR 033	Pure or applied research
*VAR 034	Experimental or theoretical research
VAR 035	Chemical Abstracts indexes used
VAR 036	Attitude towards cross-referencing
VAR 037	Non-fiction reading
*VAR 039	Journal biases
*VAR 040	Author biases
*VAR 041	Use of Current Contents

* = Dichotomous variable

TABLE 9

Grouping of variables

a RELEVANCE JUDGEMENT

Decision categories 1 to 5

b DEMOGRAPHIC VARIABLES

Age	Position
Length of time in field	Education level

c INFORMATION GATHERING HABITS

Number journals scanned	C.A. Indexes used
Hours/week information gathering	External contacts
Use of external libraries	Use of delegated searches
Frequency interlibrary loan	Frequency photocopying
Frequency use abstracts journal	Use of Current Contents

d ATTITUDES TOWARDS INFORMATION

ABNO-OBNA	Written or verbal sources
Journal and Author biases	Attitude towards cross-referencing
If enough time spent reading	Experience of information gathering

e TYPE AND NATURE OF WORK

Subject area	Stage of research
Pure or applied	Size of work group
Experimental or theoretical	Access to languages
Time limit on research	

f EXPERIENCE AND ATTITUDES

Experience of teaching	Speed of reading
Experience of Industrial R & D	Non-fiction reading
Language ability	Experience of inf. gathering
User as information source	

the subject expertise of the research students in this sample who were all of a similar age.

The variables relating to relevance judgements made by the users were the decision categories. The relevance judgements on titles were each assigned to one of five decision categories [see page 70]. This type of variable has not been examined before.

The variables regarding information gathering habits were designed to discover how often various sources were used, how much time was spent on literature activities and whether certain sources were used. These variables gave an indication of how literature conscious the individual users were. One would expect that the extent of literature usage would influence relevance judgements. The use of primary and secondary sources was examined by the variables on the number of journals scanned, the frequency of use of abstracts journals and the use of Current Contents. The persistence of users in attempting to obtain documents was indicated by the use of external libraries and the frequency of use of ILL.

The users attitudes towards information in general and to particular sources can have an effect on their relevance judgements. One of the more important variables within this group was the amount of information that the users wanted to receive. Users were asked if they looked at all items of possible interest, only items of obvious interest or whether

they compromised between these two approaches. Using the notation of Fairthorne [1964], these two approaches were called ABNO (all but not only) and OBNA (only but not all) respectively. Leggate [1971] also thought that this variable was important. Biases for and against specific authors and journals can affect relevance judgements and must be taken into account. The other variables in this group of attitudes towards information were, opinion of cross-referencing in the abstracts journals used; the use of written, verbal or written and verbal sources, and whether the user felt he spent enough time on literature activities. The users were also asked to give their subjective opinion of their experience in information gathering activities.

The user's type of work and working environment may have an effect on his relevance judgements. Most previous studies have included the users subject area as a variable. Some have included the nature of research, whether it is pure or applied and experimental or theoretical [Clague 1971]. The stage at which a researcher is at in his project affects his information need, and hence his relevance judgements [Orr 1970]. The size of the group within which the user works affects the flow of informal communications [Menzel et al 1960], and may possibly affect the relevance judgements made. This is particularly noticeable in senior members of a group who note articles of interest to other members of their group. The users were asked if there were a time limit on their research, to

see whether they were under pressure to complete their work by a certain date.

The general experience of users and their individual abilities may influence their relevance judgements. The experience variables included in this group were; experience of teaching, industrial R and D and experience of information gathering. The variables relating to individual abilities were; language ability and speed of reading. This latter variable was included as it was thought that people who read quickly would tend either to read more or to spend less time on information gathering, than people who read slowly. Also included in this group were the non-fiction reading (unconnected with work) of users, and whether the user was used as a source of information by other people.

Most of the variables included in the present research have been used in previous studies. However, there are a few new variables which, it is suggested, may affect relevance judgements. These are;

Decision categories 1 to 5
Experience of information gathering
Access to language ability
Frequency of use of ILL
Speed of reading

Length of time in the field

Non-fiction reading

Use of Current Contents

Several of the variables have been divided into discrete categories to standardize the users replies and to facilitate coding. [see Appendix A1.1]. In some cases too many categories were used originally and the number of responses in each category were very small. Under these circumstances the number of categories was reduced by amalgamating adjacent categories. For example the variable on experience of information gathering originally had five categories, the two extreme ones (None and very extensive) contained very few responses and they were merged with their adjacent categories to give a final variable with only three categories, (slight, moderate and extensive). Other variables where the number of categories was reduced were; frequency of photocopying, frequency of use of abstracts journal and Chemical Abstracts indexes used. The main reason for making sure that the categories contained several responses was that the Chi squared statistical test requires that the expected cell frequency is at least five. [Siegel 1956].

6.2 Statistical tests

A variety of statistical tests were used in order to establish;

i Whether the variables were independent of each other

ii Whether the variables influence relevance judgements

Wherever possible non-parametric statistics were used, as they make no assumptions about the distributions of the variables and they can be used with variables having a relatively low level of measurement. Most parametric tests require an interval or ratio level of measurement (ie. continuous variables), whereas non-parametric tests can be used with an ordinal or even, in the case of Chi squared, a nominal level of measurement. [Nie et al 1975].

All of the continuous variables in this study have a ratio level of measurement. Most of the discrete variables that are not dichotomous have categories that can be ordered or ranked, hence an ordinal level of measurement. Dichotomies can be regarded as having an interval or even in some cases a ratio level of measurement [Nie et al 1975]. Only three variables have the lowest level of measurement (nominal) these are, subject area, ABNO-OBNA and non-fiction reading.

The tests carried out to determine whether the variables are independent were the Chi squared test and Spearman ranked correlation coefficients. Both of these tests were

carried out on only the discrete variables. Factor analyses and principal component analyses were performed on the variables to try to find if there were any underlying factors, and to cluster the variables [see Section 7]. The use of factor analysis to clarify the nature of variables involved in relevance decisions has been suggested by Davidson [1974].

To establish whether the variables had any affect on relevance judgements, several tests were employed using the variables and the measures of relevance (conditional probabilities, sensitivity and bias). Pearson correlation coefficients were calculated for the continuous variables and Spearman correlation coefficients calculated for all of the variables. The Mann-Whitney test was carried out on the dichotomous variables and the Kruskal-Wallis test carried out on the remaining variables, the continuous variables were tested by first putting the data into categories. The Kruskal-Wallis test is an extension of the Mann-Whitney test to more than two samples or categories. A more detailed description of these statistical tests is given in Section 3.

The Pearson and Spearman correlation coefficients and the Chi squared test were calculated using SPSS; the Mann-Whitney and Kruskal-Wallis tests were calculated by hand. All of the variables listed in Table 7 were examined in these tests except the use of Current Contents which was omitted from the Chi squared and Spearman correlation between variables tests. The variables author and journal

bias and the use of Current Contents were omitted from the factor analysis.

The choice of significance levels for statistical tests is arbitrary and depends on the variables involved, the sample size and type of analysis being carried out. A convention has been established to use the 0.01 and 0.05 levels and occasionally the 0.1 level of significance. [Siegel 1956]. In the present analysis the 0.01 and 0.05 levels have been adopted, although the levels from 0.02 to 0.05 have been given. The levels were extended to 0.06 and 0.07 but not as far as 0.1. It was rather difficult to use the 0.07, rather than the 0.1 level, for the Mann-Whitney and Kruskal-Wallis tests as the tables of critical values list only the 0.05 and 0.1 levels. This problem was overcome by regarding as significant those values that were in the lower half of the 0.05 to 0.1 level range. It was considered that the 0.1 level was not sufficiently stringent considering the very small sample size. Noting the significance levels at 0.01 intervals was done in order to obtain more detailed information regarding the relationships involved, although in an exploratory study such as the present one, no great importance should be attached to this fine distinction of levels. The important point is whether the variables are significant at the 0.01, 0.05 or 0.07 levels.

6.3 Results and discussion

The results of the various statistical tests used on

the variables are given in Appendix A2. Only those results which were statistically significant have been included.

6.31 Relationships between the variables

The Chi squared test of the independence of variables indicated that almost all of the variables were not independent, and that many of them formed a complex pattern of inter-relationships. Of the 29 variables tested only one (enough time information gathering) was independent. A total of 435 pairs of variables were examined and 41 of these were statistically significant at the 0.07 level or better.

Spearman correlation coefficients were calculated for all those pairs of variables significant on the Chi squares test, together with several additional pairs. The number of significant pairs was 60. Comparisons between pairs of variables significant on Chi squared and Spearman tests are given in Table 10. Only pairs involving the discrete variables were examined as Chi squared values were calculated on these variables only.

The degree of dependence of each variable was calculated by assigning numerical values to the pairs of variables, according to the level of significance. The values for each variable were added up to give an overall

TABLE 10

Comparison of tests for related variables

Variable pair	Significance Chi^2	Significance Spearman
Position-Time limit	0.01	0.02
Access to language - Pure or applied	0.01	0.02
Position - Education	0.02	0.01
Freq. photocopying - External contacts	0.02	0.01
User as inf. source - External contacts	0.02	0.01
Expt. or theor. - Time limit	0.03	0.01
Stage research - Work group	0.01	0.05
Ind. R & D - User as inf. source	0.05	0.02
Exp. inf. gath. - External contacts	0.05	0.02
Education - Written/verbal	0.04	0.05
Exp. inf. gath. - Freq. abst. j.	0.07	0.03
Pure or applied - External contacts	0.07	0.03
Outside libs. - External contacts	0.07	0.03
Freq. ILL - User as inf. source	0.07	0.04
Position - Teaching	0.01	NS
Subject - Pure or applied	0.01	NS
Freq. photocopying - Freq. ILL	NS	0.01
Position - Size work group	0.02	NS
Subject - Time limit	0.02	NS
ABNO - OBNA - Journal bias	0.02	NS
Deleg. search - Freq. ILL	NS	0.02
Freq. abst. j. - Freq. photocopying	NS	0.03
Expt. or theor. - ABNO - OBNA	NS	0.03

Variable pair	Significance Chi^2	Significance Spearman
Education - Exp. teaching	NS	0.04
Language ability - Cross refs.	NS	0.04
Outside libs. - Access to language	NS	0.04
Deleg. search - Freq. photocopying	NS	0.04
Education - Stage research	0.05	NS
Access to language - Speed reading	0.05	NS
Speed reading - Non-fiction	0.04	NS
Freq. ILL - Freq. abst. j.	0.05	NS
Freq. abst. j. - Cross refs.	0.05	NS
Written or verbal - Journal bias	0.05	NS
Expt. or theor. - C.A. indexes	0.05	NS
Freq. ILL - Author bias	NS	0.05
Pure or applied - User as inf. source	NS	0.05
Expt. or theor. - Subject	NS	0.05
Outside libs. - User as inf. source	NS	0.05

NS - Not significant at 0.07 level.

figure of dependence. The numerical values used were

0.01 level = 2

0.02 to 0.05 level = 1

0.05 to 0.07 level = ½

The degree of dependence of each variable is given in Table 11. A diagrammatic representation of these relationships is given in Figure 8.

From these results it is obvious that there are several highly complex relationships between the variables. Only one of the 29 discrete variables, and none of the continuous variables were completely independent.

It was because of the highly related nature of the variables that factor analysis was used. This technique allows the clustering of variables and indicates the underlying factors that may emerge from a large number of variables.

From the results of the Chi squared and Spearman correlation tests, groups of highly inter-related variables can be extracted. These groups consist of one or two 'key' variables, to which all of the remaining group members are related; and variables significantly related to each other and to the key variable(s). There were eleven groups identified amongst the variables. These groups are not mutually exclusive, and variables can belong to more than one group. The eleven groups are given in Table 12

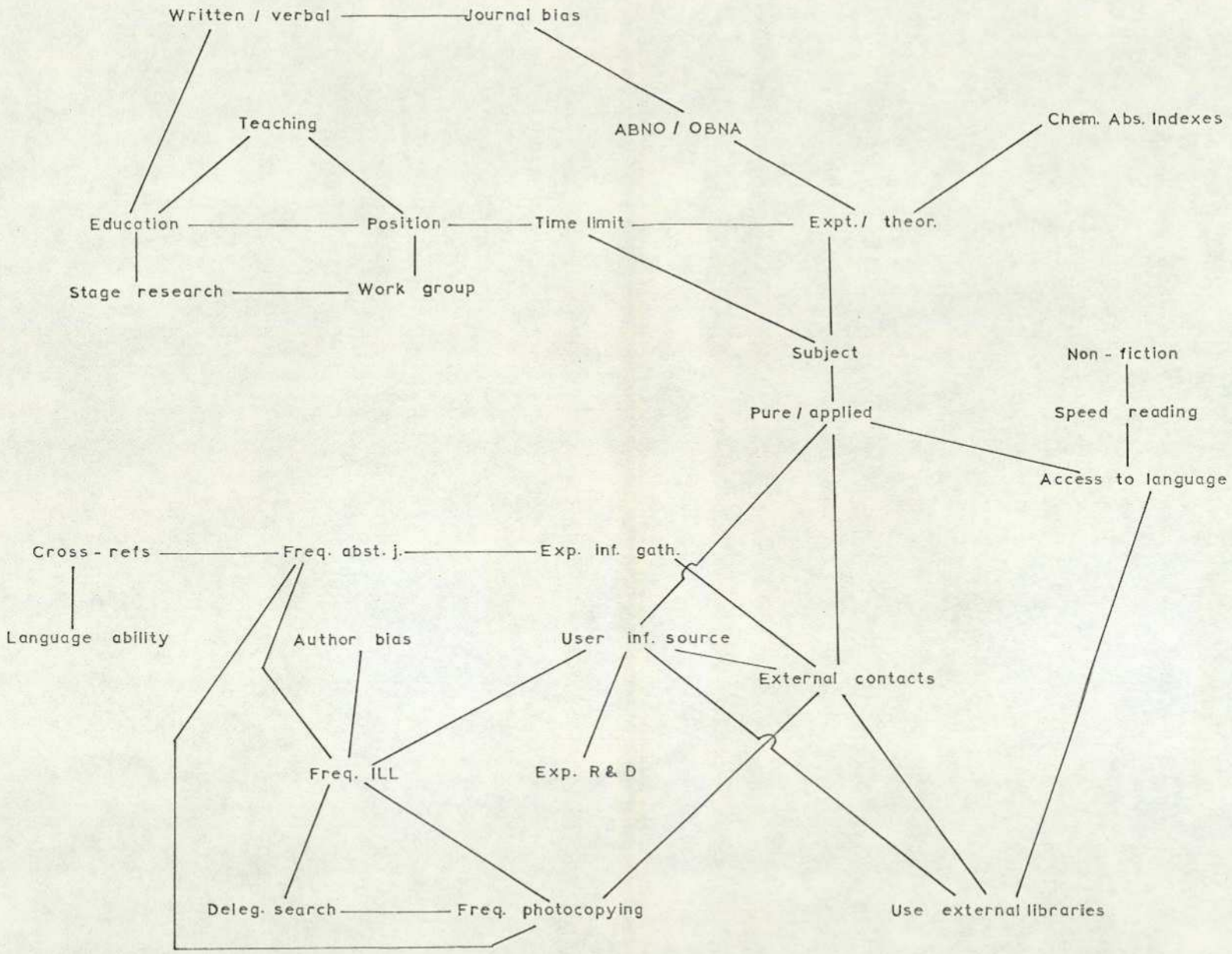


Figure 8 Relationships between the variables

TABLE 11

Degree of dependence of highly related variables

Variable	Dependence
External contacts	11
Position	9
User as information source	8½
Pure or applied	7½
Time limit	7
Frequency of photocopying	7
Frequency of ILL	6½
Experimental or theoretical	6
Access to language	5
Frequency abstracts journal	4½
Education	4
Subject area	4
Work group	4
Exp. information gathering	3½
Use outside libraries	3½

TABLE 12

Groups of variables from Chi squared and Spearman tests

1. *Education
   *Position
   Exp. teaching
   Work group
   Stage research
   Age
   Hr/wk inf. gathering
   No. journals scanned

2. *Experimental or theoretical
   Time limit
   Subject
   ABNO - OBNA
   Chemical Abstracts indexes
   Age

3. *Pure or applied
   Subject
   Access to language
   User as information source
   External contacts

4. *External contacts
   Pure or applied
   Use outside libraries
   Experience inf. gathering
   Freq. photocopying
   User as information source
   Hr/wk inf. gathering

5. *Freq. abstract journal
   Cross-references
   Experience inf. gathering
   Freq. ILL
   Freq. photocopying

6. *Freq. ILL
   Author bias
   Freq. abstracts journal
   Delegated search
   User as inf. source
   Freq. photocopying

7. *User as inf. source
   Pure or applied
   External contacts
   Use outside libraries
   Freq. ILL
   Exp. industrial R and D
   Hr/wk inf. gathering

8. *Freq. photocopying
   External contacts
   Freq. ILL
   Delegated search
   Freq. abstracts journal
   No. journals scanned

9. *Hr/wk inf. gathering
   Use outside libraries
   External contacts
   User as inf. source
   Education
   No. journals scanned

10. *No. journals scanned
    Hr/wk inf. gathering
    Freq. photocopying
    Experience inf. gathering
    Education

11. *Time in field
    *Age
    Education
    Exp. industrial R and D
    Position
    Time limit
    Experimental or theoretical

* = Key variables

6.32 Variables and measures of relevance

All of the variables, except the five decision categories, were tested against the measures of relevance to discover whether any of them had a significant effect on these measures. Of the 34 variables examined, 11 were significant on two tests, 9 were significant on one test and 14 were not significant on any test. The variables within each group are given in Table 13.

In order to determine which variables affect relevance, those which were significant on more than one test and with more than one measure of relevance were examined. These stringent requirements are necessary since in any statistical test there is always a slight chance that a significant relationship is discovered purely by chance. If this stringent requirement is applied to variables being tested, only eight can be confidently said to affect relevance. These are:

Number of journals scanned
Pure or applied research
Frequency of inter library loan
Frequency of use of abstracts journal
Access to language ability
Position
Experience of industrial R and D
Use of written or verbal sources

TABLE 13

Significant variables

i	Significant on two tests (0.07 level or better)	
	Pure or applied	S, M-W
	No. journals scanned	P,S
	Frequency ILL	S, K-W
	Frequency abstracts journal	S, K-W
	Stage of research	S, K-W
	Written or verbal sources	S, M-W
	Language ability	S, K-W
	Access to language	S, K-W
	Position	S, K-W
	ABNO or OBNA	S, K-W
	Experience industrial R and D	S, M-W
ii	Significant on one test (0.07 level or better)	
	Work group	K-W
	Chemical Abstracts indexes	K-W
	Time limit	M-W
	Experimental or theoretical	S
	Age	P
	Length of time in field	S
	Education	S
	Frequency of photocopying	S
	User as information source	S

iii Not significant on any test

Hr/wk information gathering

Experience of teaching

Subject area

Use of outside libraries

Delegated searches

Experience of information gathering

Enough time on information gathering

Cross-referencing

Non-fiction reading

Journal biases

Use of Current Contents

Speed of reading

Author bias

External contacts

S = Spearman correlation coefficient

P = Pearson correlation coefficient

M-W = Mann-Whitney test

K-W =Kruskal-Wallis test

At a lower level of confidence one could take those variables that are significant on two tests, or that have at least 2 measures significant on one test, as influencing relevance. There are seven such variables namely:

Language ability
ABNO or OBNA
Stage of research
Time limit
Size or work group
Chemical Abstracts indexes
Experimental or theoretical

The variables that are not significant on any test do not affect relevance judgements. The position of those variables that do not meet either of the criteria of significance outlined above is doubtful. All that can be said about them is that there is not sufficient evidence to suggest that they influence relevance judgements.

Subject area was not found to be a significant variable in affecting relevance judgements. This is contrary to the results of studies by Morikawa [1974] and Davidson [1974]. The contradiction is probably due to the fact that in these two studies judges rather than users made the relevance judgements, and that the subject specialities of the judges were more widespread than in the present study.

Carrington [1973] found that variations in relevance judgements by a group of users in the same research field were due to differences in experience. This is reflected in the present study by the significance of position in affecting relevance measures. Position is the variable that distinguishes between research students and the more experienced users in the sample. However the length of time that a user had been working in his field was not significant.

The factors that were important in influencing relevance judgements, were the nature of the user's work and his use of various information sources. Amongst the variables relating to the users work, six (pure or applied, experimental or theoretical, stage of research, size of work group, time limit and position) had significant effects on the relevance measures. Of the variables relating to the information sources used, five (written or verbal, number of journals scanned, frequency of use ILL, frequency of use of abstracts journal and Chemical Abstracts indexes used) were significant.

It was interesting that the two variables concerned with language ability and accessibility of linguistic assistance were both significant. Elwen [1972] found that relevance judgements were strongly affected by the language of the original document.

7. FACTOR ANALYSIS

The Chi squared test of the independence of the variables under examination indicated that many complex relationships existed between these variables. In order to try and simplify and sort out these variables, it was decided to use a clustering technique. The method proposed for the analysis of the variables is the multivariate statistical technique of factor analysis. The aim of factor analysis according to Harman [1967] is

> "The principal concern of factor analysis is the resolution of a set of variables linearly in terms of (usually) a small number of categories or factors."

The factors obtained convey all the essential information of the original variables. From such an analysis it should be possible to obtain a single underlying factor from a set of related variables.

There are two main reasons for using factor analysis; one is to try and find the underlying relationships between variables (exploratory), the other is to confirm hypotheses made about the variables (confirmatory). The use of factor analysis in examining variables relating to relevance judgements is exploratory.

Factor analysis has had little use in the field of information retrieval; Cuadra and Katter [1967] used it in a study of the attitudes of the judges they used in their study of relevance judgements. Davidson [1974] suggested that factor analysis should be used to clarify the variables involved in relevance judgement experiments.

7.1 Outline of factor analysis

The basic characteristics and methods of factor analysis will be dealt with, it is not intended to present a detailed or highly mathematical exposition of this subject. A more detailed account can be found in Child [1970] or Harman [1967].

There are three major steps in factor analysis

i Preparation of the correlation matrix

ii Extraction of initial factors

iii Rotation to a terminal solution

In order to find out if a group of variables have anything in common it is necessary to know the nature of the correlations between each pair of variables. This is done by constructing a matrix of Pearson correlation coefficients. The variables in this matrix should be experimentally independent, if they are not spurious correlations may arise [Guilford 1954].

Most distributions of scores on the variables are acceptable provided that they are not excessively skewed, truncated or multimodal [Harman 1967]. It is important that the relationships between the variables are linear, as it is assumed in factor analysis that the correlations are obtained from variables which are linearly related.

7.11 Extraction of factors

Two types of factors are distinguished in the factor analysis model, these are common factors and unique factors. Common factors (C_F) account for the intercorrelations amongst the variables, the unique factors result from the individual properties of the variables. The unique factors account for the residual variance after the common factors have been extracted. The unique factors can be divided into variance due to the nature of the variable (Specificity, U_s) and variance due to errors in measurement etc. (unreliability, (U_u) All of the unique factors should be independent of the common factors.

Each variable has a certain amount of variance associated with it, this variance can be expressed in terms of common and unique factors. Thus the variance of each variable can be expressed as

$$\text{Variance} = C_F + U_s + U_u$$

Each variable is represented by one or more common factors. The sum of the common factor variance for each variable is called its communality. When a variable contains more than one common factor it is regarded as being complex, as opposed to a simple variable which has only one common factor. The complexity of a variable is defined as the number of common factors involved in the description of that variable. Variables usually have a complexity greater than one (ie. they are not simple variables).

The techniques for extracting factors generally aim to take out as much common variance as possible in the first factor. There are several methods for extracting factors, one of the most popular is the principal factor method. (For other methods see Harman [1967]. There are two types of principal factor solution

a. Factor analysis, which takes into account the unique variance
b. Component analysis, which merges the unique variance with the common variance.

In practice the difference between these two types of solution lies in the treatment of the correlation matrix. In factor analysis the main diagonal of the correlation matrix is replaced by estimates of the communality of each variable; whereas in component analysis the main diagonal is replaced by unities. A major problem in factor analysis is finding suitable estimates for the communality of each variable.

Several estimating methods have been proposed, these include using the squared multiple correlation (SMC) or using repeated approximations from an informed guess. The problem of obtaining communality estimates is considered in detail by Harman [1967].

During factor analysis the correlation matrix is inverted, this is not done during component analysis. This inversion can present problems when highly correlated variables are involved, therefore to analyse such variables component analysis must be used.

The number of factors to be extracted is determined using a criterion such as Kaiser's criterion. In this case only factors with eigen values greater than one are considered as being common factors. This criterion is regarded as being most reliable for analyses having 20 to 50 variables [Child 1970]. The eigen value (or latent root) is the ratio of the percentage variance to the number of tests. Each factor is obtained using the following principles:

a. The factors are not correlated
b. Those variables which are most highly inter-correlated are combined within a single factor
c. The factors are derived in such a way that maximizes the percentage of the total variance of the variables attributable to each successive factor.

The rationale for the extraction of principal components has been expressed succinctly by Harman [1967].

> "The first principal component is that linear combination of the original variables which contributes a maximum to their total variance; the second principal component, uncorrelated with the first, contributes a maximum to the residual variance; and so on until the total variance of all n principal components is equal to the sum of the variance of the original variables."

The first factor obtained tends to be a general factor, the remaining factors are more specific and some have several variables with significant negative projections. These latter factors are called bipolar factors.

7.12 Rotation to a terminal solution

The factor matrix derived from the correlation matrix is not very stable, and depends heavily on the relative number of variables examined. This matrix can also be difficult to interpret. These problems can be overcome by altering the frame of reference (or reference axes) of the factor solution; this is called rotation.

For the principal factor solution the reference axes are the major and minor axes of the elliptical distributions. The simplest method of rotation is orthogonal rotation, where the

axes are rotated, but they are kept at right angles to each other. In this case the axes are independent of one another. Alternatively the axes can be rotated at various angles to each other, this is called oblique rotation. In both rotation methods each axis is rotated in turn with all of the remaining axes.

The aim of rotation is to transform the initial solution into a simple structure solution, such as the one proposed by Thurstone [Child 1970]. This simple structure should not only reduce the number of related variables to a small number of independent factors, but should also be invariant.

One of the most commonly used methods of orthogonal rotation is Varimax rotation, another frequently used method is Quartimax rotation. Varimax rotation aims to achieve a simple solution by simplifying the columns (or factors) of the factor matrix. This involves increasing the large factor loadings and decreasing the small loadings for each variable in the original factor matrix. Quartimax rotation aims to simplify each row (or variable) of the factor matrix. Quartimax allows a general factor to emerge from the analysis; Varimax does not permit the emergence of a general factor but it is better at approximating to the classical simple structure principle [Harman 1967].

The exact configuration of a factor structure is not unique and can be transformed into others without violating

the basic assumptions involved. This indeterminancy means that there is no 'best' solution, therefore a rotational method must be chosen to give a terminal solution that satisfies the theoretical and practical problem under investigation [Nie et al 1975].

There are several criteria for deciding on the significance of factor loadings when interpreting the results of a factor analysis. A simple 'rule of thumb' which is frequently used is to consider as significant only those factor loadings greater than 0.3, providing that the sample size is not too small. Interpretation and naming of factors obtained from the terminal solution is carried out by looking at the variables with significant loadings on each factor, and seeing whether these variables fall into recognizable groups.

The factor analysis model requires very few assumptions about the data being analysed. The model is violated if the unique variance components are correlated or if the row vectors of the correlation matrix are not linearly independent. [Pawlik 1973]. Generally an analysis should have three times as many variables as factors. The factors obtained should be regarded as empirical rather than theoretical constructs, too much attention should not be paid to the fine details of the numbers obtained from factor analysis. This point is clearly made by Levy and Pugh [1972].

"(One) should be careful about attempting to achieve solutions which are optimal for the data available; achieving lower coefficients but making fewer data specific assumptions will increase the likelihood that the solutions will hold up in other studies that might be carried out."

Particular care must be taken when using small samples as the effects of errors can be high. This is especially important when dealing with human behaviour where the margin of error is high.

7.13 Analysis of variables in the present study

The factor analysis program of SPSS was used to analyse the variables related to relevance judgements [Nie et al 1975]. Both factor analysis and principal component analysis (PCA) were carried out on the variables. Three sets of variables were examined, these sets contained different numbers of variables. The variables in each set were

Set (i)	16 variables	VAR 007, 009, 010, 012, 015, 016, 018, 021, 023, 024, 026, 027, 028, 030, 031, 032
Set (ii)	27 variables	VAR 011 to VAR 037
Set (iii)	36 variables	VAR 002 to VAR 037

Set (ii) contained only discrete variables, sets (i) and (iii) contained a mixture of discrete and continuous variables.

Extraction of the factors underlying the variables examined involved constructing a matrix of Pearson correlation coefficients for each pair of variables. The initial factor solution matrix was obtained by replacing the main diagonals of the correlation matrix with communality estimates or with values of one; and then extracting factors based on the variance of each variable. The initial factor solution matrix was rotated to get the final factor solution. All of these processes were carried out by computer. Details of the methods used are given in table 14, following the procedure recommended by Child [1970].

7.2 Results

Each of the three sets of variables were analysed using PCA and factor analysis, with both Varimax and Quartimax rotation. This produced a total of 12 analyses, the results of which are given in Appendix 3. In all cases there was little difference between the factor loadings for the two rotation methods. As Varimax rotation approximates better to the theoretical simple structure solution, only results from this rotation were considered in detail.

The number of factors obtained and the percentage variance accounted for was the same in both PCA and factor analysis within each set of variables, but varied with the number of variables in the set.

TABLE 14

Methods used for factor analysis and PCA

Method	PCA	Factor analysis
Factoring method	Principal factors	Principal factors
Diagonal entries	Unities	Communality estimates from SMC with iteration
Criterion for number of factors	Eigen value $\geq$ 1	Eigen value $\geq$ 1
Significance of factor loadings	0.33	0.33
Rotation method	Varimax and Quartimax	Varimax and Quartimax

Set (i)	16 variables	7 factors	79.1% of variance
Set (ii)	27 variables	9 factors	81.2% of variance
Set (iii)	36 variables	11 factors	86.8% of variance

In the analyses of sets (ii) and (iii) the correlation matrix could not be inverted, so the results of the factor analysis may have been unreliable. There was no matrix inversion problem with Set (i). As the correlation matrix could not be inverted for two of the three sets of variables the factor analyses were not examined in detail. Only the results of the PCA for all three sets were considered in detail. The factors extracted by this method were used to interpret the relationships between the variables studied.

The loading of variables on each factor for the three sets of variables examined are given in Appendix 3. The variables are listed in decreasing order of their factor loadings, up to a maximum of six variables. The criterion of significance used was a factor loading of 0.33 or more. All of the variables that were initially included in each set appear in the factor loadings.

Each factor in each of the analyses had at least one simple variable. The first factor extracted had several simple variables.

Set (i)	had	8 simple variables,	50% of the number analysed
Set (ii)	had	18 simple variables,	66.7% of the number analysed
Set (iii)	had	21 simple variables,	58.3% of the number analysed

The results from Set (ii) were compared with those from Set (iii) to see if they produced similar factors, as all of the variables of Set (ii) were included in Set (iii). Such comparisons could not be made with Set (i) as it contained only 16 variables, not all of which were included in Set (ii).

Factors 1 and 2 extracted from Set (iii) were unique as they contained the continuous variables not included in Set (ii). Of the remaining nine factors, six were similar in both sets and three were dissimilar. The six similar factors are given in Table 15, which also shows the variables common to those factors from Sets (ii) and (iii).

7.3 Discussion

At attempt was made to use the multivariate statistical techniques of factor analysis and principal component analysis (PCA) to cluster the variables hypothesized to affect relevance judgements. On the whole the analyses performed were successful in clustering the variables, and a smaller number of factors were extracted.

TABLE 15

Loadings of variables on the factors

FACTOR A	FACTOR B
Cross references (-)	Chemical Abstracts indexes
Language ability (-)	Non-fiction reading
Written or verbal	User as information source
Education	

FACTOR C	FACTOR D
Enough time information gathering	Access to languages
Subject area	Pure or applied (-)
Speed of reading	Teaching experience (-)
	Written or verbal
	Non-fiction reading

FACTOR E	FACTOR F
ABNO or OBNA	Frequency of photocopying
Experimental or theoretical	Frequency of inter-library loan
External contacts	External contacts
	Delegated searches
	Frequency use abstracts journal

(-) indicates a negative factor loading

The analysis was hampered by the small sample size compared with the number of variables. This together with the low level of measurement and unknown distributions of many of the variables gave rise to a situation where factor analysis could not meaningfully be used. In the two sets of variables where there were a large number of variables the correlation matrix could not be inverted; this was caused by the fact that there were almost the same number of variables as cases (ie users). A proper factor analysis could only be carried out on Set (i), where there were half as many variables as cases.

The factor analysis and principal component analysis programs of SPSS could deal with virtually any type of variable and with most levels of measurement. The correlation coefficient matrix was computed using Pearson correlations, with suitable modifications where the variables would not normally have been used for this particular correlation [Nie et al 1975]. This meant that all of the variables examined which had at least an ordinal level of measurement could be correctly used in this clustering technique.

Because of the limitations of the data being analysed and the small sample size, only the order and relative sizes of the factor loadings derived from PCA were examined in detail. To make inferences using the actual values of the factor loadings requires making assumptions about the original data that would probably not be valid.

The factors obtained by PCA from sets (ii) and (iii) were in most cases very similar. Two factors extracted from set (iii) contained the continuous variables, which were examined in only this set. Of the remaining nine factors from set (iii) and the nine factors from set (ii), six have several variables in common. These were combined to give factors A to F [see Table 15].

The analyses carried out on the relevance judgement variables produced an unusually large number of simple variables. According to Nie et al [1975] the complexity of each variable in a factor analysis is usually greater than one (ie. they are not simple variables). In each of the three sets examined 50% or more of the variables were simple. This greatly facilitated the interpretation and naming of the factors. Another feature than can be useful in the interpretation of factors is the occurrence of bipolar factors. However in the present analysis only two bipolar factors were extracted, and this did not help in the interpretation of the factors.

Some of the factors extracted by PCA contained groups of variables that were related to a particular activity or attitude. Five of the six combined factors (A to D and F) and the first two factors of set (iii) were interpreted as reflecting an underlying influence. Factor 1 of set (iii) had high loadings of four of the five variables concerned with the categories of relevance judgements. Thus this

factor relates to the type of relevance judgements made by users. The second factor of set (iii) contained three variables which were measures of time so this factor reflected the effects of time.

Factors A and D contained variables which deal with written and verbal aspects of information seeking activities. These could be regarded as 'linquistic (written)' for factor A and 'linguistic (verbal)' for factor D; although there was no clear-cut distinction between these two. For example the non-fiction reading variable appeared in the 'linguistic (verbal)' group..

Factor C may relate to a person's ability to cope in keeping up to date with new information, as it contained variables on speed of reading and whether enough time was spent on information activities. The success in keeping up to date depended on subject, since some subject areas were much faster moving than others.

The two variables ABNO - OBNA and experimental or theoretical always appeared together in all of the analyses carried out. There was no immediately obvious reason for this co-occurrence. It has been suggested by McAlpine [1972] that experimentalists use the literature less than theoreticians, and that they tend to get their information from colleagues rather than from written sources. This view ties in with the

variables of factor E, and even provides a good reason for the presence of the variable relating to contacts outside of the person's immediate research group.

Factor F contained variables which all reflected information gathering habits. Three of the five variables were quantitative, that is they measured how often a particular source or item was used.

There were two factors from set (ii) that were not combined with factors from set (iii), which appear to have an underlying structure. The first factor (Factor 2 in Appendix A3.2) related to the user's working situation and the type of research work he was doing. The variables concerned in this factor were stage of research, size of work group, time limit on research and experimental or theoretical. The second factor (number eight in the original PCA) contained variables connected with the person's rank or seniority together with their use as a source of information. The more senior a person is the more likely he is to be asked for information, provided that he is involved with research rather than administration. This last factor may involve the role of 'gatekeepers' and/or 'invisible colleges' in information gathering. The names of the factors extracted by PCA are as follows;

Factor A and Factor B	Linguistic (written)
Factor C	Ability to keep up to date

Factor D		Linguistic (verbal)
Factor E		Attitudes towards information
Factor F		Information gathering (quantitative)
Set (ii)	Factor 2	Work situation
	Factor 8	Seniority
Set (iii)	Factor 1	Relevance decisions
	Factor 2	Time effects

8. RELEVANCE JUDGEMENTS

8.1 Relevance judgements and decision categories

Relevance judgements were obtained from 31 users of three different sources of SDI material over a period of about six months. Judgements were given on titles, abstracts and on titles again, after a break of about three months. A three point scale of relevance was used (major value, minor value and irrelevant). Details of the procedure for obtaining the SDI material and the collection of relevance judgements are given in Section 3.

The precision of titles and abstracts for each user is given in Appendix A4.1. The precision of titles was affected by the number of references retrieved, larger outputs had a lower precision than small outputs. [see Table 16]. The average precision of titles, where the number of references retrieved was less than 50 was 26.2%; and was 26.3% for 50 - 100 items. The average precision dropped uniformly as the number of items retrieved increased from 100. There was a very marked fall in average precision if more than 200 references were retrieved; from 22.6% for 150 - 200 items to 14.4% for 200 - 250 items. The precision of T_2 was lower than abstracts for 22 of the 31 users, and was lower for all but two of the users who examined the same number of titles in T_1 and T_2.

The majority of users were happy using a three point scale of relevance. At times some users lapsed into giving

TABLE 16

Effect of output size on precision of titles

Output size	No. users	Average precision (%)
Under 50	7	26.2
50 - 100	4	26.3
101 - 150	10	23.5
151 - 200	2	22.6
200 - 250	6	14.4
251 - 300	-	-
Over 300	2	8.49

just relevant or irrelevant decisions. One user commented that he would prefer more categories to grade relevance, and one user actually gave relevance judgements using five levels of relevance.

The positive relevance judgements made by users on the first set of titles were assigned to one of five decision categories. These categories were a subjective assessment on the part of the author, of the type of decision made by the user. These categories were;

1. Obvious from the information request statement (IRS)
2. Hopeful due to vaguely worded title
3. Using prior knowledge of the subject
4. No apparent reason or relevant to a colleague
5. Possibly of interest (a hopeful decision)

The number of judgements in each category for each user are given in Appendix A4.2.

Relevance judgements on titles (T_1), abstracts (A) and the second set of titles (T_2) were examined with respect to the decision categories assigned to T_1. Relevance judgements on titles and abstracts (T_1-A) and both sets of titles (T_1-T_2) were examined for each decision category. (see Appendix A4.3). The results showed that items which were relevant on both titles and abstracts (T_1-A) were classified

as being either obvious from the IRS (Category 1) or obvious using subject knowledge (Category 3). Judgements regarded as being hopeful (Category 5) had more reverses in relevance (higher R - I values). This tendency of the hopeful decisions was shown more clearly in comparisons of the relevance judgements on the first and second set of titles (T_1 - T_2).

The majority of category 1 judgements remained relevant on abstracts and the repeated titles, and most were of major value on all three occasions. There was a wider range of major and minor value decisions on titles and abstracts for category 3 judgements, however the vast majority of these judgements on T_1 remained relevant on abstracts and titles (T_2). Relevance judgements made for no apparent reason (Category 4), showed, as expected the greatest number of changes from relevant to irrelevant (R - I). Items judged relevant that had vague titles (Category 2) generally remained relevant, although many that were initially regarded as being of major value were demoted to minor value on seeing the abstract or title again. Most of the hopeful relevance judgements (Category 5) were of only minor value. Of the few that were of major value, almost all became minor value or irrelevant when seen again as abstracts or the second set of titles. There was a division between agreement and disagreement of decisions on titles and abstracts (T_1-A) and between decisions on the two sets of titles (T_1-T_2). There were more reversals of judgement on the second set of titles

than on the abstracts when compared with the original titles, for category 5 decisions.

The items that most closely matched the IRS occurred in category 1 and, as would be expected, they were more highly relevant than other items. Relevance judgements made hopefully (Category 5) were less likely to remain relevant when examined again with more information available, as in the case of abstracts.

The effects of three variables on the number of judgements in categories 1, 3 and 5 were examined (see Table 17). There was a noticable difference in the average number of judgements in the decision categories between users who changed their stage of research during this study and those who did not change. The only statistically significant difference was in category 5, where those who changed their stage made almost twice as many hopeful decisions as those who had not changed their stage of research. This latter group included the six users who were supervisors or group leaders, and who made fewer hopeful decisions.

Users who worked alone tended to make more hopeful decisions than those who worked in a group (averages of 10 and 8 respectively). This second group included the supervisors, who may have affected this result. There was no significant difference between the average number of judgements in categories 1 and 3.

TABLE 17

Effects of variables on decision categories

	Category 1	Category 3	Category 5
Change in stage of research	3.40	11.7	11.7
No change	4.57	8.76	6.81
Working alone	3.50	9.0	10.0
Working in group	4.36	9.88	8.0
ABNO	6.56	12.2	9.78
OBNA	3.33	9.17	3.25
Compromise	3.10	8.10	8.30

The users attitude towards the amount of material read (ABNO or OBNA) had a significant effect on the number of hopeful (Category 5) judgements. Those who looked at everything of possible interest (ABNO) made a considerably larger number of hopeful decisions than those who only looked at items of particular interest (OBNA). Those users who compromised between these two attitudes made quite a lot of hopeful decisions. The average number of judgements in category 5 for these groups of users were 9.78, 3.25 and 8.30 respectively. The users with the ABNO approach made a larger number of category 1 and category 3 judgements than those users who had an OBNA or compromise approach. Statistically significant differences in the number of judgements in the different categories occurred only for category 5. The differences in approach were not statistically significant for categories 1 and 3.

8.2 Conditional probabilities

Conditional probabilities of hits and false alarms were calculated for relevance judgements on titles and abstracts (T_1-A, T_2-A) and for judgements on both sets of titles (T_1-T_2). Details of these calculations are given in Section 3. These probabilities were used as measures of relevance against which the variables were tested. The probabilities for each user are given in Appendix A4.4. A problem was encountered in the large number of probabilities with values of zero or one,

these values are invalid, and can present difficulties in the calculation of statistical measures. In a total of 186 conditional probabilities there were 8 zeros and 13 values of one.

The distributions of these conditional probabilities were studied by using rankit plots [Colquhoun 1971]. These were graphs of probability against the corresponding rankit. The zero value probabilities were omitted in the plotting of these graphs. Graphs were plotted linear and semilog scales [see Appendix A5.3].

The linear plots gave curves with 'kinks' in them. These kinks were caused mainly by tied values. The semilog plots tended to be linear except for the first one or two points; again there were kinks due to tied values. All of the probabilities, except T_2-A false alarm, had reasonably linear semilog plots and could be described as having lognormal distributions. The exception (T_2-A false alarm) had a smoothly curved semilog plot, which indicated that this measure did not have a lognormal distribution.

The average hit and false alarm probabilities for the three sets of output (T_1, T_2 and A) were;

0.758 and 0.207 (T_1-A)
0.586 and 0.167 (T_2-A)
0.751 and 0.217 (T_1-T_2)

These values were less extreme than those found by Leggate [1972] in the B.A. Previews experiment. He found an average hit rate of 0.899 and a false alarm rate of 0.101.

The hit and false alarm probabilities for T_1-A, T_2-A and T_1-T_2 were compared to see if these three groups (or treatments) were equivalent using the Wilcoxon matched-pairs signed ranks test. [Siegel 1956]. The results are given in Table 18; if these groups are equivalent the null hypothesis (H_o) is true. There was no difference between the probabilities associated with T_1-A and T_1-T_2 decisions, but there was a significant difference between the probabilities of T_2-A and T_1-T_2 decisions. Thus the treatments T_2-A and T_1-T_2 are equivalent but the treatment T_2-A is not equivalent.

It was thought that a change in the users stage of research might affect his relevance judgements on the second set of titles (T_2), compared with the original titles (T_1). However there was no significant difference in the hit, false alarm, miss and correct rejection rates on T_1-T_2, between users who had changed their stage of research and those who had not. There were some differences in hit and correct rejection rates but these were not statistically significant. Also no correlations were found between hit and false alarm probabilities of T_1-T_2 and stage of research using the Spearman correlation and Kruskal-Wallis tests.

TABLE 18

Wilcoxon matched- pairs signed-ranks test

Pair		N	T	Z	H_o
T_1-A, T_2-A	hit	24	-27	-3.51	Reject
	false alarm	27	-125	-1.54	Accept
T_2-A, T_1-T_2	hit	28	80	-2.80	Reject
	false alarm	31	106	-2.78	Reject
T_1-A, T_1-T_2	hit	28	180	-0.52	Accept
	false alarm	30	153	-1.64	Accept

H_o is rejected at 0.05 level of significance if Z -1.96

8.3 Titles and abstracts

The output received from each of the SDI profiles contained keywords in addition to title and bibliographic citation. The use of keywords to retrieve references in the profiles varied from 20% to 75% of the total number of citations. In this latter case, the profile (on adhesion) retrieved a large number of patents. Patents often have very vague titles and hence retrieval is generally achieved by the use of keywords. The profile that used keywords the least for retrieval was on automated analysis. Most of the papers in this field contained the terms automat or continuous in the title, and these terms were included in the search strategy. The use of keywords to retrieve relevant references and to make relevance judgements was also examined. The results are given in Appendix A4.5.

In most cases less than 10% of the relevant references were retrieved from the data base using keywords. The main exception was the adhesion profile, where such a large number of items were retrieved by keywords, that it was inevitable that most of the users examining this profile should find that a large number of relevant references were retrieved using keywords. Apart from this profile, only two users had more than 10% of their relevant items retrieved using keywords. The use of keywords in making relevance

judgements was quite low. Only four users used the keywords in more than 10% of the references to make relevance decisions, two of whom looked at the adhesion profile. This analysis of the use of keywords was not carried out for the UKCIS Macroprofiles as they do not have the keywords with the title and citation, although the Macroprofiles are searched against the Chemical Abstracts Condensates data base, which includes keywords.

The frequency of occurrence of search profile terms in the total output received and in the relevant items only was counted. The results showed that in seven of the nine profiles, the most frequently occurring terms in the relevant references had been specifically mentioned in the users' information request statements (IRS). In the case of the GABA profile, the terms not specifically mentioned were alternative names for GABA and had been included during profile construction to increase recall. The other profile (on molecular orbitals) had one extra term added (LCAO) which is a fairly common term in this particular field.

The relevance judgements on titles (T_1) and abstracts were compared to see how leniently the users judged titles. Most of the sample (14, 45.2%) judged titles more leniently than abstracts, 10 (32.3%) used approximately the same strictness and 7 (22.6%) judged titles more strictly than abstracts. Those who judged titles leniently included eight

of the thirteen students in the sample. None of the students judged abstracts more leniently than titles.

Comparisons of relevance judgements on the two sets of titles (T_1 and T_2) showed that 21 users (67.7%) judged T_1 more leniently than T_2. This group included all of the research students. Only 7 (22.6%) used a similar degree of strictness and 3 (9.68%) judged T_1 more strictly than T_2.

The relevance judgements made by users sharing profiles were examined to identify a 'core' of relevant references, and to discover the extent of agreement between the users. Only those profiles shared by more than two people were examined; these were combustion, adhesion, nitration, automated analysis and electrochemical reactions.

In the nitration, electrochemical reactions and combustion profiles there was the greatest degree of agreement between the experienced members of the groups. The least experienced people, usually students,(particularly students having recently started their research,) showed the greatest variation in relevance decisions.

In the adhesion profile there was general agreement amongst three of the four users. The fourth person was highly experienced in the field and not involved in practical experimental work. Greatest agreement in the automated analysis profile was between the group leader and the user

with the shortest length of time in the field. The person who agreed least with the other members of this group was the oldest. The differences in agreement here appear to be due to level and recency of education, rather than experience in the field.

The language of items retrieved noticably affected the relevance judgements of some users. Five people commented on the language of a paper as they made relevance judgements on titles (or in one case when looking at abstracts). The usual comment was that they would not make much effort to read papers in Japanese, Russian or other slavonic languages. Two people made comments on the information content of foreign papers.

8.4 Discussion

The precision of the SDI profiles for each user depended on output size, with higher precision figures for smaller outputs. This is in agreement with the findings of the INSPEC SDI evaluation. [Clague 1971].

From the results obtained, it would appear that searching a data base which includes keywords as well as titles, increased the recall by about 40%, but only increased the precision by 7%. Lancaster, Rapport and Penry [1972] found that the use of index terms plus titles increased recall by about 6%, compared with using only index terms. Persson

/¯1974_7 found that the relevance predictability of titles compared with the full document was increased by about 10% when the title was augmented with either keywords or the abstract. The use of keywords for recall is important in areas of applied science, where there are a large number of patents, as patents have relatively uninformative titles, and they are poor for information retrieval purposes. /¯Bottle and Seeley 1970_7.

The use of decision categories to provide a rough guide to the types of relevance judgements made by the users in this study was quite successful. Those judgements made on titles that were similar to the original question (Category 1) and those that were obviously relevant to someone with knowledge of the subject (Category 3), were generally of major value and they remained relevant on abstracts and on the repeated set of titles (T_2). The judgements made in the hope of finding something of interest in a reference (Category 5), were generally of only minor value and were most likely to be judged irrelevant on abstracts or the second set of titles.

The effects of the variables stage of research, ABNO or OBNA and size of work group on the number of relevance judgements in decision categories 1, 3 and 5 were slight. The only statistically significant effects were in the category 5 judgements made by users who had changed their stage of research during the investigation and those who had not. Also between the users who adopted ABNO or OBNA approaches towards information.

The distributions of the probabilities of hits and false alarms were lognormal, except for the T_2-A false alarms. The possible lognormal distributions of measures associated with information retrieval have also been found by King and Bryant [1971] and by the Westat researchers. [Westat 1968]. It was found that the conditional probabilities of T_1-A and T_1-T_2 were equivalent, but that the T_2-A probabilities were not.

Comparisons of the relevance judgements made by users working in the same group indicated that agreement on the relevance of references depends mainly on the experience of the users in their particular subject area. The users with the most experience tend to agree on the relevance of items and the greatest number of disagreements occur with the least experienced researchers.

Relevance judgements were, on the whole, made more leniently on the first set of titles than on abstracts or the second set of titles. The research students made the most lenient judgements on the original titles (T_1).

9. DISCUSSION

9.1 Variables

9.11 Groupings of variables

The 39 variables examined in this study were initially placed into six groups, these were;

Relevance judgement (ie. decision category)
Demographic variables
Information gathering habits
Attitudes towards information
Type and nature of work
Experience and abilities

These variables were examined to see if they were independent of each other, or whether there were any interrelationships between them, using the Chi squared test and Spearman correlation coefficients. Results of these tests indicated that there were eleven groups of highly related variables (see page150) These groups were not mutually exclusive and twelve variables occurred in three or more groups. These were frequency of inter-library loan (ILL), external contacts, age, education, number of journals scanned, hours per week information gathering, pure or applied research, user as information source, use of external libraries, experience of information gathering, frequency of photocopying and frequency

of use of abstracts journal. The high degree of redundancy of these groups was indicated by the fact that all the variables in three groups appeared in other groups.

Factor analysis and principal component analysis (PCA) were used to try and clarify these relationships between the variables, and to discover the underlying factors involved. Due to the large number of variables and small sample size it was not possible to carry out a correct factor analysis, so only PCA was performed. A total of ten factors were extracted, six of these appeared in the PCA of the three sets of variables examined. The remaining four factors appeared as distinct factors on the analyses of the individual sets. Not all of the factors appeared on all of the sets as the variables examined in each set were not the same.

The factors extracted were named as follows;

Linguistic - written (two factors)
Linguistic - verbal
Ability to keep up to date
Attitude towards information
Information gathering - quantitative
Work situation
Seniority
Relevance judgements
Time factors

There was only a slight correspondence between the groups of variables (factors) obtained by PCA and the groups found using Chi squared and Spearman correlation coefficients; only two groups were equivalent. These were related to time (age, length of time in field and time limit) and to information gathering habits (frequency of photocopying, frequency of ILL, external contacts, delegated searches and frequency of use of abstracts journal. There were three partially equivalent groups which contained the variables ABNO-OBNA and experimental or theoretical; pure or applied and access to language; education and position. The remaining five factors and six groups had, at most, only one variable in common.

The factors obtained by PCA were similar to the initial grouping of the variables. The factors corresponded to four of the six original groups of variables, the remaining two groups did not correspond to any of the factors extracted by PCA. Of the five decision categories that were labelled relevance judgements, four appeared together on one factor. The demographic variables appeared on two factors. All of the variables on factor F related to information gathering habits, particularly to the frequency of use of various information sources. Most of the variables regarding the type and nature of work appeared on one factor. There was no single factor

which included mainly variables relating to the users' attitudes towards information or to the users' experience and abilities.

The groups obtained by the Chi squared and Spearman tests did not really correspond to the original groupings or to the factors. Only the demographic variables and the information gathering variables appeared as distinct groups using all three methods of grouping the variables.

9.12 Influence on relevance judgements

The effects of the variables on the measures of relevance (probabilities, sensitivity and bias) were examined using a variety of mainly non-parametric statistical tests. Non-parametric tests were used as they do not make assumptions about the distributions of the variables, and they accept a lower level of measurement than parametric tests. However they suffer from a loss of power compared with the corresponding parametric test. In the present study, which was purely exploratory and had a small sample size, this loss of power was not considered important.

For a variable to be regarded as affecting the relevance measures, it had to be statistically significant on more than one test or significant on one test with more than one relevance measure. Adopting these criteria of

significance, only 15 of the 39 variables tested were found to affect the relevance judgements made by the users studied.

The variables that significantly affected the relevance measures were related to the nature of the user's work and his use of various information sources. Six variables relating to work and five variables relating to information sources were significant. Two variables were concerned with foreign language ability and accessibility. The remaining variables were position and experience of industrial R and D. These results were similar to those of Cuadra and Katter [1967], who found that no one variable affected the relevance judgements made by judges, and that influences from many sources could affect these judgements.

The interrelationships between the variables could affect the significance of individual variables in influencing the measures of relevance. A variable could appear significant solely because it was strongly related to a variable that had a significant effect on the relevance measures. The possibility of this 'carry-over' effect, due to the relationships between the variables, was anticipated by Cuadra and Katter in 1967. They also noted that generalizations cannot be made from the results of one experiment, as the influence of all the variables may not be known.

9.2 Relevance

9.21 Relevance judgements

The distributions of five of the six conditional probabilities, used as measures of relevance, were lognormal. Westat [1968] and King and Bryant [1971] noted that most measures used in information retrieval analyses were skewed to the right, and were best described by lognormal distributions.

The logarithmic nature of estimates made by people has been discussed by Brookes [1976]. He suggested that all human decision making can be expressed by a logarithmic metric. As decisions associated with I.R. systems are made by people, they will follow this metric. The logarithmic scale is 'natural', whereas the physical metric used in science and mathematics is man-made. The results obtained in this present study of relevance decisions support this proposal of Brookes, and the observations of Westat [1968] and King and Bryant [1971].

An attempt was made to discover the types of decisions made by users when assessing the relevance of references to their information requirements. The judgements made were assigned to one of five decision categories. This technique proved to be successful, despite the subjective and arbitrary method of analysing these judgements. The

decisions made on titles that matched the query, either directly or by inference (Categories 1 and 3), were generally of major value and remained relevant when the abstract was seen. The other decision categories contained judgements that were mainly of only minor value, and were subject to a larger number of decision changes on seeing the corresponding abstracts.

It would be interesting to analyse in detail the reasons why users made hopeful decisions (Category 5). A detailed analysis of the data collected was not possible, as the users were not asked to explain in depth the reasons why they decided that certain references were relevant. Such an analysis of the reasons for making decisions would almost certainly affect the user's decision making process.

Some possible reasons for users making hopeful decisions are;

i Insufficient information available

ii Assumptions about the content of the document

iii Decision using peripheral information (eg. author or institution)

iv Assumptions regarding details in the document (eg. spectroscopic data, melting points)

The amount of material normally looked at by the users (ABNO or OBNA) significantly affected the number of hopeful decisions made. Those people who looked at all references of possible interest (ABNO) made, on average, twice as many category 5 decisions as those people who only looked at references of specific interest (OBNA).

The number of items judged relevant by each user was affected by the size of the output received. This effect was very marked when more than 200 references were retrieved. Here the average precision fell from 26% for 150 to 200 references to 14% for 200 to 250 references retrieved. The effect of output size on precision was found in the evaluation of the INSPEC SDI service [Clague 1971], but was not found by Rowlands [1970]. However the figures for output size and precision for individual profiles are rarely published in evaluations of I.R. systems.

The use of keywords to retrieve references from the INSPEC and Chemical Abstracts Condensates databases increased the recall by 43% and the precision by 7%. The high use of keywords for recall was affected by the adhesion profile. This profile used keywords to recall 75% of the total output as it retrieved a large number of patents, which have vague titles and have to be recalled using the added keywords. The users did not make much

use of the keywords when making relevance judgements. On average less than 10% of positive judgements were made after consideration of the keywords. However it was difficult at times to establish whether the keywords did in fact influence the relevance decision. It became obvious during the course of this study, that in some cases keywords were used to decide that a reference was irrelevant.

The relevance judgements made by users looking at the same profile showed a high degree of agreement between those with several years experience in their field. The least experienced members of the group tended to disagree with the other members of the group. A similar effect was found by Carrington [1973] and by Cuadra and Katter [1967]. In the present study the effects of experience could be explained by the fact that the least experienced users were mainly research students, who tended to work in a very narrow field and did not have the wider interests of more experienced research workers. It was interesting to find that in one group (automated analysis) the level and/or recentness of education, rather than experience, was the factor most likely to explain the observed disagreements in relevance judgements.

Almost half of the sample of users judged titles more leniently than abstracts, whilst only 23% judged abstracts more leniently than titles. The majority of the students

gave generous estimations of the relevance of titles; none judged abstracts more leniently than titles. Comparisons of the relevance decisions on both sets of titles (T_1 and T_2) showed that 68% of users judged T_1 more leniently than T_2. All of the students gave more lenient judgements on T_1.

There was a marked difference in the precision of the first and second sets of titles, and between the abstracts and second set of titles (T_2). In both cases the precision of T_2 was lower. It was rather surprising to find that the precision of T_2 was lower than abstracts for two thirds of the sample. The reasons for this may be due to slight changes in the information requirements of the users, the affects of novelty in receiving SDI material or that the references presented the second time were remembered. This latter reason is unlikely as few users said that they remembered any titles more than vaguely, and also as the abstracts were judged more relevant than T_2. A factor which was important, especially with the research students, was the novelty of SDI. Most of them had never received any sort of current awareness material before and had not delegated searches, so they were totally unused to assessing the relevance of lists of titles that they had not found themselves. They reacted to this new situation by over-estimating the relevance of the first titles they received. Hence the precision of T_1 was much higher than that of T_2 or abstracts, for the students.

The judgements of references as being of minor value could have meant one of two things. Either that the reference was known to be of only minor or peripheral value, or that the user was not sure of its degree of relevance. These two aspects of the minor value relevance judgements were not distinguished.

The present study has been concerned with relevance judgements made on titles and abstracts. The relevance measures used were conditional probabilities of hits and false alarms, and each user's sensitivity and bias towards relevance. The probabilities indicated the users' ability to predict the relevance of an abstract from the title, with, in most cases, additional keywords. These probabilities were really measures of the relevance predictability of titles for each individual user. The measures obtained using SDT assessed the users' sensitivity to relevance, that is how easily and accurately each user could determine the relevance of references to his information need. The bias or criterion of relevance determined how much information the user was prepared to accept. Thus the probabilities were concerned with relevance predictability and the SDT measures related to relevance judgements. It was unfortunate that sensitivity and bias values could not be obtained for all of the users, using the SDT models.

9.22 Signal detection theory

An attempt has been made to measure the users' sensitivity towards relevance using three signal detection theory (SDT) models and a model developed by Robertson [1975a] to explain the behaviour of I.R. systems. A high degree of consistency was found in the sensitivity measures obtained from the three SDT models, but the sensitivity from Robertson's model differed considerably. The reason for this difference may have been due to the smoothing of the conditional probabilities that was done prior to calculating the sensitivity and bias measures. This smoothing process pulled all of the probabilities on to a single line, which in effect reduced the degrees of freedom of the sensitivity and bias measures from two to one. [Robertson 1976]. This reduction in the degrees of freedom may account for the differences observed between Robertson's model and the SDT models. Another possible reason for the discrepancies may be that, although the models were superficially similar, in that they involved probabilities derived from two by two tables, they were derived using totally different assumptions and approaches.

It would appear that the SDT models give a consistent evaluation of an individual's sensitivity and bias to relevance. It is therefore necessary to decide which of the three SDT models used in this study is most applicable to the analysis of relevance judgements.

The SDT unequal variance model has two major disadvantages. Firstly the likelihood ratio is not monotonic with x which is a serious theoretical flaw. Secondly it requires the drawing of O.C. curves and calculation of the sensitivity and bias measures from these curves, this is both tedious and time consuming. In many relevance judgement experiments, including the present one, there are not sufficient valid hit and false alarm probabilities to plot the O.C. curves.

Luce's model does not suffer from these two disadvantages, but it has been found to be incompatible in tests where there are a fair proportion of hits together with a low proportion of false alarms [Laming 1973]. However Luce's theory requires that the O.C. is symmetric about the anti-diagonal. This will not necessarily occur, and is not the case in the present study.

The exponential model does not require the drawing of O.C. curves, and it makes the fewest assumptions about the signal and noise distributions. It also allows for asymmetric O.C. curves; and is particularly applicable in response-response situations (which includes the relevance judgement situation). One problem associated with this model is the lack of a recognized method for calculating the response bias. This model has had comparatively little use.

Relevance judgement tasks are similar to word recognition tasks in that they both require not only the detection of a signal (stimulus) but also the discrimination between a correct and an incorrect response. In this situation the individual O.C. is based on response-response rather than stimulus-response contingencies [Green and Swets 1966]. The mathematical modelling of the response-response situation is difficult. Clarke, Birdsall and Tanner [1959] produced a model in which the assumptions they made led to the use of exponential distributions.

A relevance decision consists of two distinct processes, these are the detection and the decision processes. The detection process involves deciding whether a signal has been presented, this depends on the individual's innate ability to detect the signal. The decision process entails deciding on the response to give (ie. whether a reference is relevant or not), this process is affected by the individual's bias or criterion of relevance. A person with a lax criterion tends to judge more references relevant, whereas a person with a strict criterion is very selective and judges comparatively few references relevant. This is basically the same as the ABNO or OBNA approaches to the amount of material looked at. The only difference is that the SDT bias is adopted subconsciously and the ABNO-OBNA approach of each person was given consciously in answer to the questionnaire.

9.3 Questionnaire

The time spent per week on literature activities and the number of journals scanned by the research workers in the present study was rather low compared with other surveys of physical scientists. The low figures are accounted for by the high proportion of research students in the sample (38.7%). These students spent, on average, less time information gathering and scanned fewer journals, than the more established research workers.

It was interesting to find that the journal scanned regularly by the largest number of users was an abstracts journal (Chemical Abstracts). This finding ties in with the suggestion by Meadows [1074] that some researchers scan abstracts journals for their current awareness rather than scanning a large number of primary publications. Current Contents, another type of secondary journal (which is highly advertised), was only scanned by four users. They worked mainly in a diverse field where relevant references appeared in a wide variety of journals, many of which were not held by their institution's library.

Nearly three quarters of the sample (74.2%) could understand papers in French, 45.2% could understand papers in German and only 10% had even a very slight knowledge of Russian. These results were similar to those found in

earlier surveys [Bernal 1948, Wood 1967].

About 40% of the world literature in the sciences is not written in English, and on present estimates only about 15% is in French or German [Wood 1967]. This means that British scientists, with their very restricted knowledge of foreign languages, are cut off from about 25% of the scientific literature. This amount would be drastically reduced if there were a greater ability to cope with Russian language material. However this will not happen until there is a change in attitudes in schools and universities, and the teaching of Russian is much more freely available to science students. At present in the British educational system nearly all school children are taught French and many science students learn German either at school or at University. Only language specialists are taught Russian at school and there is little or no facility or encouragement for science students to learn it at university.

Most of the sample (87.1%) used Chemical Abstracts, but very few used any index other than the subject and author indexes. The organic chemists used the formula index and a few people used the index guide. The findings in this study indicate the lack of awareness of the various indexes available and how they can be useful in literature searching. This is a manifestation of the lack of instruction

in the use of the scientific literature given to students at most universities. Most of the indexes were used by researchers with experience in literature searching.

All of the users in the sample except one used written sources of information and 42% used both written and verbal sources. The majority of users who used verbal sources of information had contacts outside of their institution who they discussed their work with. These users are part of an informal communication network, sometimes referred to as an 'invisible college'. The students did not, on the whole, have external contacts and relied mostly on written information sources. Those users involved in informal communications tended to be heavier literature users.

It was found that all of the users made photocopies in connection with their work, but that the students made fewest copies. This is probably due to the tighter control exercised by university departments on the photocopying by students than on the photocopying by members of staff. None of the students had to pay for their photocopying. There was a generally high level of use of the inter-library loan (ILL) service. The frequency of use of ILL was correlated with the amount of material written in the user's particular field.

Libraries outside of the users' institution were used by 71% of the sample. This reflects the ease of access to good scientific libraries in London and to a lesser extent the limited periodical holdings of the institutional libraries (especially the City University library). The libraries used most often were the Science Reference Library (Holborn division) and the Chemical Society library.

The amount of material examined by users depended on their stage of research and on the type of work they did. Those who looked at only references of obvious relevance (OBNA) generally were involved in both teaching and research and so had little time to spend on literature searching. All of the section or group leaders looked at all references of possible interest (ABNO). These users were responsible for the information gathering of their group, so they had wider interests and were less restrictive in the amount of material they examined.

Most of the scientists in this study worked in a small group, only 16% worked alone. Working in a group allows discussion and cross fertilization of ideas. According to Menzel [1960] there is an optimum mixture of personnel working together, for a scientist to derive maximum benefit from informal discussions. This mixture consists of one or two colleagues working in the same research area, together with a larger number working in related areas. The

scientists who work alone obviously do not have the benefits of these informal discussions.

9.4 Applications of the information seeking and acquisition model

In order to understand how and why certain relevance judgements are made by people, it is very important to have some knowledge of the underlying mental processes involved. The model proposed in Section 2 aims to provide an insight into these mental activities. With a greater understanding of the internal factors involved in making relevance judgements, it should be possible to explain and eventually (hopefully) to predict the effects of external variables on the judgements made by users of information services.

The model of information seeking and acquisition consists of a knowledge structure which represents the way in which information is stored in a person's long term memory. This knowledge structure is built up from information on a particular subject obtained from a wide variety of sources. The structure also contains the means whereby new information can be classified into various subgroups, and these subgroups can be related to give a complete structure.

It is on the basis of this knowledge structure that a person perceives an information need. This need is then transferred into a question or a personal search strategy.

It is usually necessary to modify the original information need to fit in with the requirements of the information sources available.

Relevance judgements are made on the material retrieved by a literature search. Again the knowledge structure is important as it governs the relevance decisions made.

A person's information need and the content and order of his LTM determines the nature of the relevance judgement made. People with highly ordered structures can make relevance decisions easily and have a clear idea whether a reference is relevant or not relevant, providing of course, there is sufficient information available to assess the contents of the document. Users with less ordered structures have greater difficulty in determining the relevance of a reference to their information requirements, as they are unsure as to how the information received might fit into their knowledge structure.

On the basis of a relevance judgement, new relevant information is identified and incorporated into the person's knowledge structure, thus producing a slightly modified structure. In most cases the modification is very slight. However when major changes take place, the person's information needs alter noticeably, at this point it

becomes necessary to modify a user's SDI profile if they have one. The most rapid changes in knowledge structure occur with people who are inexperienced in a particular subject area and are building up their knowledge structures, or in subject fields which are very rapidly moving. It is important to appreciate the continual changes in knowledge structure, and hence information need, of users who receive current awareness material, and to modify the SDI profile as the information need changes.

The knowledge structure of each individual person is different, and so is their ability to express their information need. People also differ in their ability to read and assimilate new information, as Bourne said as early as 1962.

> "Another factor (in I.R. systems) is the basic limitation that the user has in his ability to read and absorb differing amounts of information. The human's input channel capacity and processing rates ... might be considered in the delivery of information (to the user.)"

10. CONCLUSIONS

10.1 Conclusions

The distributions of the conditional probabilities of hits and false alarms were, in all except one case, lognormal. This is in agreement with the suggestion by Brookes [1976] that all processes involving estimation rather than absolute measurement follow a logarithmic metric, rather than a linear metric. The distribution of the signal (%) in the SDT method was also lognormal.

The relevance judgements on titles and abstracts were analysed using three different signal detection models and a model developed to explain the behaviour of I.R. systems. It was found that the exponential SDT model was most applicable to relevance decision tasks, despite its lack of use in other fields. The only other area where it has been used to any extent is in the recognition of words. Both word recognition and relevance judgement tasks are response-response situations. In these situations the person has to decide not only whether a signal has been presented, but also has to give a correct response.

The SDT unequal variance model has had the most use despite its severe theoretical limitations. Luce's model could be used in some relevance experiments. However it

makes greater assumptions about the signal and noise distributions than the exponential model, and it also requires a symmetric operating characteristic (O.C) curve. It has been shown in this study that signal detection methods can be applied to relevance judgements, and that a high degree of consistency was obtained in the sensitivity values calculated for each user on the different models.

The information gathering habits of the established research workers in the present sample were basically similar to those found in previous surveys of chemists and physical scientists. A sharp distinction was found between established researchers and research students. This was particularly noticeable in the use of formal and informal sources of information and in the awareness of the various sources of information available. The students had not yet been accepted into the informal communication network that existed in their specific research field. As a result the students had few contacts outside of their immediate group and needed to rely more on written sources of information. However the students spent relatively less time on literature searching, despite their need to rely on formal sources. This suggests a basic inexperience in using scientific literature and a feeling that literature work is not important. This reaction to information work is probably due to the lack of instruction in literature searching and

availability common in most British universities.

A total of 39 variables were examined to see if they affected relevance judgements. Measures of relevance were used in order to carry out various statistical tests. These measures were the conditional probabilities of hits and fasle alarms, together with the sensitivity and response bias values obtained from the signal detection models. Fifteen variables were found to have a statistically significant effect on the relevance measures. These variables were;

Number of journals scanned

Use of written or verbal sources

Use of inter-library loan

Use of abstracts journal

Chemical Abstracts indexes used

Language ability

Access to language ability

ABNO or OBNA

Pure or applied research

Experimental or theoretical research

Stage of research

Size of work group

If time limit on research

Position

Experience of industrial R and D

These variables fall into two groups; the nature of the user's research and his information gathering habits. These two areas would have been expected to have some effect on relevance judgements.

There were fourteen variables that were not significant on any test. This group included, rather surprisingly, the time spent on literature searching, speed of reading, external contacts and the user's experience of information gathering.

The variables studied in this project were found to be highly interrelated. This makes it difficult to state categorically that any single variable affects relevance judgements, as the affects of one highly influential variable may possibly carry over to variables that are related to it. All that can be said is that these fifteen variables probably affect relevance judgements, but the strength of their influence is unknown.

Because of the relationships between the variables, factor analysis was used to try and reduce the complexity of the variables and discover the underlying factors involved. This technique and the more general method of principal component analysis were used; from a total of 37 variables ten factors were extracted. The factors were named

Linguistic (written) - two factors

Linguistic (verbal)

Attitude to information

Information gathering (quantitative)

Ability to keep up to date

Work situation

Seniority

Relevance decisions

Time effects

These two clustering methods were used although the sample size was really too small for the number of variables involved. Despite this difficulty the results gave an indication of the groupings of the variables that are important when looking at relevance judgements.

An attempt was made to examine the types of relevance judgements made using five categories of judgement. This appears to have given quite consistent results, when the judgements on titles and abstracts were compared. In an attempt to elucidate some of the processes involved in making relevance judgements a model of information seeking and acquisition has been proposed. Using this model it is possible to explain the nature of information need and its transformation into a question, and how this information need affects relevance judgements. The ways in which information is sorted out and stored in the human memory

are shown to affect both relevance judgements and information needs.

10.2 Suggestions for future work

In this present project the application of SDT to relevance judgements was studied for the first time. The use of signal detection methods in relevance experiments could be extended to rating scale experiments. In these experiments the user would be asked to give relevance judgements on a rating scale, rather than the simple relevant-irrelevant dichotomy. For example a user could be requested to give decisions in one of several categories, such as;

definitely relevant
possibly relevant
possibly irrelevant
definitely irrelevant

The advantage of rating scale experiments is that they require fewer relevance judgements to plot operating characteristic curves, and to calculate sensitivity and bias values. This type of experiment should eliminate the problems of probabilities of zero and one.

It has been found that the exponential SDT model is most applicable to relevance judgement tasks. However there is at present no method available for calculating response bias

using this model. A small but important contribution to the use of this model would be to develop a method for calculating response bias, particularly one which is designed specifically with relevance decision tasks in mind.

It would be interesting to study a group of research students from the start of their PhD research project to a few years after their PhD was completed, to see how their information gathering habits and attitudes towards information change. One could see how dependent they are initially on their supervisor for literature searching, and at what point they do all, or almost all, of their own literature searching and current awareness. In a study of research students it would be possible to see at what stage and by what processes a research student or junior research worker becomes accepted into an informal communication network in his field of work.

The knowledge of how relevance judgements are made and what affects them would be considerably increased if more were known about the mental processes involved in information storage in memory and information seeking. Information acquisition by people is affected by their cognitive style, learning strategy and their rate of assimilation of information. These factors affect a person's information seeking habits and probably the relevance judgements they made. A very interesting but difficult area of study would be to examine the effects of

these various factors on information gathering and storage in the memory, and also their effects on relevance judgements. In order to do this, the methods and techniques of psychology and especially cognitive psychology would have to be adapted and used.

APPENDICES

CONTENTS

A1.1 Questionnaire given to users

BACKGROUND QUESTIONNAIRE

1. Name

2. Date of birth

3. Formal education at University

First degree	Higher degree(s)
Subject	Subject
Date	Date
University	University

4. Have you experience of a) Industrial R & D
b) Teaching (not demonstrating)

5. In what subject area is your present research

6. Is your research mainly theoretical experimental

7. Is your research pure applied

8. Do you work mainly a) Alone
b) In a small group (less than 5)
c) In a large group (5 or more)

9. Is there a specific time when your research must be finished

10. How long have you been in your present line of research

11. Which journals do you scan regularly

12. Do you use libraries outside of your institution

13. If YES to 12,
Which other libraries do you use

14. Have you ever used a) Computer information services
b) Internally produced information bulletins

15. Have you ever delegated literature searches (eg to librarians)

16. Which foreign languages can you read technical papers

17. For which foreign languages can you get a colleague to indicate the contents of a paper

18. Have you any biases towards specific a) Journals
b) Authors

19. Approximately how often do you use the interlibrary loan service

More than monthly	Less than once in 3 months
About monthly	Never
About once in 3 months	

20. How often do you make photocopies connected with your work

More than weekly	Monthly
Weekly	Less than monthly
Less than weekly	

OPINION QUESTIONNAIRE

1. How fast do you read technical papers
   a) Quickly
   b) Moderately fast
   c) Slowly

2. What is your experience of literature searching

None	Extensive
Slight	Very extensive
Moderate	

3. How much time per week do you think you spend on information gathering

4. When doing literature searches or looking through references do you
   a) Look at all references of possible interest
   b) Only look at references of particular interest
   c) Compromise between these two approaches

5. Which abstracts journal, if any, do you normally use

6. Do you use Current Contents

7. How often have you used an abstracts journal in the last 3 months

Not at all	Frequently
Occasionally	Very frequently
Often	

8. Do you find the cross referencing in the abstracts journal you use
   Too little
   About right
   Too much

9. For users of Chemical Abstracts; Which of the following indexes have you used

Ring	Patent concordance
HAIC	Numerical patent
Source journals	Index guide

10. Which type of source do you use mainly
    Written
    Verbal
    About equal use of written and verbal

11. Do you have contacts outside of your work place who provide you with information

12. Do people come to you for information regarding their work

13. Do you feel that you spend enough time on literature work

14. Do you read non-fiction material other than that connected with your work. If YES, in what general areas do you read

A1.2 Results from questionnaire

1. Number of journals scanned

None	3	(9.7%)
1 to 5	16	(51.6%)
6 to 10	12	(38.7%)
Mean 4.7	Range 0 to 10	

2. Time spent per week literature searching

Less than $2\frac{1}{2}$ hrs	13	(41.9%)
$2\frac{1}{2}$ to 5 hrs	13	(41.9%)
More than 5 hrs	5	(16.1%)
Mean 3.5 hrs	Range 0.5 to 20 hrs	

3. Age

Under 25	8	(25.8%)
26 to 30	8	(25.8%)
31 to 35	5	(16.1%)
Over 35	10	(32.2%)
Mean 33.2 yrs	Range 21 to 63 yrs	

4. Length of time in field

Less than 2 yrs	10	(32.2%)
2 to 5 yrs	9	(29.0%)
6 to 10 yrs	8	(25.8%)
Over 10 yrs	4	(12.9%)
Mean 5.8 yrs	Range 0.5 to 25 yrs	

5. Educational level achieved

BSc	13	(41.9%)
MSc	7	(22.6%)
PhD	11	(35.5%)

6. Position and place of work

Research student	12	(38.7%)
University lecturer	8	(25.8%)
Research fellow	6	(19.4%)
Government institution	5	(16.1%)

7. Experience of industrial R and D

YES	14	(45.2%)
NO	17	(54.8%)

8. Experience of teaching

YES	17	(54.8%)
NO	14	(45.2%)

9. Size of work group

Working alone	5	(16.1%)
Small group	22	(71.0%)
Large group	4	(12.9%)

10. Use of external libraries

YES	22	(71.0%)
NO	9	(29.0%)

11. External libraries used

SRL	16	(51.6%)
Chem. Soc. (CS)	10	(32.3%)
RAPRA	2	(6.5%)
Others	3	(9.7%)
None	9	(29.0%)
SRL and CS	5	(16.1%)

12. Use of delegated searches

YES	6	(19.4%)
NO	25	(80.6%)

13. Language ability

French	24
German	20
Russian	3
Other	3
French and German	11
None	3

14. Access to language ability

French	7	Czech	3
German	8	Polish	2
Russian	9	Spanish	2
Chinese	3	Other	3
Japanese	2	None	10

15. Frequency of use of inter-library loan

More than monthly	10	(32.3%)
Monthly	8	(25.8%)
3 monthly	4	(12.9%)
Less than 3 monthly	4	(12.9%)
Never	5	(16.1%)

16. Frequency of photocopying

More than weekly	11	(35.5%)
Weekly	3	(9.7%)
Fortnightly	9	(29.0%)
Monthly	5	(16.1%)
Less than monthly	3	(9.7%)

17. Speed of reading

Quickly	8	(25.8%)
Moderately	14	(45.2%)
Slowly	9	(29.0%)

18. Experience of information seeking

Slight	9	(29.0%)
Moderate	14	(45.2%)
Extensive	8	(25.8%)

19. ABNO or OBNA

ABNO	9	(29.0%)
OBNA	12	(38.7%)
Compromise	10	(32.3%)

20. Frequency of use of abstracts journal

Not at all	9	(29.0%)
Occasionally	12	(38.7%)
Often	5	(16.1%)
Frequently	5	(16.1%)

21. Written or verbal sources

Written	17	(54.8%)
Verbal	1	(3.2%)
Both	13	(41.9%)

22. External contacts

YES	21	(67.7%)
NO	10	(32.3%)

23. User as information source

YES	25	(80.6%)
NO	6	(19.4%)

24. If enough time spent on literature work

YES	13	(41.9%)
NO	18	(58.1%)

25. Pure or applied research

Pure	10	(32.3%)
Applied	21	(67.7%)

26. Experimental or Theoretical research

Experimental	26	(83.9%)
Theoretical	5	(16.1%)

27. Chemical Abstracts indexes used

Subject and author	14	(45.2%)
One other	6	(19.4%)
More than one other	5	(16.1%)
Not applicable	6	(19.4%)

28. Attitude towards cross-referencing

Too much	2	(6.5%)
Alright	22	(71.0%)
Too little	3	(9.7%)
Not applicable	4	(12.9%)

29. Author and journal biases

Author bias	8	(25.8%)
Journal bias	11	(35.5%)

30. Use of Current Contents

YES	4	(12.9%)
NO	27	(87.1%)

31. Use of computerized I.R. services

YES	5	(16.1%)
NO	26	(83.9%)

32. Subject area

Organic	6	(19.4%)
Analytical	6	(19.4%)
Electrochemistry	5	(16.1%)
Combustion	4	(12.9%)
Adhesion	4	(12.9%)
Inorganic and nuclear	6	(19.4%)

33. Non-fiction reading

Scientific	5	(16.1%)
Social science	7	(22.6%)
Humanities and arts	7	(22.6%)
General	6	(19.4%)
None	6	(19.4%)

34. Abstracts journal used

Chemical	27	(87.1%)
Physics	2	(6.5%)
RAPRA	1	(3.2%)
None	1	(3.2%)

35. Stage of research

Preliminary	1	(3.2%)
Experimental	11	(35.5%)
Data analysis	1	(3.2%)
Writing up	1	(3.2%)
Supervisor	6	(19.4%)
Prelim/expt.	1	(3.2%)
Expt/analysis	3	(9.7%)
Analysis/writing	4	(12.9%)
Expt/writing	3	(9.7%)

36. If a time limit on research

YES	16	(51.6%)
NO	15	(48.4%)

37. Amount written in field

Very little	3	(9.7%)
Little	7	(22.6%)
Average	9	(29.0%)
Fair amount	9	(29.0%)
Great deal	3	(9.7%)

38. Journals scanned regularly (by at least two users)

Chemical Abstracts	6
The Analyst	5
J.Amer.Chem.Soc.	5
Analytical chem.	4
J.Polymer Sci.	4
Combust. & Flame	4
Electrochim. Acta	4
J.Electrochem. Soc.	4
Current Contents	4
J.Chem.Soc. Perkin I & II	3
J.Appl.Polymer Sci.	3
J.Organic Chem.	3
J.Chem.Soc. Faraday	3
Anal. Biochem.	2
Chem. & Ind.	2
Chem. in Britain	2
Tetrahedron	2
Tetrahedron Lett.	2
Aust. J.Chem.	2
European Polym.J.	2
J.Thermal Analysis	2
J.Fire & Flame	2
Corrosion Science	2
J.Appl.Electrochem.	2
J.Phys.Chem.	2
J.Catalysis	2

Nature	2
Brain Research	2
Nuclear Inst. Methods	2

A2.1 Spearman correlation coefficients (variables)

Variable pair	R_s	Signif.
Age, Time in field	.788	0.01
Age, Position	.682	0.01
Age, time limit	.676	0.01
No. journals scanned, Hr/wk inf. gath.	.668	0.01
Freq. photocopying, External contacts	.593	0.01
Freq. photocopying, Frequency ILL	.553	0.01
Expt. or theor, Current contents	-.542	0.01
User inf. source, External contacts	.535	0.01
Exp. industrial R&D, Time in field	-.534	0.01
Exp. industrial R&D, Age	-.526	0.01
Time in field, Education	.517	0.01
Time in field, Time limit	.512	0.01
Hr/wk inf. gath.,Exp. inf. gathering	.510	0.01
Position, Education	.499	0.01
No. journals scanned, Exp. inf. gath.	.492	0.01
Hr/wk inf. gath., External contacts	.485	0.01
Time in field, Position	.483	0.01
Expt. or theor., Time limit	-.474	0.01
No. journals scanned, Freq. photocopying	-.470	0.01
Time in field, Exp. inf. gath.	.445	0.02
Exp. industrial R&D, User inf. source	.445	0.02
Pure or applied, Access language	-.445	0.02
Age, Education	.436	0.02
Exp. inf. gath., External contacts	-.435	0.02
Position, Time limit	.430	0.02
Delegated search, Frequency ILL	.423	0.02
No. journals scanned, Education	.421	0.02
Hr/wk inf. gathering, Education	.415	0.03
Pure or applied, External contacts	-.410	0.03
Hr/wk inf. gath., User inf. source	-.413	0.03
External libraries, External contacts	.410	0.03
Exp. inf. gath., Freq. abstracts j.	.404	0.03
Freq. abstracts j., Freq. photocopying	-.400	0.03
Expt. or theor., ABNO or OBNA	.398	0.03
Language ability, Cross-references	.414	0.04

A2.1 Cont.

Variable pair	R_s	Signif.
Freq ILL, User inf. source	.381	0.04
Delegated search, Freq. photocopying	.380	0.04
Expt. or theor., Age	-.375	0.04
External libraries, Access language	.373	0.04
Exp. teaching, Education	-.373	0.04
Stage research, Size work group	-.385	0.05
Expt. or theor., Subject area	.361	0.05
External libraries, User inf. source	.361	0.05
Pure or applied, User inf. source	.360	0.05
Education, Written or verbal	.360	0.05
Frequency ILL, Author bias	.357	0.05
Hr/wk inf. gath., External libraries	-.357	0.05
Delegated search, Cross-references	-.372	0.06
No. journals scanned, Position	.354	0.06
Exp. teaching, Access to language	-.354	0.06
Hr/wk inf. gath., Freq. photocopying	-.354	0.06
Exp. inf. gath., Delegated search	-.353	0.06
No. journals scanned, Time in field	.349	0.06
Pure or applied, Exp. industrial R&D	-.349	0.06
Exp. teaching, Age	-.349	0.06
Exp. teaching, Language ability	.347	0.06
Position, Freq photocopying	-.347	0.06
Time in field, Stage of research	.359	0.07
Hr/wk inf. gath., Time limit	.338	0.07
Time in field, External contacts	-.333	0.07

A2.2 Chi squared test

Variable pair	X^2	D.F.	Signif
Position, Time limit	17.5	3	0.01
Position, Exp. teaching	15.3	3	0.01
Access language, Pure or applied	12.2	3	0.01
Subject area, Pure or applied	17.7	5	0.01
Stage research, Size work group	27.6	8	0.01
External contacts, User inf source	6.22	1	0.02
ABNO or OBNA, Journal bias	8.48	2	0.02
Freq. photocopying, External contacts	12.1	4	0.02
Subject area, Time limit	13.4	5	0.02
Position, Size work group	15.2	6	0.02
Education, Position	16.5	6	0.02
Time limit, Expt. or theor.	4.78	1	0.03
Education, Written or verbal	9.94	4	0.04
Exp. industrial R&D, User inf. source	4.07	1	0.05
Exp. inf. gath., External contacts	6.12	2	0.05
Written or verbal, Journal bias	6.10	2	0.05
Expt. or theor., Chem. Abs. indexes	9.55	4	0.05
Access language, Speed of reading	12.9	6	0.05
Speed of reading, Non-fiction	16.3	8	0.05
Education, Stage research	15.6	8	0.05
Frequency ILL, Freq. abstracts j.	21.3	12	0.05
Freq. abstracts j., Cross-references	21.1	12	0.05
Size work group, Time limit	5.96	2	0.06
Exp. inf. gath., Journal bias	5.96	2	0.06
External libraries, Exp. inf. gath.	5.83	2	0.06
External contacts, Non-fiction	9.01	4	0.06
Education, Language ability	15.5	8	0.06
Language ability, Exp. inf. gath.	15.2	8	0.06
Position, Subject area	24.8	15	0.06
Subject area, Stage research	31.3	20	0.06
External contacts, Pure or applied	3.50	1	0.07
External libraries, External contacts	3.49	1	0.07
Exp. industrial R&D, Size work group	5.60	2	0.07
Time limit, ABNO or OBNA	5.57	2	0.07
External contacts, Education	5.52	2	0.07

A2.2 Cont.

Variable pair	X^2	D.F.	Signif
Position, Delegated search	7.19	3	0.07
Freq. abstracts j., External contacts	7.10	3	0.07
Frequency ILL, User inf. source	8.90	4	0.07
Exp. inf. gath., Freq. abstracts j.	11.9	6	0.07
ABNO or OBNA, Chem. Abs. indexes	14.9	8	0.07
Chem. Abs. indexes, Non-fiction	25.3	16	0.07

A2.3 Pearson correlation coefficients

Variable	Measure	R	Signif.
Decision category 1	False alarm T_2-A	.443	0.02
Decision category 1	False alarm T_1-T_2	.419	0.02
Decision category 2	False alarm T_2-A	.595	0.01
Decision category 2	False alarm T_1-A	.377	0.04
Decision category 2	Bias SDT	-.502	0.04
Decision category 3	False alarm T_1-A	.506	0.01
Decision category 3	False alarm T_2-A	.542	0.01
Decision category 3	False alarm T_1-T_2	.645	0.01
Decision category 3	Bias SDT	-.556	0.02
Decision category 3	Hit T_1-T_2	.413	0.03
Decision category 5	False alarm T_2-A	.745	0.01
Decision category 5	False alarm T_1-A	.632	0.01
Decision category 5	False alarm T_1-T_2	.573	0.01
Decision category 5	Bias SDT	-.542	0.03
Decision category 5	Hit T_2-A	.409	0.03
Decision category 5	Hit T_1-T_2	-.426	0.04
Decision category 5	Bias Luce	-.426	0.04
Age	Bias Luce	-.503	0.02
No. journals scanned	False alarm T_2-A	.396	0.03
No. journals scanned	False alarm T_1-A	.369	0.05

A2.4 Spearman correlation coefficients

Variable	Measure	R_s	Signif.
Decision category 1	False alarm T_2-A	.517	0.01
Decision category 1	Sensitivity Rob.	-.444	0.02
Decision category 1	Sensitivity Luce	.413	0.05
Decision category 1	False alarm T_1-T_2	.367	0.05
Decision category 2	False alarm T_2-A	.567	0.01
Decision category 2	False alarm T_1-A	.523	0.01
Decision category 2	Bias SDT	-.497	0.02
Decision category 3	False alarm T_1-T_2	.679	0.01
Decision category 3	Sensitivity Rob.	-.649	0.01
Decision category 3	Bias SDT	-.635	0.01
Decision category 3	False alarm T_1-A	.561	0.01
Decision category 3	False alarm T_2-A	.531	0.01
Decision category 3	Bias Luce	-.490	0.02
Decision category 3	Hit T_1-T_2	.373	0.05
Decision category 4	False alarm T_1-A	.408	0.03
Decision category 4	False alarm T_1-T_2	.378	0.04
Decision category 5	False alarm T_1-A	.807	0.01
Decision category 5	Sensitivity Rob.	-.771	0.01
Decision category 5	False alarm T_1-T_2	.661	0.01
Decision category 5	False alarm T_2-A	.636	0.01
Decision category 5	Bias SDT	-.538	0.02
Decision category 5	Bias Luce	-.503	0.03
No. journals scanned	Bias Luce	-.541	0.03
No. journals scanned	False alarm T_2-A	.404	0.03
Time in field	False alarm T_2-A	.416	0.02
Education	False alarm T_2-A	.365	0.05
Position	Bias SDT	-.502	0.04
Position	False alarm T_2-A	.363	0.05
Exp. industrial R&D	Hit T_2-A	-.381	0.04
Exp. industrial R&D	False alarm T_2-A	-.370	0.05
Subject area	Sensitivity SDT	.474	0.05
Stage of research	False alarm T_2-A	.437	0.02
Access to language	Bias SDT	-.513	0.03
Frequency ILL	False alarm T_1-T_2	-.483	0.01
Frequency ILL	Sensitivity Rob.	.393	0.03
Frequency ILL	Hit T_1-T_2	-.366	0.05

A2.4 Cont.

Variable	Measure	R_s	Signif
Freq. photocopying	False alarm T_1-T_2	-.359	0.05
ABNO or OBNA	Hit T_1-A	-.432	0.02
Freq. abstracts j.	Sensitivity SDT	.587	0.02
Written or verbal	Bias Luce	.482	0.03
Written or verbal	Sensitivity SDT	.460	0.05
User inf. source	Bias SDT	.477	0.05
Pure or applied	Sensitivity Luce	-.437	0.04
Pure or applied	False alarm T_1-T_2	.359	0.05
Expt. or theor.	False alarm T_1-A	.475	0.01
Expt. or theor.	False alarm T_1-T_2	.434	0.02
Expt. or theor.	Bias Luce	-.424	0.04
Decision category 2	Sensitivity SDT	-.468	0.06
Decision category 2	Sensitivity Rob.	-.352	0.07
Age	False alarm T_2-A	.352	0.06
Time in field	Hit T_2-A	.339	0.07
Position	Hit T_2-A	.348	0.06
Language ability	Hit T_1-A	.332	0.07
Access to language	Hit T_1-A	.346	0.06
ABNO or OBNA	Hit T_2-A	-.335	0.07
Written or verbal	Hit T_1-T_2	-.340	0.07
Enough time	Bia s Robertson	-.344	0.06
Expt. or theor.	Sensitivity Rob.	-.338	0.07
Current contents	Bias Luce	.380	0.06

A2.5 Mann-Whitney U test

Variable	Measure	U	Signif.
Pure or applied	Sensitivity Luce	18	0.05
Pure or applied	Sensitivity SDT	22	0.1
Pure or applied	False alarm T_1-T_2	59	0.1
Pure or applied	False alarm T_1-A	63	0.1
Written or verbal	Hit T_1-T_2	67	0.05
Written or verbal	Sensitivity SDT	21	0.1
Exp. industrial R&D	Hit T_2-A	57	0.02
Exp. industrial R&D	False alarm T_2-A	66.5	0.05
Time limit	False alarm T_2-A	75	0.1
Time limit	Bias SDT	20	0.1
Expt. or theor.	False alarm T_1-A	48.5	0.1
Author bias	Bias Luce	31	0.1
Author bias	Sensitivity expon.	32	0.1
User inf. source	Bias SDT	2	0.1
Current contents	Bias Luce	16	0.1

A2.6 Kruskal-Wallis one-way analysis of variance

Variable	Measure	H	DF	Signif.
Frequency ILL	False alarm T_1-T_2	13.9	4	0.01
Freq. abstracts j.	Hit T_1-A	14.6	3	0.05
Freq. abstracts j.	False alarm T_1-T_2	13.2	3	0.05
Freq. abstracts j.	False alarm T_2-A	12.6	3	0.05
Freq. abstracts j.	False alarm T_1-A	7.07	3	0.07
Freq. abstracts j.	Sensitivity SDT	6.38	3	0.1
Access language	False alarm T_1-T_2	13.0	3	0.01
Access language	Bias SDT	7.01	3	0.07
Access language	False alarm T_1-A	6.33	3	0.1
Size work group	False alarm T_2-A	8.56	2	0.03
Size work group	False alarm T_1-T_2	5.70	2	0.07
Size work group	Bias Luce	4.62	2	0.1
Chem. Abs. indexes	False alarm T_2-A	9.95	3	0.03
Chem. Abs. indexes	Hit T_2-A	7.29	3	0.07
Language ability	Hit T_1-T_2	10.1	4	0.05
Age	Hit T_2-A	6.69	3	0.1
Time in field	False alarm T_2-A	4.88	2	0.1
Position	False alarm T_2-A	7.07	3	0.07
Position	False alarm T_1-T_2	6.97	3	0.07
Freq. photocopying	False alarm T_1-T_2	7.79	4	0.1
ABNO or OBNA	Hit T_1-A	5.77	2	0.07
Stage of research	False alarm T_2-A	7.37	3	0.07
Speed of reading	Bia s SDT	4.68	2	0.1

A3.1 Principal component analysis set i

Factor 1
No. journals scanned*
Hr/wk information gathering*
Exp. information gathering
Freq. use abstracts journal
External contacts (-)
Frequency of photocopying (-)

Factor 2
Frequency use ILL*
Subject area (-)
Frequency of photocopying

Factor 3
Stage of research*
Subject area
Language ability (-)
Use external libraries (-)
Freq. use abstracts journal

Factor 4
Position* (-)
Use external libraries (-)
Language ability
Enough time inf. gathering

Factor 5
Length of time in field*
Experience information gathering
Enough time inf. gathering

Factor 6
User as information source*
Use external libraries
External contacts

Factor 7
ABNO or OBNA*
External contacts
Language ability (-)
Freq. use abstracts journal

* Indicates a simple variable
(-) indicates a negative factor loading

A3.2 Principal component analysis set ii

Factor 1
Frequency photocopying*
Exp. information gathering* (-)
Frequency of ILL* (-)
External contacts
Freq. use abstracts journal*
Delegated searches

Factor 2
Stage of research*
Size of work group* (-)
Time limit* (-)
Speed of reading
Experimental or theoretical
Written or verbal (-)

Factor 3
Chem. Abs. indexes*
Non-fiction reading (-)
Exp. Industrial R&D*
User as information source
Pure or applied (-)

Factor 4
Access to languages*
Experience of teaching* (-)
Written or verbal
Pure or applied (-)
Non-fiction reading

Factor 5
Cross-references* (-)
Language ability* (-)
Written or verbal
Education

Factor 6
ABNO or OBNA*
Experimental or theoretical
Delegated searches(-)
External contacts

Factor 7
Enough time inf. gathering*
Subject area*
Speed of reading

Factor 8
Position* (-)
Education (-)
User as inf. source

Factor 9
Use external libraries*
Pure or applied

* Indicates a simple variable
(-) Indicates a negative factor loading

A3.3 Principal component analysis set iii

Factor 1
Decision category 1*
Decision category 2
Decision category 3*
Decision category 5*
Stage of research

Factor 2
Length of time in field*
Age*
Time limit*
Size work group
Exp. industrial R&D (-)

Factor 3
Cross-referencing* (-)
Language ability* (-)
Subject area
Written or verbal
Education

Factor 4
Frequency of ILL*
Delegated searches
User as information source

Factor 5
Chem. Abs. indexes*
Non-fiction reading
User as information source
Size work group (-)
Hr/wk inf. gathering

Factor 6
Access to language*
Pure or applied* (-)
Exp. of teaching (-)
Written or verbal
Non-fiction reading

Factor 7
Frequency of photocopying* (-)
External contacts * (-)
Freq. use abstracts journal
No. journals scanned
Exp. information gathering

Factor 8
Enough time inf. gathering*
Subject area*
Position (-)
Use external libraries
No. journals scanned
Speed of reading

Factor 9
Decision category 4*
Decision category 2
Speed of reading
Exp. industrial R&D (-)
Education

Factor 10
Use external libraries*
Stage of research (-)
Size work group
Freq. use abstracts journal (-)
User as information source

A3.3 Cont.

Factor 11

ABNO or OBNA*

Experimental or theoretical*

Speed of reading

External contacts

Freq. use abstracts journal

* Indicates a simple variable

(-) Indicates a negative loading

A3.4 Factor loading of simple variables

Set i

Factor 1	No. journals scanned
	Hr/wk information gathering
Factor 2	Frequency of ILL
Factor 3	Stage of research
Factor 4	Position
Factor 5	Length of time in field
Factor 6	User as information source
Factor 7	ABNO or OBNA
Factor 8	Position
Factor 9	Use of external contacts

Set ii

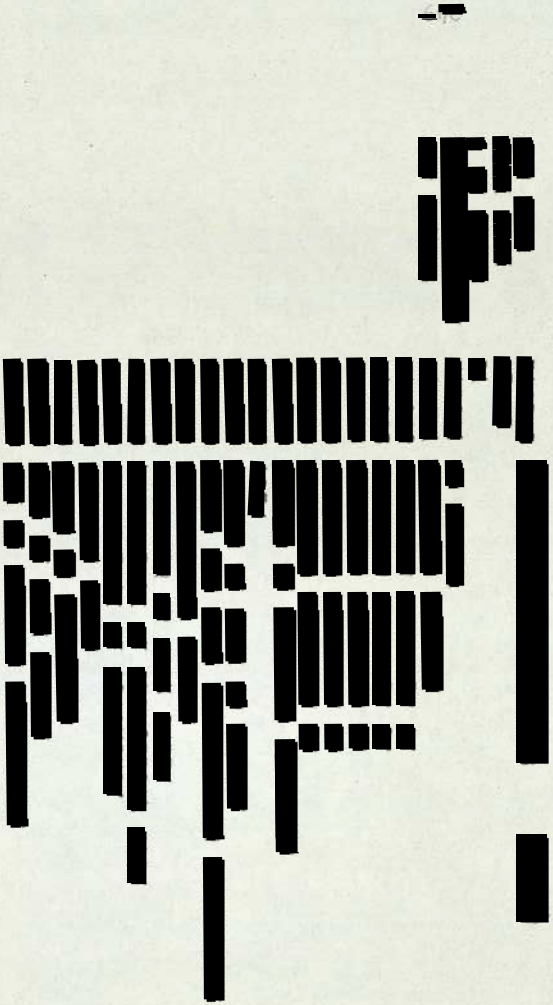
Factor 1	Freq. of photocopying
	Exp. information gathering
	Frequency ILL
	Freq. use abstracts journal
Factor 2	Stage of research
	Size work group
	Time limit
Factor 3	Chem. Abs. indexes
	Exp. industrial R&D
Factor 4	Access to language
	Experience of teaching
Factor 5	Cross-references
	Language ability
Factor 6	ABNO or OBNA
Factor 7	Enough time inf. gathering
	Subject area
Factor 8	Position
Factor 9	Use of external libraries

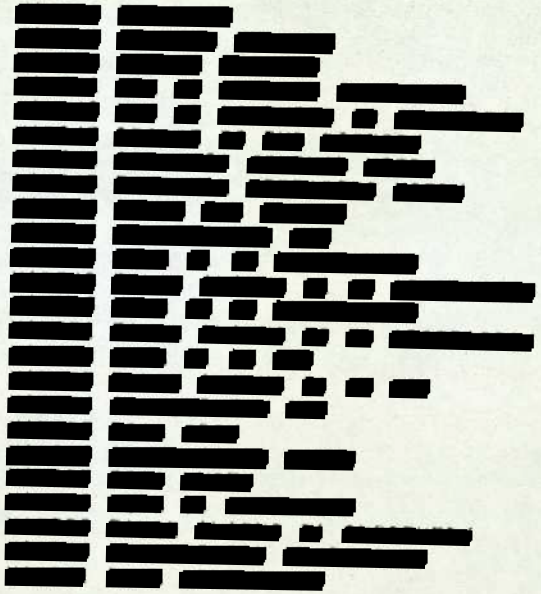
A3.4 Cont.

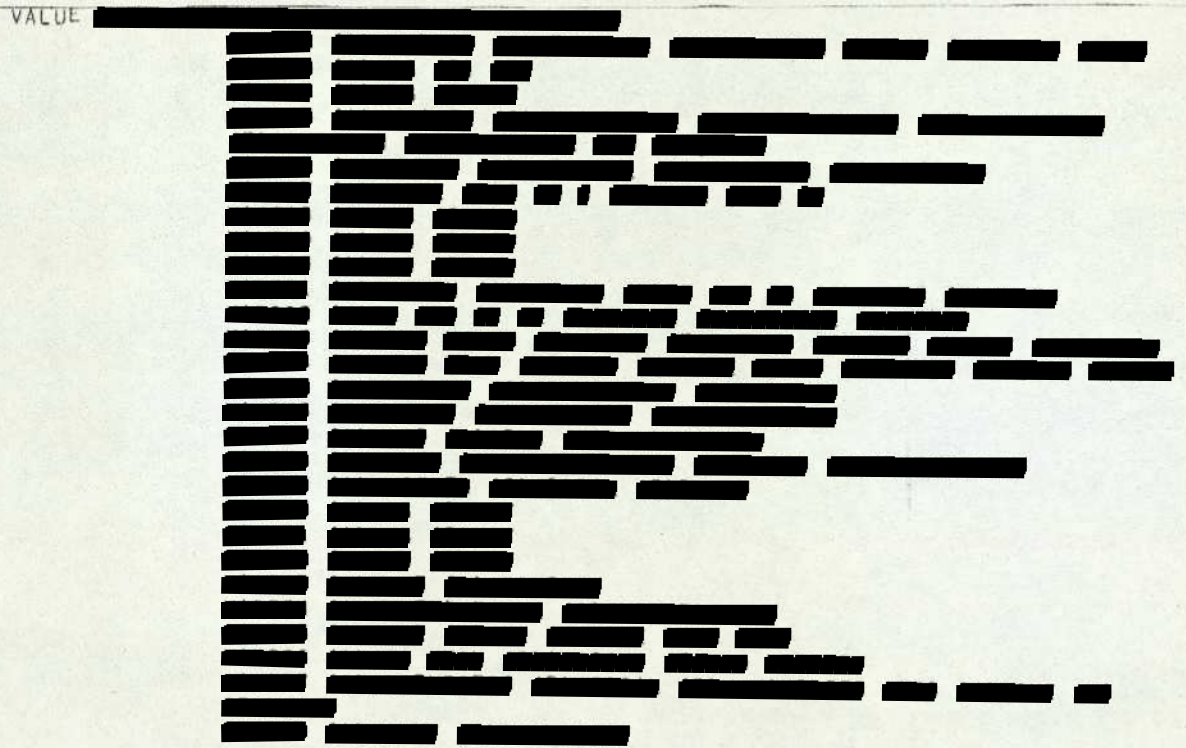
Set iii

Factor 1	Decision category 1
	Decision category 3
	Decision category 5
Factor 2	Length of time in field
	Age
	Time limit
Factor 3	Cross references
	Language ability
Factor 4	Frequency ILL
	Delegated searches
Factor 5	Chem. Abs. indexes
Factor 6	Access to language
	Pure or applied
Factor 7	Freq. of photocopying
	External contacts
Factor 8	Experience of information gathering
	Subject area
Factor 9	Decision category 4
Factor 10	Use external libraries
Factor 11	ABNO or OBNA
	Experimental or theoretical

A3.5 SPSS computer programs for factor analysis and principal component analysis

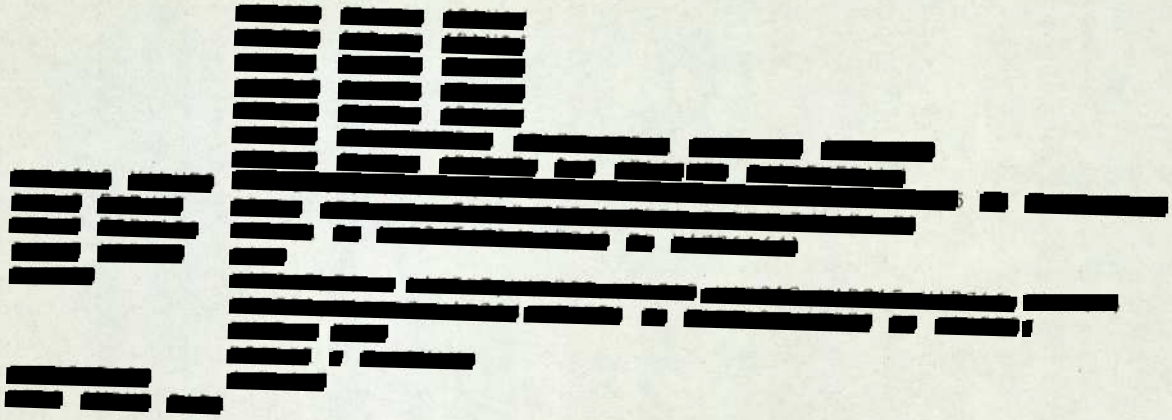





VALUE

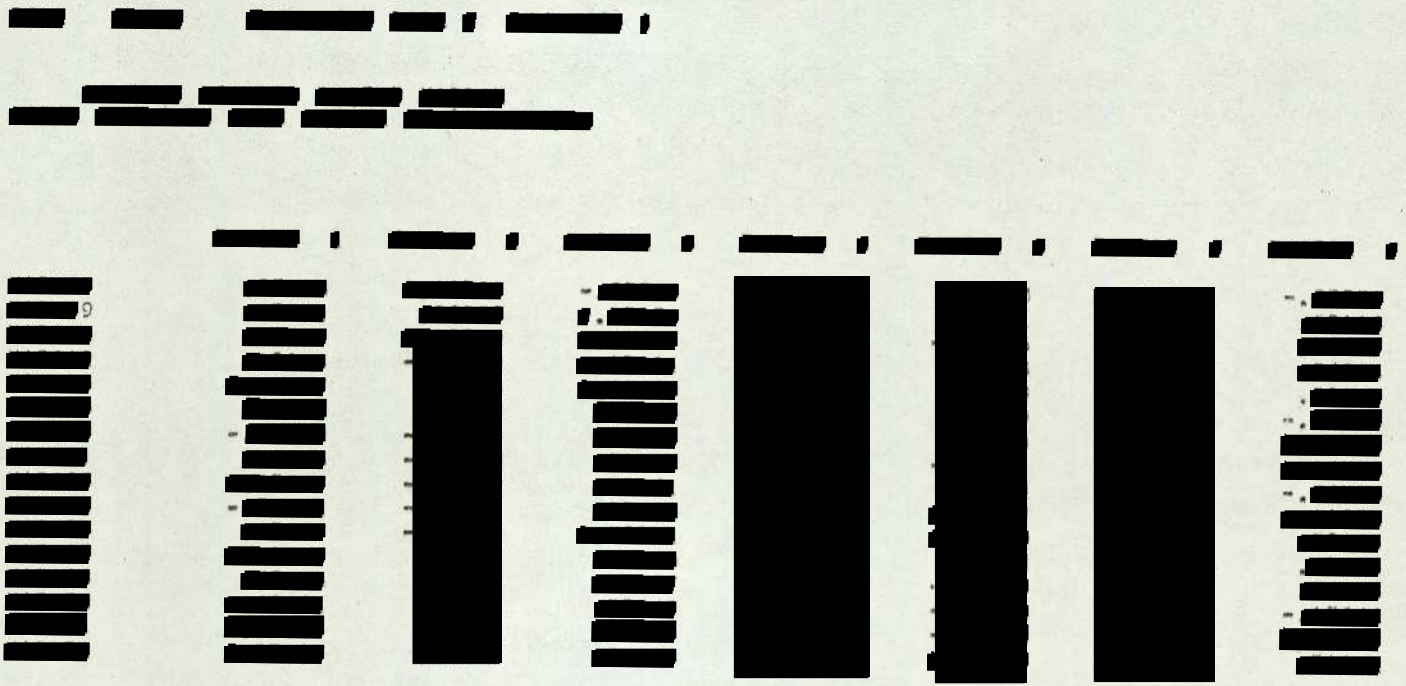
FACTOR ANALYSIS 16 VARIABLES

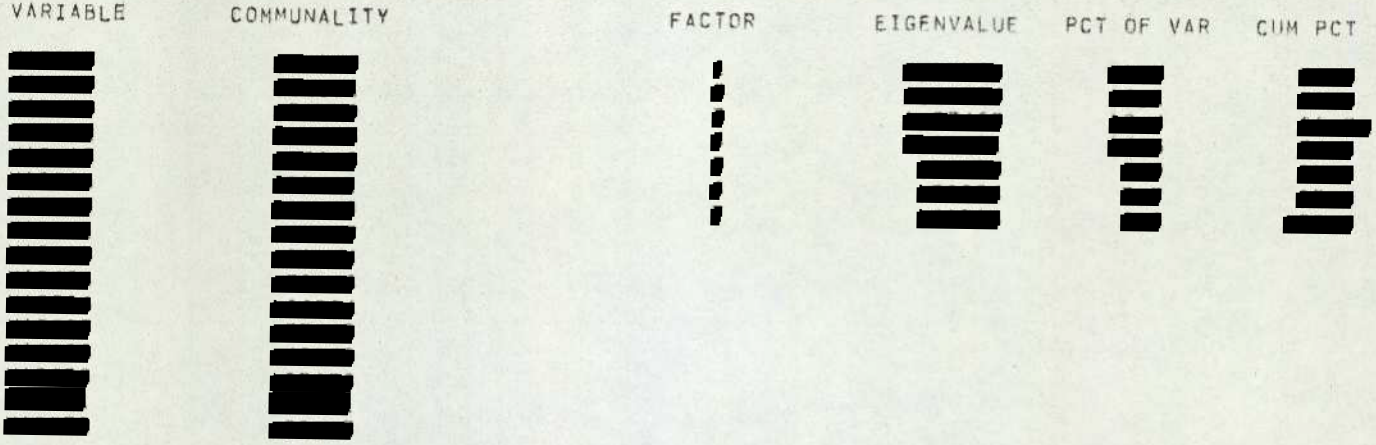
27/05/76



FACTOR ANALYSIS 16 VARIABLES Set(i)

27/05/76




VARIABLE
COMMUNALITY
FACTOR
EIGENVALUE
PCT OF VAR
CUM PCT

FACTOR ANALYSIS 27 VARIABLES [redacted]

FILE RELV3 (CREATION DATE = 27/3[redacted]

VARIMAX ROTATED FACTOR MATRIX
AFTER ROTATION WITH KAISER NORMALIZATION

VARIABLE	COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT

FACTOR ANALYSIS 35 VARIABLES Sct(iii) 26/05/76 PAGE 8

FILE RELV3 (CREATION DATE = 26/05/76)

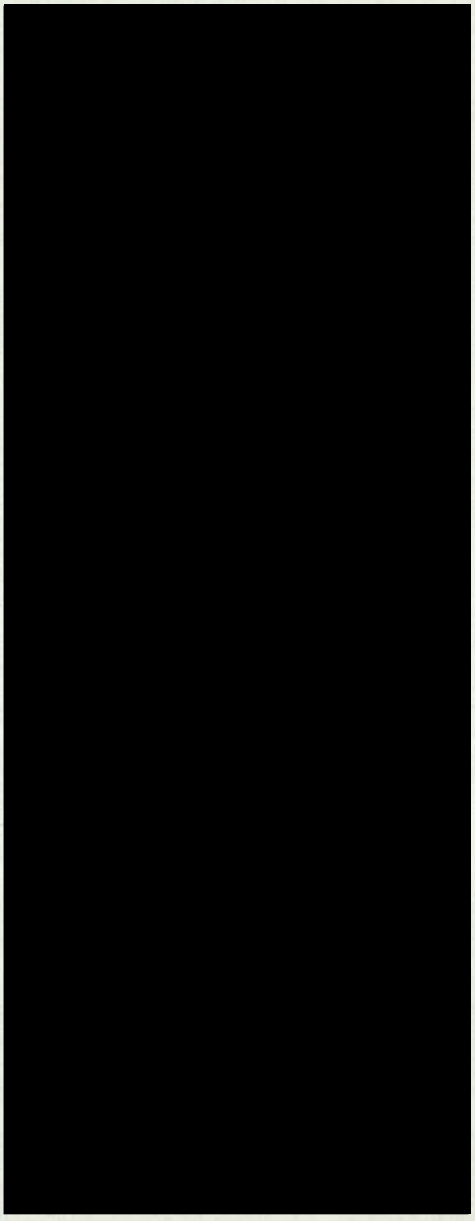
VARIMAX ROTATED FACTOR MATRIX
AFTER ROTATION WITH KAISER NORMALIZATION

FACTOR 1 FACTOR 2 FACTOR 3 FACTOR 4 FACTOR 5 FACTOR 6 FACTOR 7 FACTOR 8 FACTOR 9 FACTOR 10

FACTOR ANALYSIS 35 VARIABLES

FILE RELV3 (CREATION DATE = 26/05/76)

FACTOR 11



FILE RELV3 (CREATION DATE = 26/05/76)

VARIABLE	COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT

PRINCIPAL COMPONENTS 16 VARIABLES Set(6)

FILE RELV3 (CREATION DATE = 27/05/76)

27/05/76

VARIMAX ROTATED FACTOR MATRIX
AFTER ROTATION WITH KAISER NORMALIZATION

FACTOR 1 FACTOR 2 FACTOR 3 FACTOR 4 FACTOR 5 FACTOR

PRINCIPAL COMPONENTS 16 VARIABLES

27/05/76

FILE RELV3 (CREATION DATE = 27/05/76)

VARIABLE	EST COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT

PRINCIPAL COMPONENTS 27 VARIABLES set(ii) 26/05/76 PAGE 7

FILE RELV3 (CREATION DATE = 26/05/76)

VARIMAX ROTATED FACTOR MATRIX
AFTER ROTATION WITH KAISER NORMALIZATION

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8	FACTOR 9
VAR011	[redacted]								
VAR012									
VAR013									
VAR014									
VAR015									
VAR016									
VAR017									
VAR018									
VAR019									
VAR020									
VAR021									
VAR022									
VAR023									
VAR024									
VAR025									
VAR026									
VAR027									
VAR028									
VAR029									
VAR030									
VAR031									
VAR032									
VAR033									
VAR034									
VAR035									
VAR036									
VAR037									

FILE RELV3 (CREATION DATE = 25/05/76)

VARIABLE	EST COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT

PRINCIPAL COMPONET ANALYSIS.. SET(iii)

FILE RELV3 (CREATION DATE = 25/05/76)

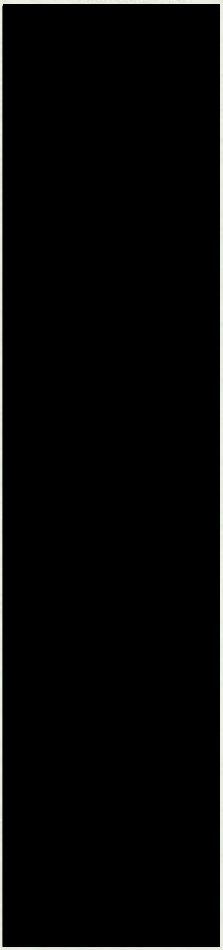
VARIMAX ROTATED FACTOR MATRIX
AFTER ROTATION WITH KAISER NORMALIZATION

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8	FACTOR 9	FACTOR 10

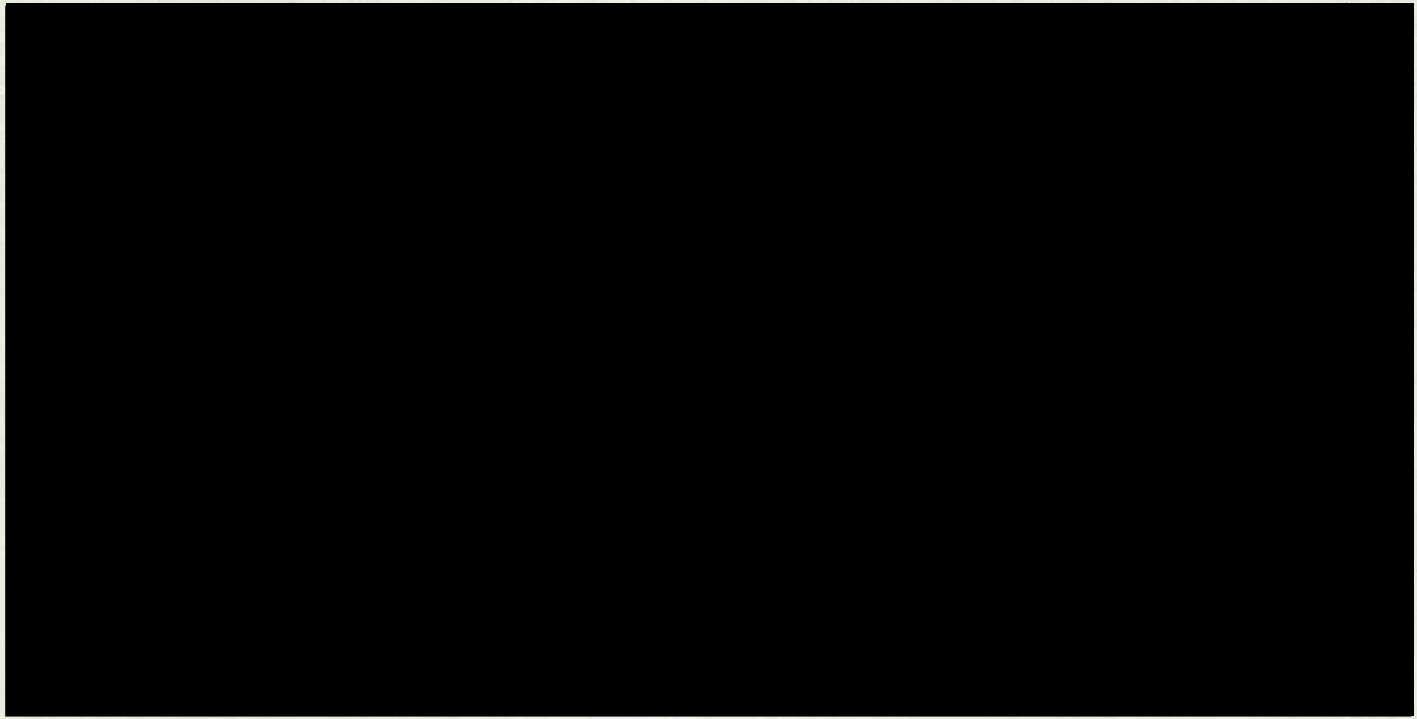
FILE RELV3 (CREATION DATE = 25/05/76)

FACTOR 11

VAR002
VAR003
VAR004
VAR005
VAR006
VAR007
VAR008
VAR009
VAR010
VAR011
VAR012
VAR013
VAR014
VAR015
VAR016
VAR017
VAR018
VAR019
VAR020
VAR021
VAR022
VAR023
VAR024
VAR025
VAR026
VAR027
VAR028
VAR029
VAR030
VAR031
VAR032
VAR033
VAR034
VAR035
VAR036
VAR037



VARIABLE	EST COMMUNALITY	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT



A4.1 Cont.

ii. Abstracts

User	No. abstracts	No. relevant	Precision (%)
AB	30	9	42.9
AC	30	4	13.3
AD	99	17	17.2
BC	38	15	39.5
CD	54	24	44.4
CE	54	5	9.26
CF	54	16	29.6
AE	32	18	56.3
CG	54	23	42.6
BD	59	49	81.4
DE	74	18	24.3
CH	68	18	26.5
AF	67	28	41.8
EF	91	8	8.79
EG	91	8	8.79
EH	123	90	73.2
EJ	91	43	47.3
AG	67	29	43.3
DF	38	1	2.63
DG	38	3	7.89
DH	38	2	5.26
DJ	38	8	21.1
FG	56	13	23.2
FH	56	9	16.1
FJ	56	5	8.93
FK	56	32	57.1
DK	55	26	47.3
BE	75	56	74.7
BF	75	37	49.3
BH	75	47	62.7
BG	75	21	28.0

A4.1 Cont.

iii. Second set of titles (T_2)

User	No. titles	No. relevant	Precision (%)
AB	30	10	33.3
AC	30	11	36.7
AD	99	13	13.1
BC	38	2	5.26
CD	54	11	20.4
CE	54	6	11.1
CF	54	13	24.1
AE	32	14	43.8
CG	54	20	37.0
BD	59	56	94.9
DE	74	16	21.6
CH	68	18	26.5
AF	67	31	46.3
EF	91	5	5.49
EG	91	7	7.69
EH	123	94	77.3
EJ	91	32	35.2
AG	67	16	23.9
DF	38	1	2.63
DG	38	2	5.26
DH	38	0	0.00
DJ	38	19	50.0
FG	56	8	14.3
FH	56	3	5.36
FJ	56	4	7.14
FK	56	29	51.8
DK	55	39	70.9
BE	75	44	58.7
BF	75	33	44.0
BG	75	34	45.3
BH	75	15	20.0

A4.2 Decision categories

User	Category 1	Category 2	Category 3	Category 4	Category 5
AB	2	1	1	0	3
AC	2	1	4	0	3
AD	12	0	17	0	13
BC	1	1	6	0	6
CD	0	3	9	2	9
CE	1	1	2	0	6
CF	0	1	6	3	8
AE	1	0	8	0	10
CG	0	2	7	6	12
BD	0	5	10	0	25
DE	3	3	6	1	7
CH	5	1	10	1	5
AF	3	3	19	2	15
EF	0	0	8	1	4
EG	1	1	5	0	6
EH	25	5	35	7	24
EJ	3	4	15	2	11
AG	2	1	16	0	9
DF	0	0	1	0	1
DG	0	0	3	1	5
DH	0	0	0	0	2
DJ	3	1	4	0	5
FG	1	0	7	0	5
FH	2	0	5	2	4
FJ	1	0	3	0	0
FK	11	0	8	0	10
DK	7	1	21	1	21
BE	12	1	21	0	11
BF	13	0	14	0	10
BG	3	0	8	0	3
BH	16	3	22	1	17

A4.3 Relevance judgements and decision categories

i. Comparison of judgements on titles and abstracts with respect to decision category. (T_1-A)

Decision	Category 1	Category 2	Category 3	Category 4	Category 5
Maj-Maj	60	4	77	2	1
Maj-Min	33	5	57	0	1
Min-Maj	7	5	24	1	22
Min-Min	22	16	69	11	128
Maj-Irr	2	4	20	0	10
Min-Irr	5	5	55	16	100
Rel-Rel	122	30	227	14	152
Rel-Irr	7	9	75	16	110
Total	129	39	302	30	262

ii. Comparison of judgements on both sets of titles with respect to decision category (T_1-T_2)

Decision	Category 1	Category 2	Category 3	Category 4	Category 5
Maj-Maj	66	4	82	1	3
Maj-Min	23	8	48	0	4
Min-Maj	4	6	23	4	27
Min-Min	18	10	63	9	92
Maj-Irr	5	1	25	1	4
Min-Irr	13	11	59	15	132
Rel-Rel	111	28	216	14	126
Rel-Irr	18	12	84	16	136
Total	129	40	300	30	262

A4.4 Conditional probabilities of hits and false alarms

User	Hit T_1-A	False alarm T_1-A	Hit T_2-A	False alarm T_2-A	Hit T_1-T_2	False alarm T_1-T_2
AB	0.677	0.048	0.556	0.238	0.600	0.050
AC	1.000	0.231	1.000	0.269	0.727	0.105
AD	0.941	0.317	0.471	0.061	1.000	0.337
BC	0.667	0.174	0.133	0.000	1.000	0.333
CD	0.625	0.267	0.417	0.033	0.818	0.326
CE	1.000	0.102	0.400	0.081	0.667	0.125
CF	0.438	0.263	0.563	0.105	0.462	0.268
AE	0.889	0.214	0.667	0.143	0.857	0.389
CG	0.696	0.355	0.609	0.194	0.850	0.294
BD	0.735	0.400	1.000	0.700	0.696	0.000
DE	0.889	0.071	0.667	0.071	0.875	0.103
CH	0.789	0.143	0.778	0.080	0.889	0.120
AF	0.821	0.487	0.750	0.256	1.000	0.306
EF	0.625	0.096	0.250	0.036	0.800	0.105
EG	0.875	0.072	0.625	0.024	0.857	0.083
EH	0.894	0.207	0.900	0.394	0.895	0.321
EJ	0.581	0.208	0.558	0.167	0.656	0.237
AG	0.690	0.211	0.483	0.053	0.813	0.294
DF	1.000	0.027	1.000	0.000	1.000	0.027
DH	0.000	0.056	0.000	0.000	0.000	0.053
DG	1.000	0.171	0.667	0.000	1.000	0.194
DJ	1.000	0.167	1.000	0.367	0.684	0.000
FH	0.667	0.149	0.333	0.000	0.231	0.233
FG	0.769	0.070	0.462	0.047	0.462	0.163
FJ	0.400	0.039	0.400	0.039	0.500	0.038
FK	0.844	0.083	0.844	0.083	0.828	0.185
DK	1.000	0.823	0.926	0.500	0.949	0.813
BE	0.714	0.263	0.661	0.368	0.773	0.355
BF	0.811	0.184	0.703	0.184	0.818	0.238
BH	0.894	0.464	0.617	0.643	0.971	0.537
BG	0.571	0.059	0.517	0.056	0.600	0.083

A4.5 Use of keywords for retrieval and relevance judgements

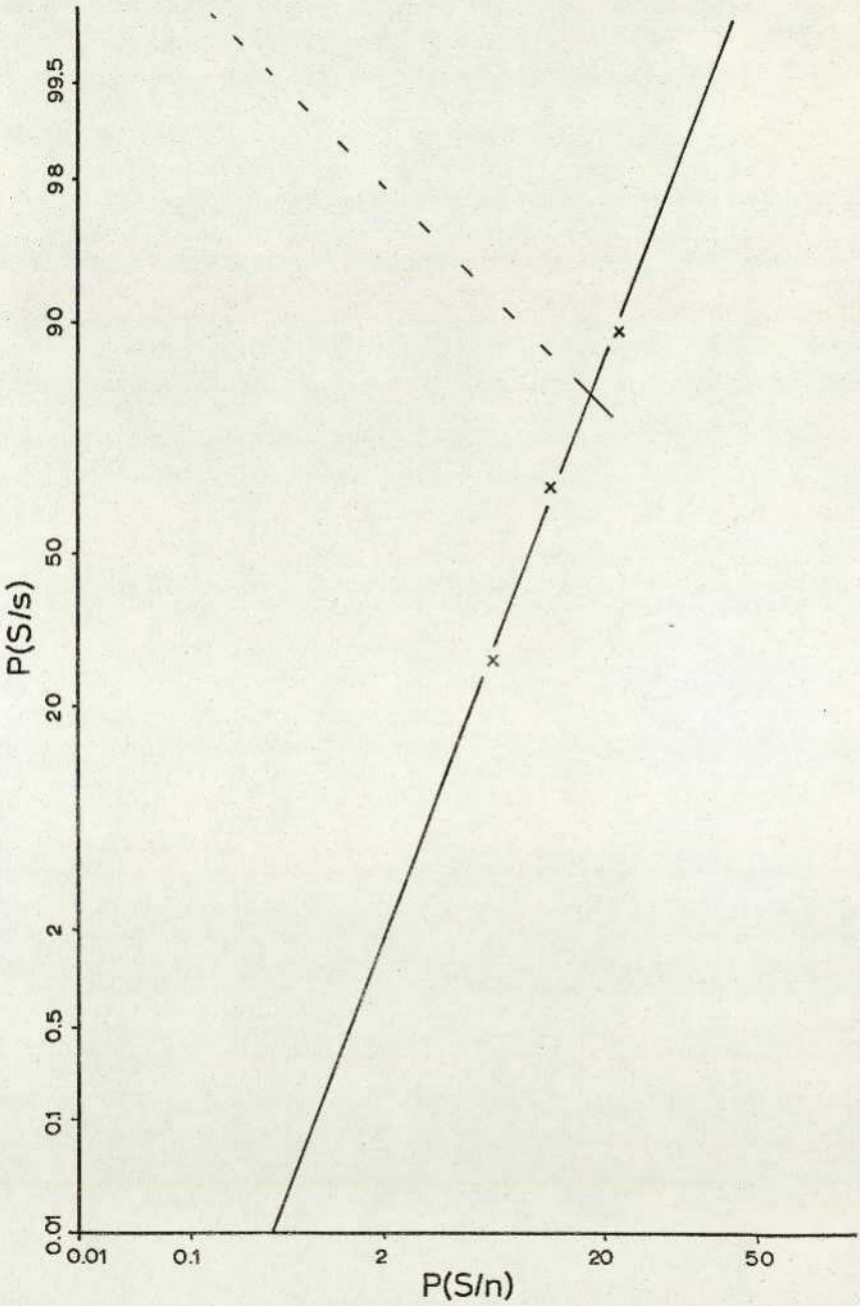
User	No. items retrieved	% Retrieved by K.W.	% Relevant retrieved by K.W.	% where K.W. for relevance
AB	30	43.3	3.33	-
AC	30	43.3	6.67	6.67
AD	115	35.7	8.70	13.9
AE	32	37.5	9.38	15.6
AF	155	40.6	7.74	1.29
AG	155	40.6	4.52	2.58
EF	262	46.9	1.53	1.15
EG	262	46.9	0.76	0.38
EH	262	46.9	4.58	4.12
EJ	262	46.9	13.7	5.73
DJ	38	47.4	5.26	2.63
DG	38	47.4	10.5	2.63
DH	38	47.4	-	-
DF	38	47.4	-	-
FG	56	75.0	14.3	12.5
FH	56	75.0	17.9	7.14
FJ	56	75.0	3.57	5.36
FK	56	75.0	37.5	19.6
BE	136	20.6	8.09	5.15
BF	136	20.6	5.88	5.15
BH	136	20.6	8.09	5.88
BG	136	20.6	-	-
DK	148	25.0	6.76	7.43
Aver.	114.5	43.0	7.16	5.43

APPENDIX 5

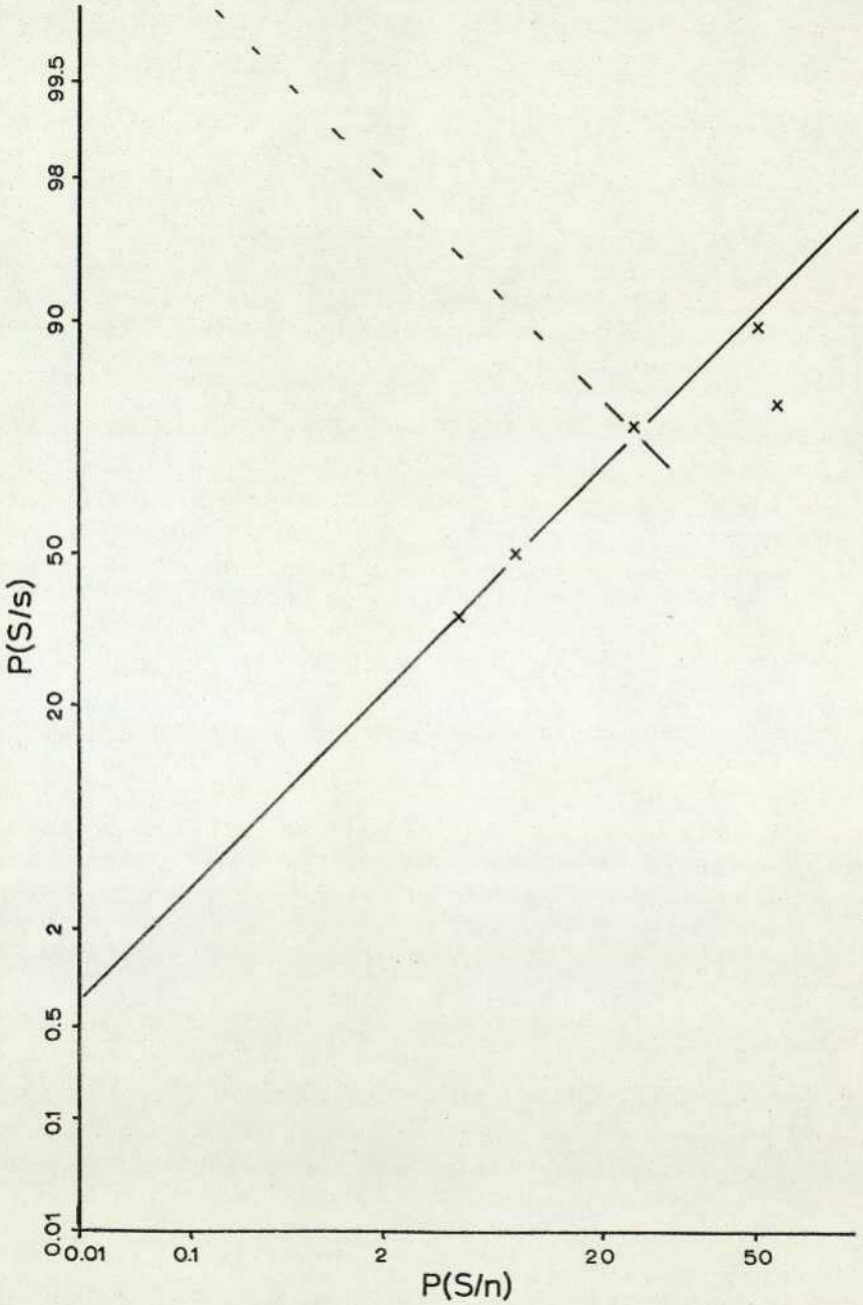
Table of graphs

A5.1 O.C. curves for eighteen users

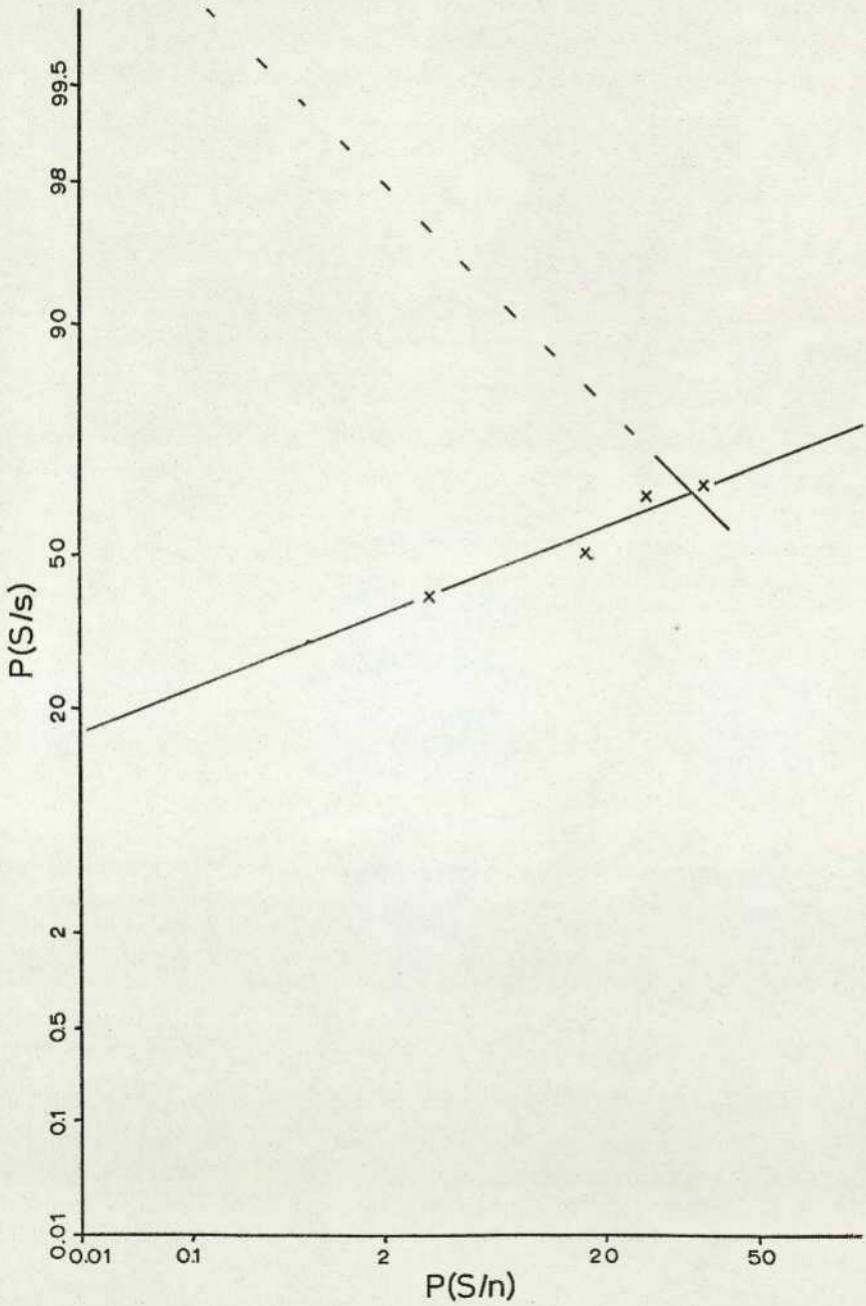
Graph 1 O.C. User AE



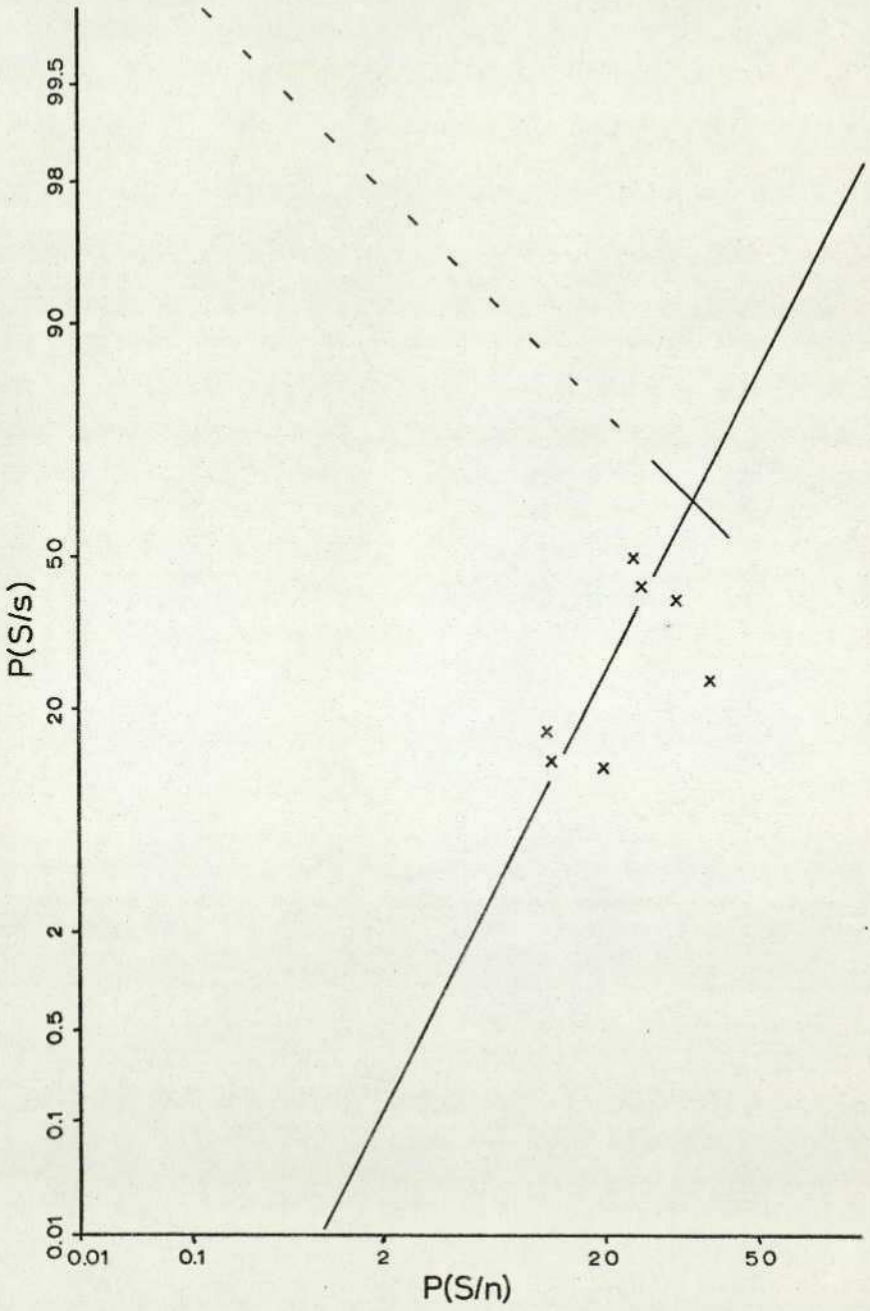
Graph 2 O.C. User AF



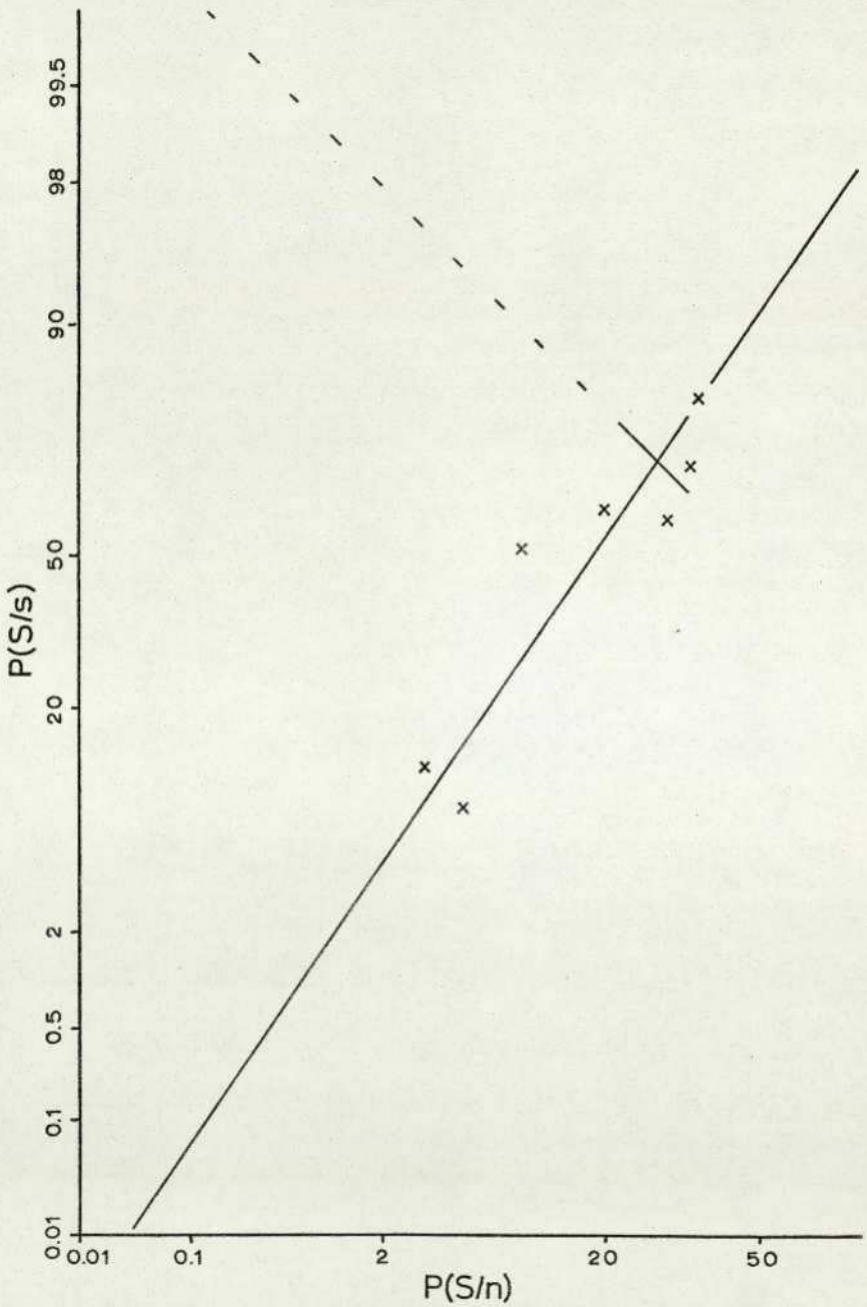
Graph 3 O.C. User CD



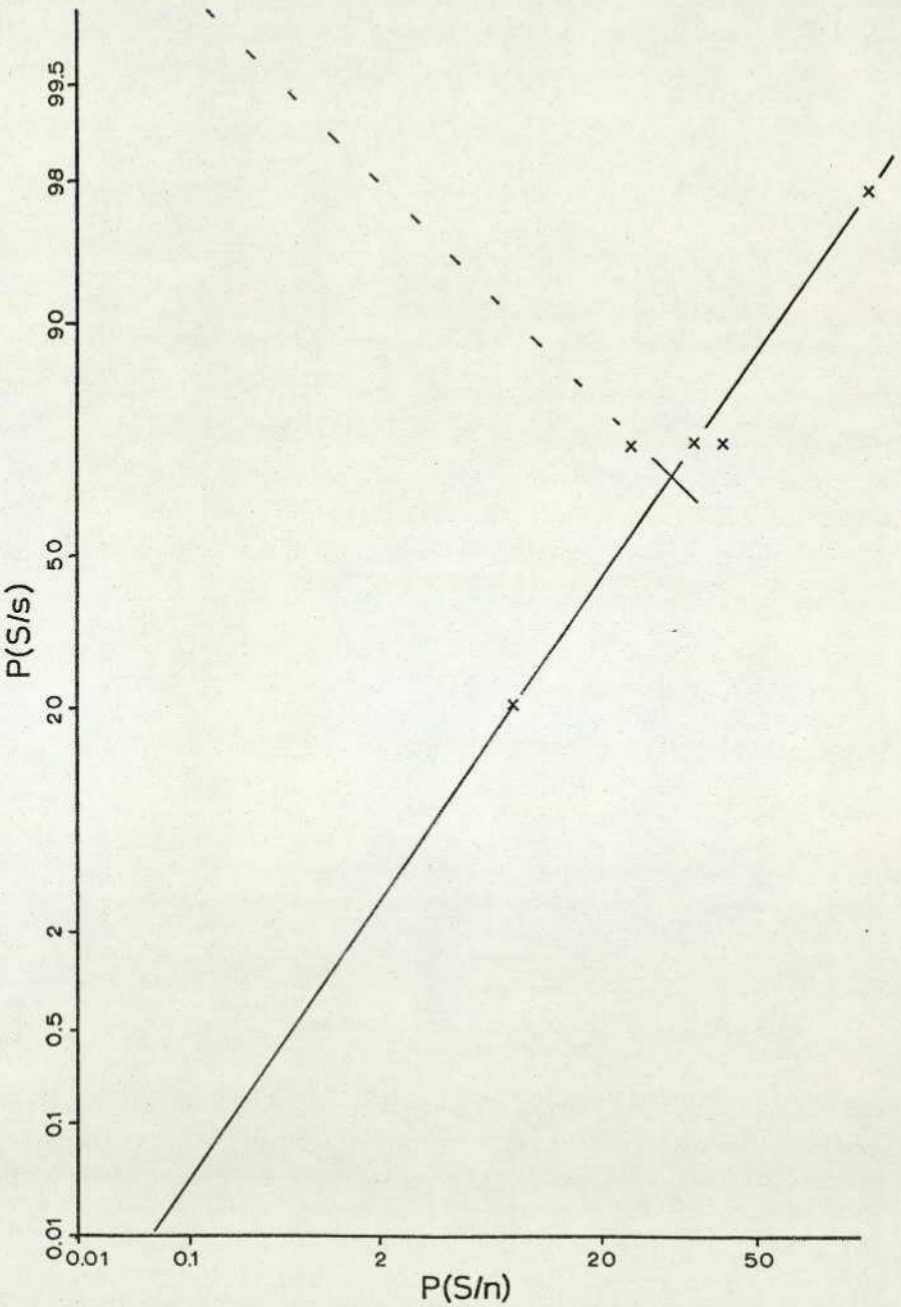
Graph 4 O.C. User CF



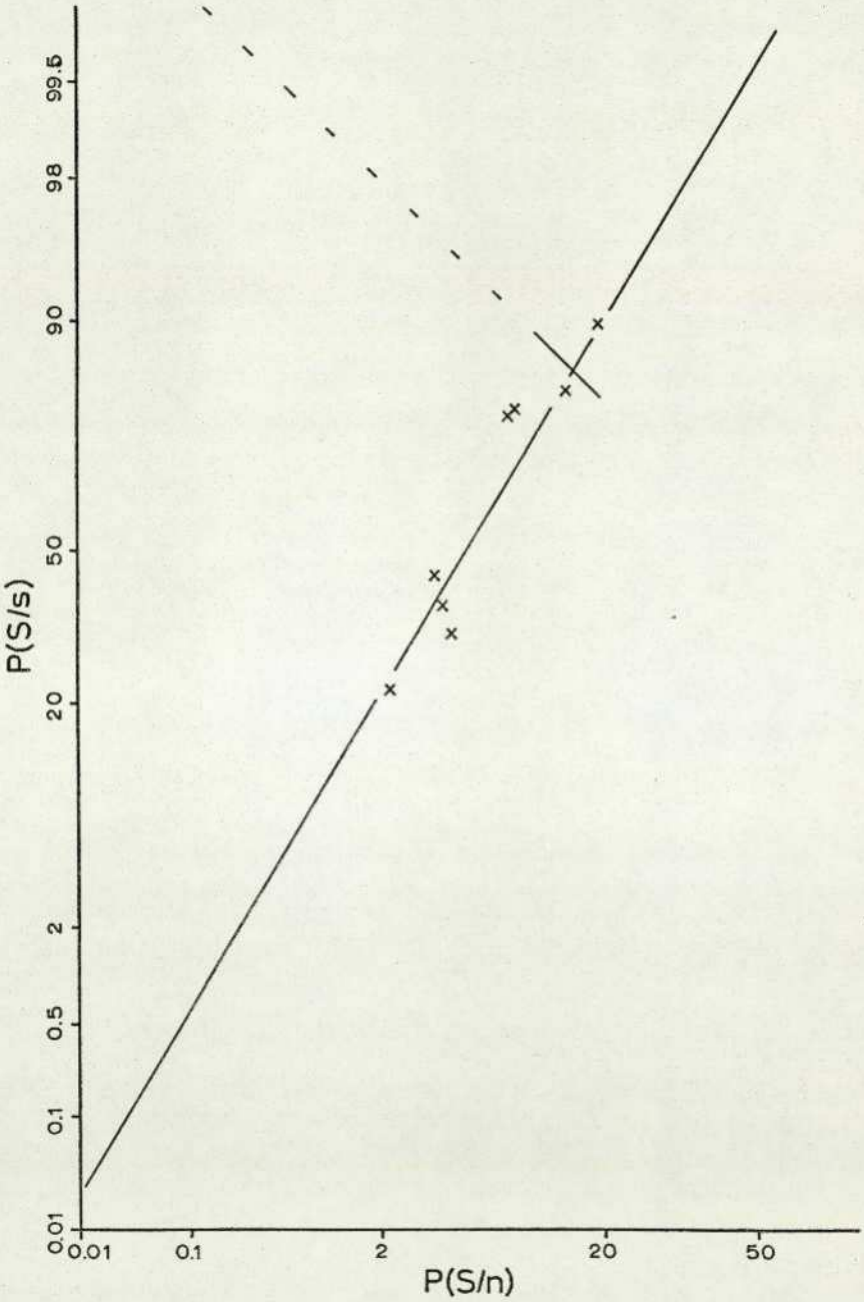
Graph 5 O.C. User CG



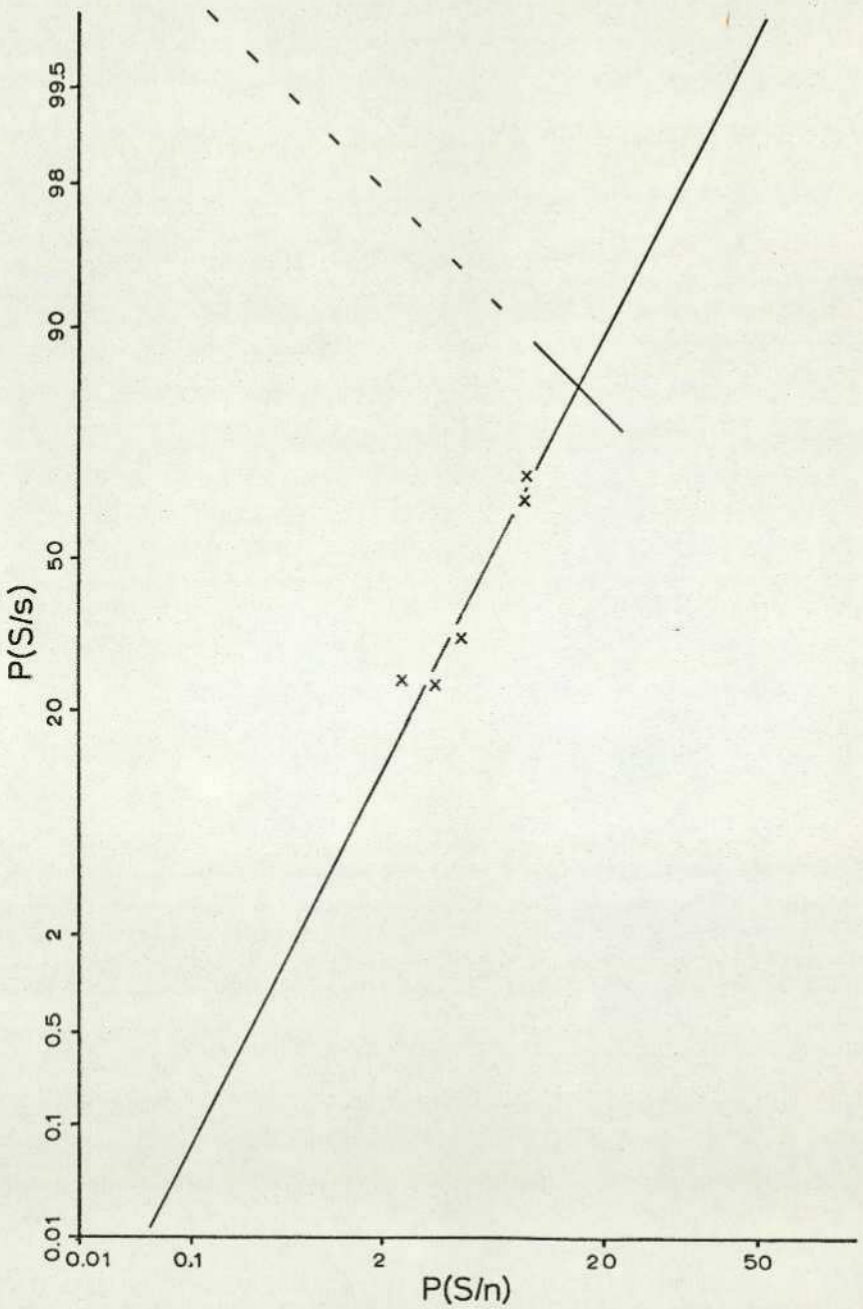
Graph 6 O.C. User BD



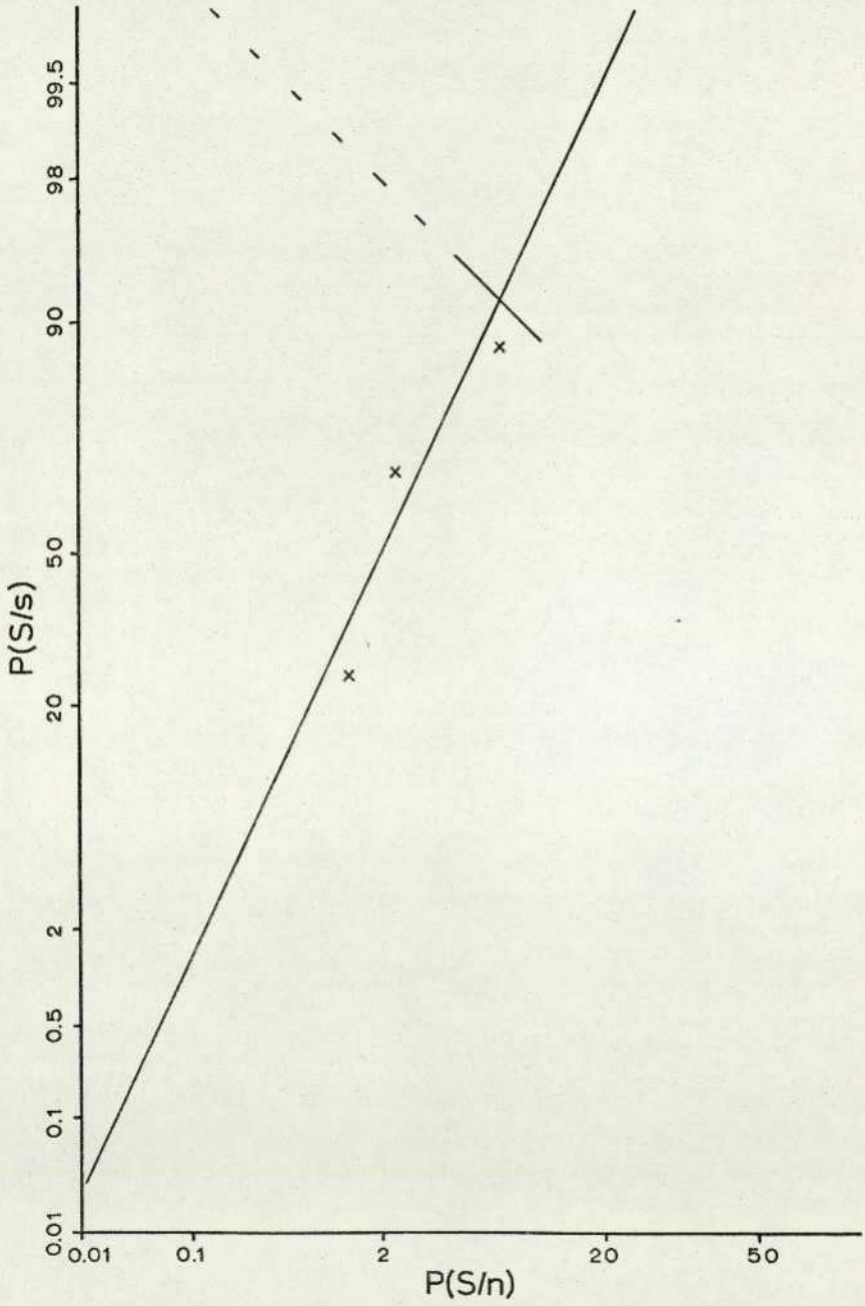
Graph 7 O.C. User CH



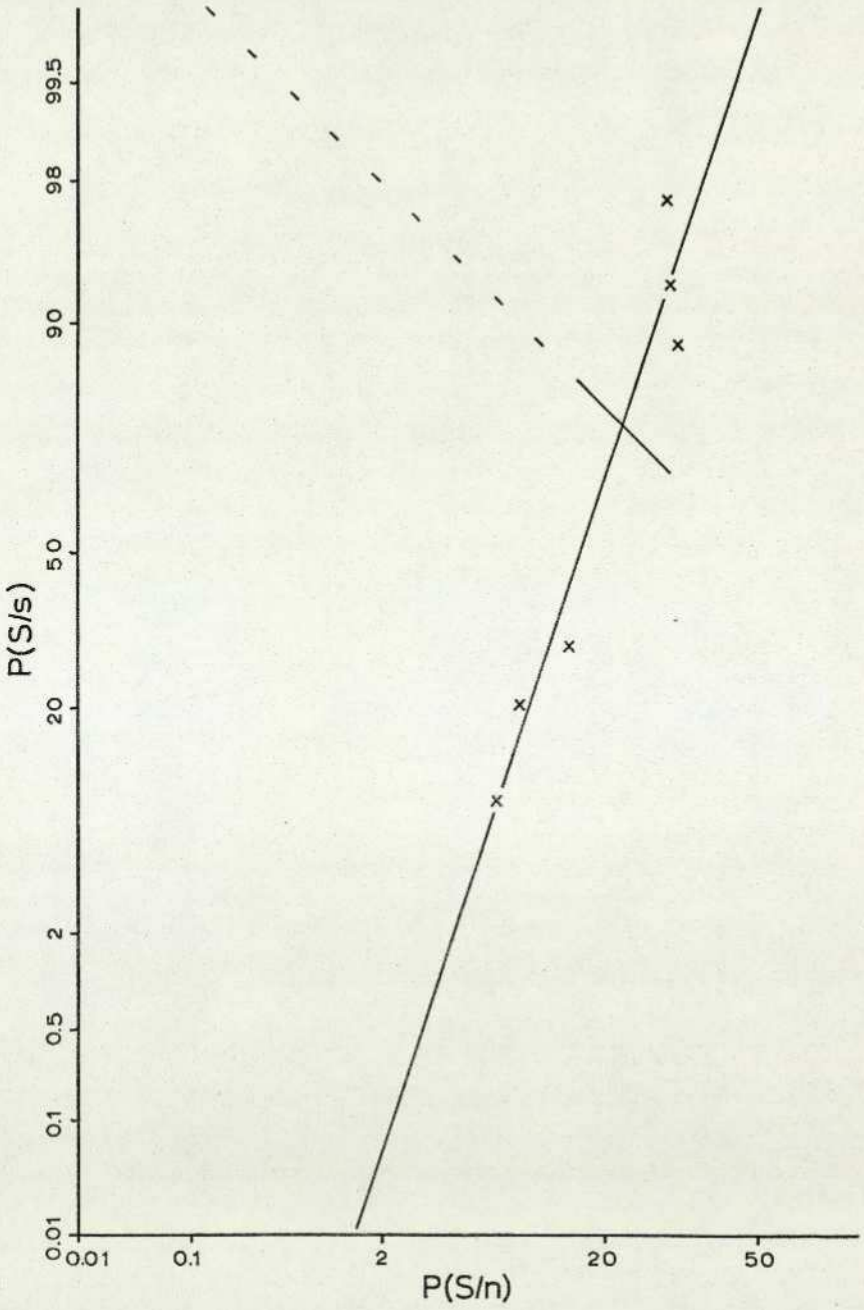
Graph 8 O.C. User EF



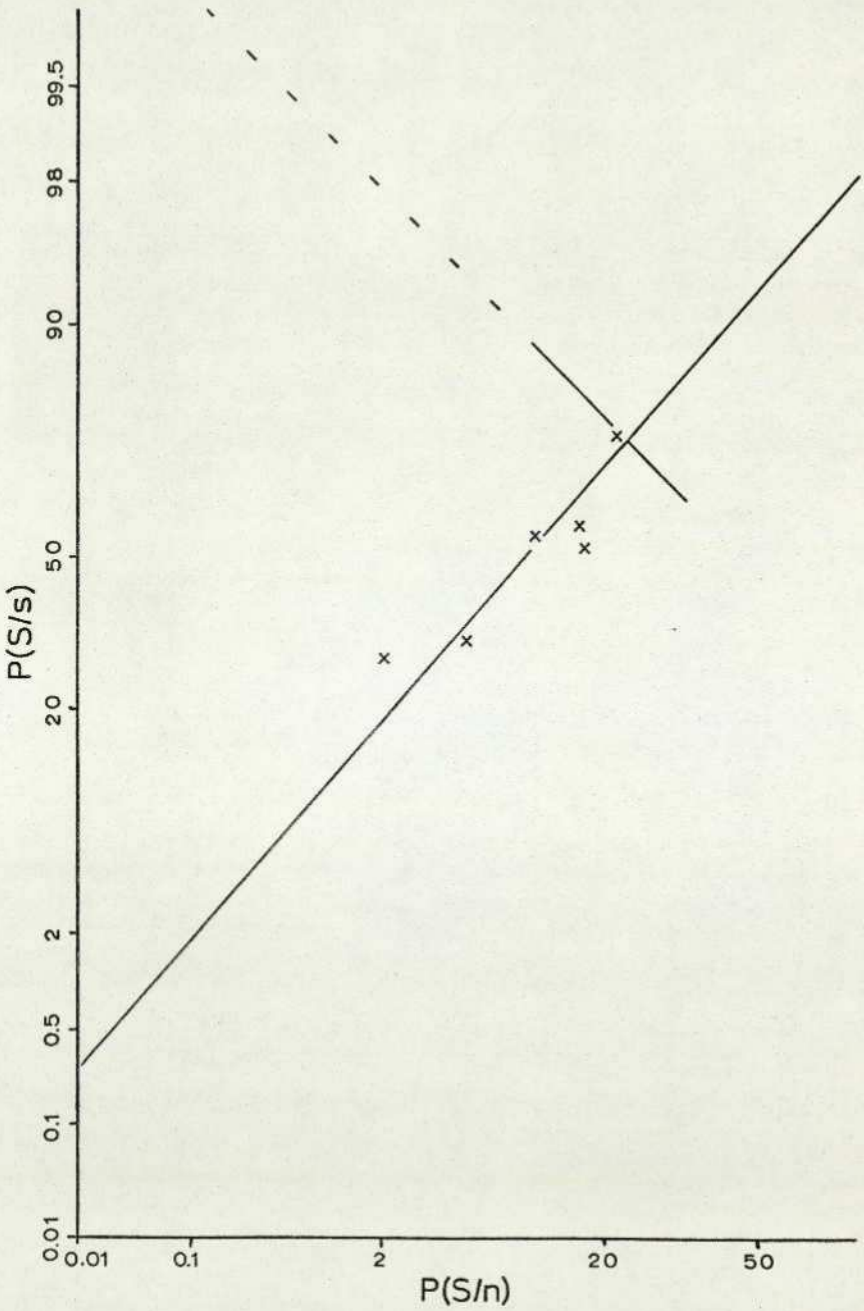
Graph 9 O.C. User EG



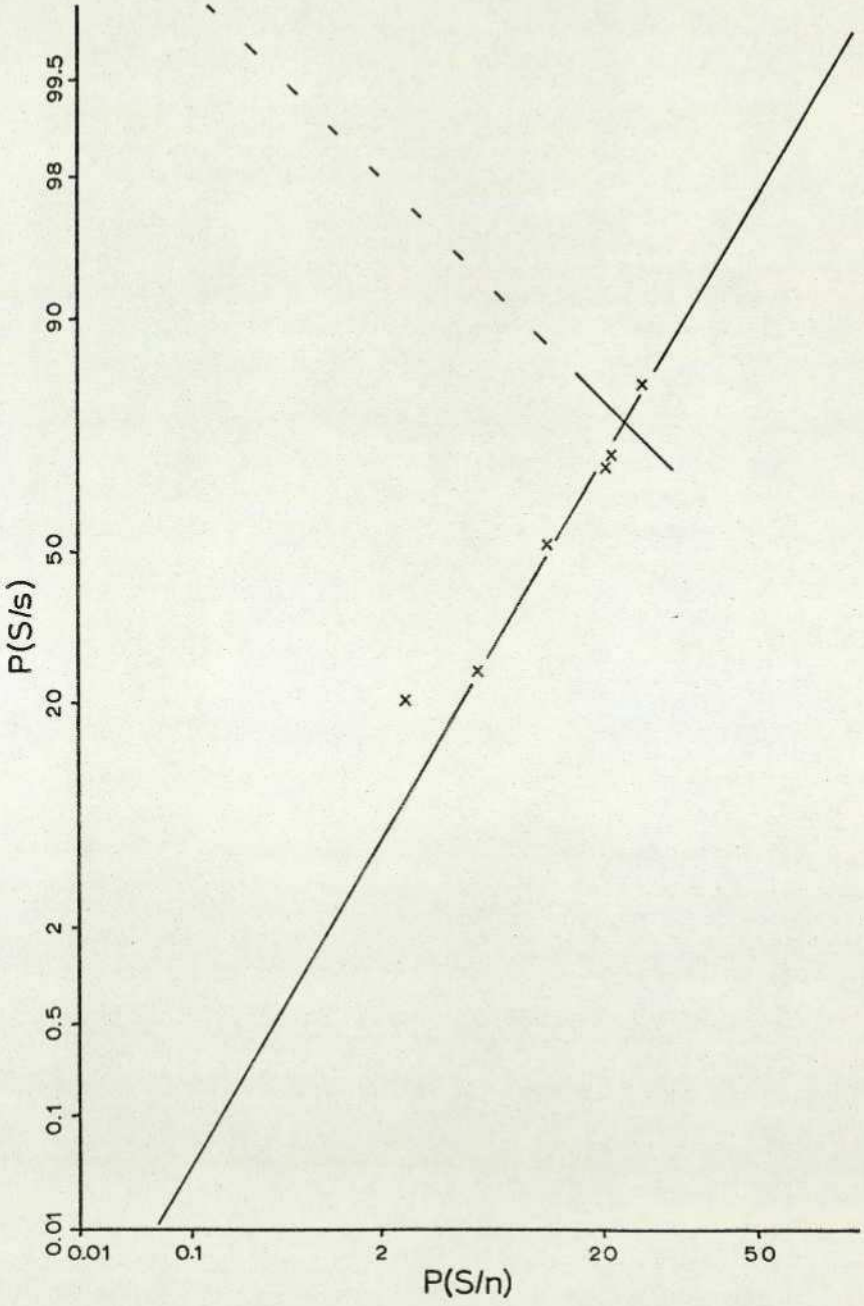
Graph 10 O.C. User EH



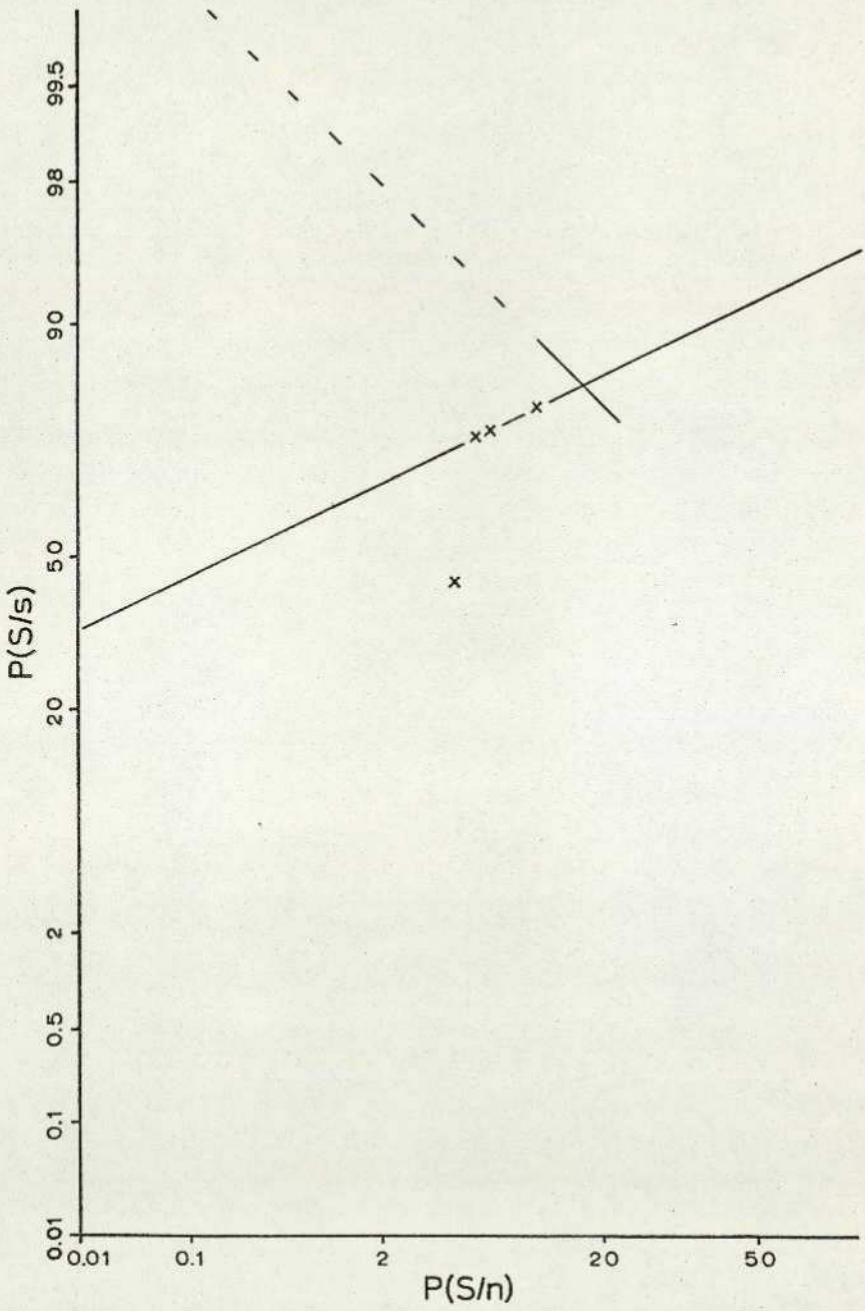
Graph 11 O.C. User EJ



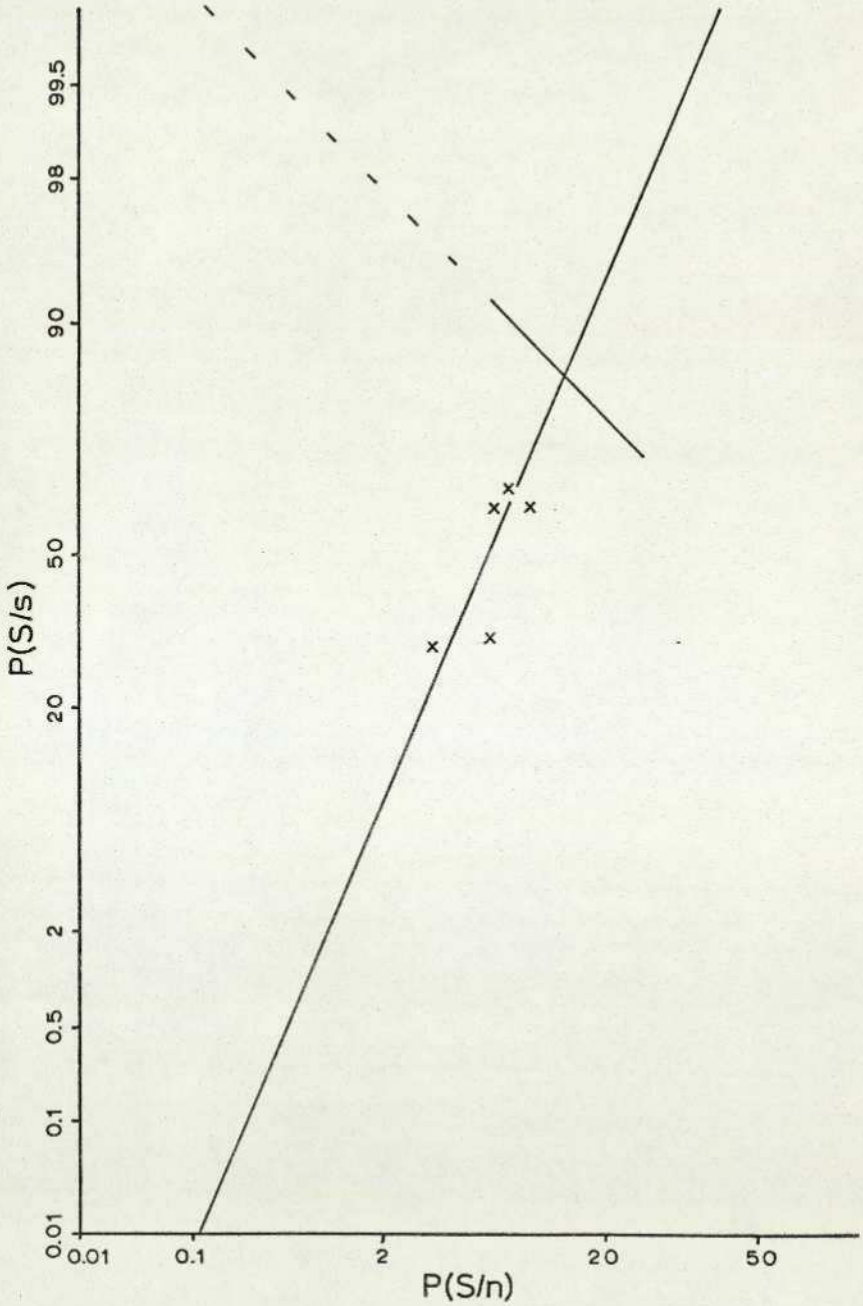
Graph 12 O.C. User AG



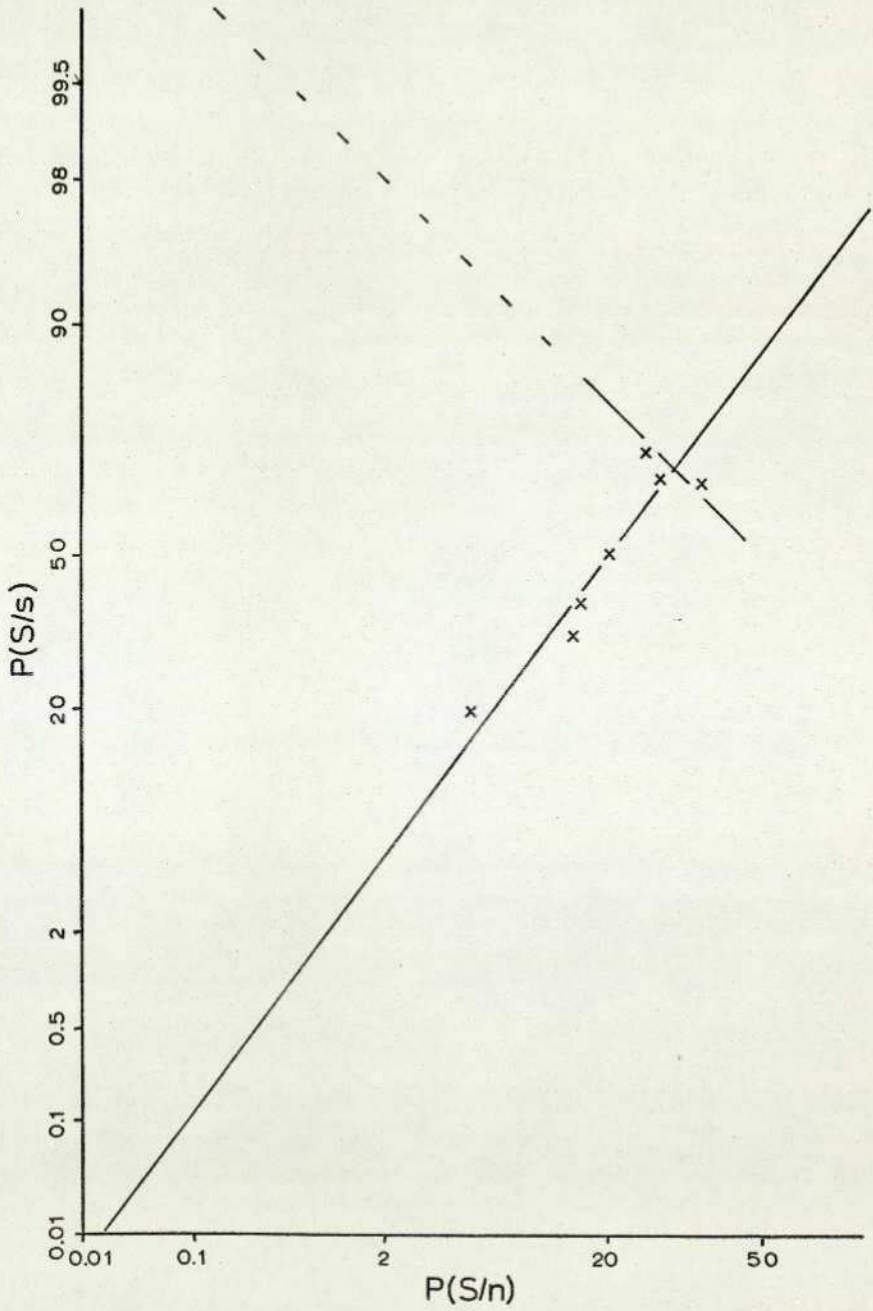
Graph 13 O.C. User FG



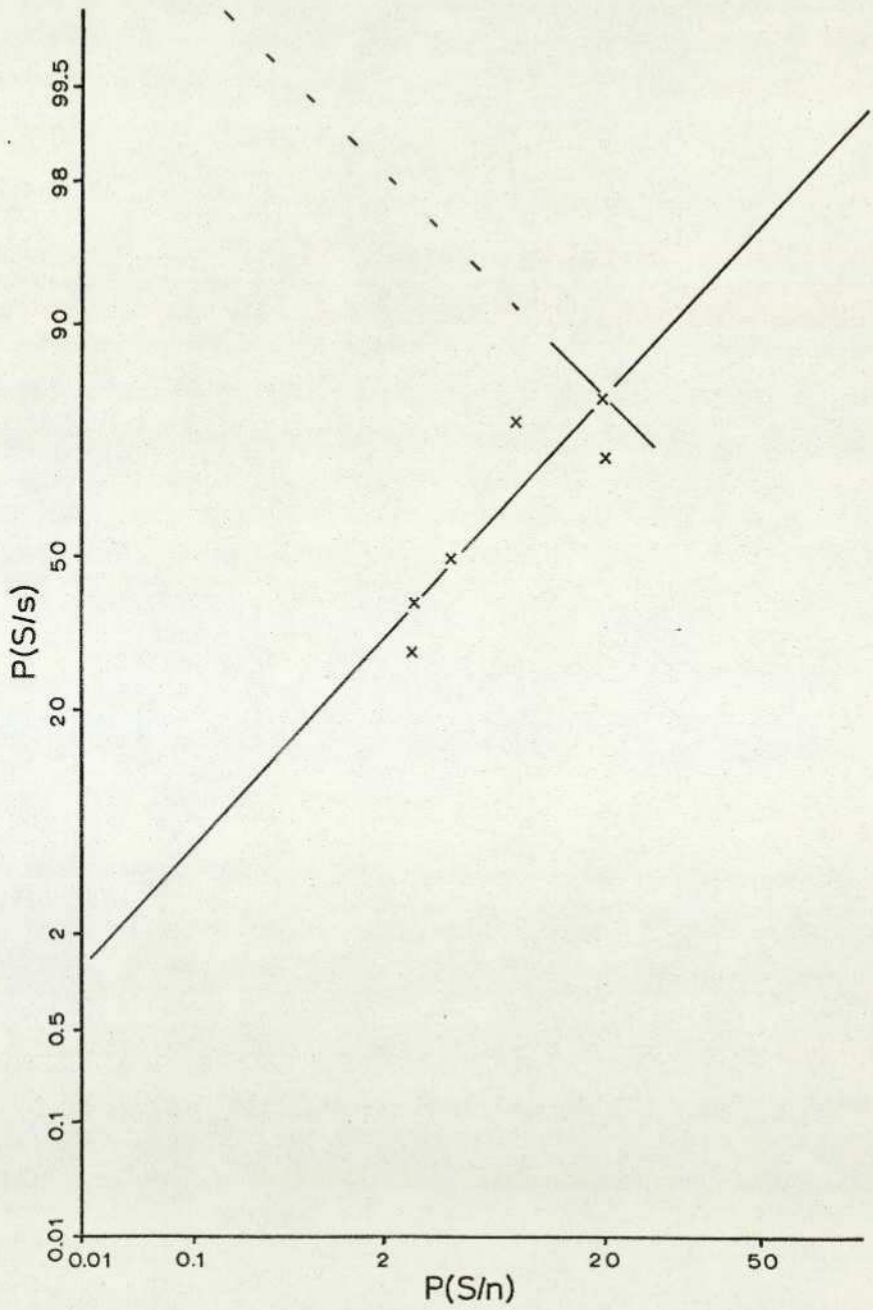
Graph 14 O.C. User FK



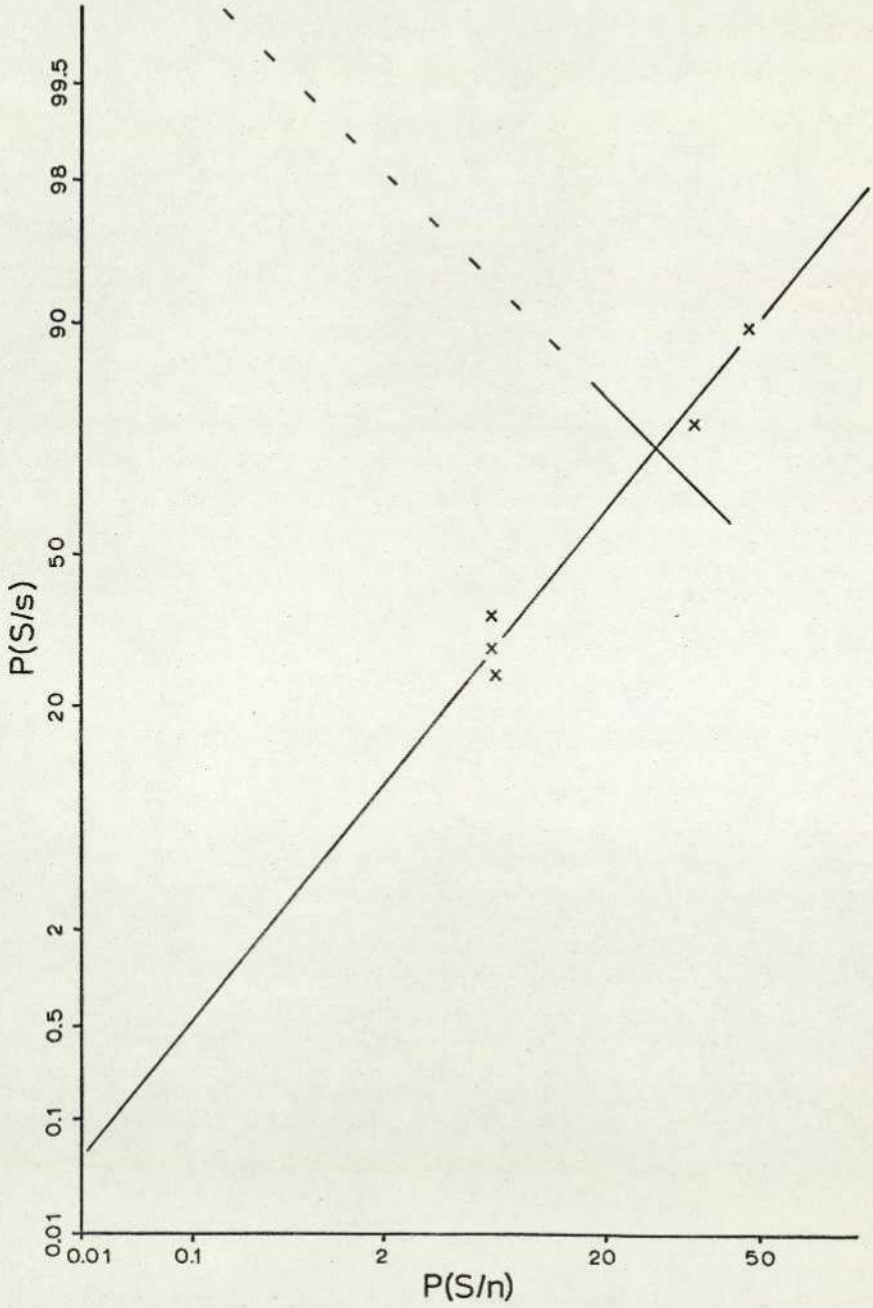
Graph 15 O.C. User BE



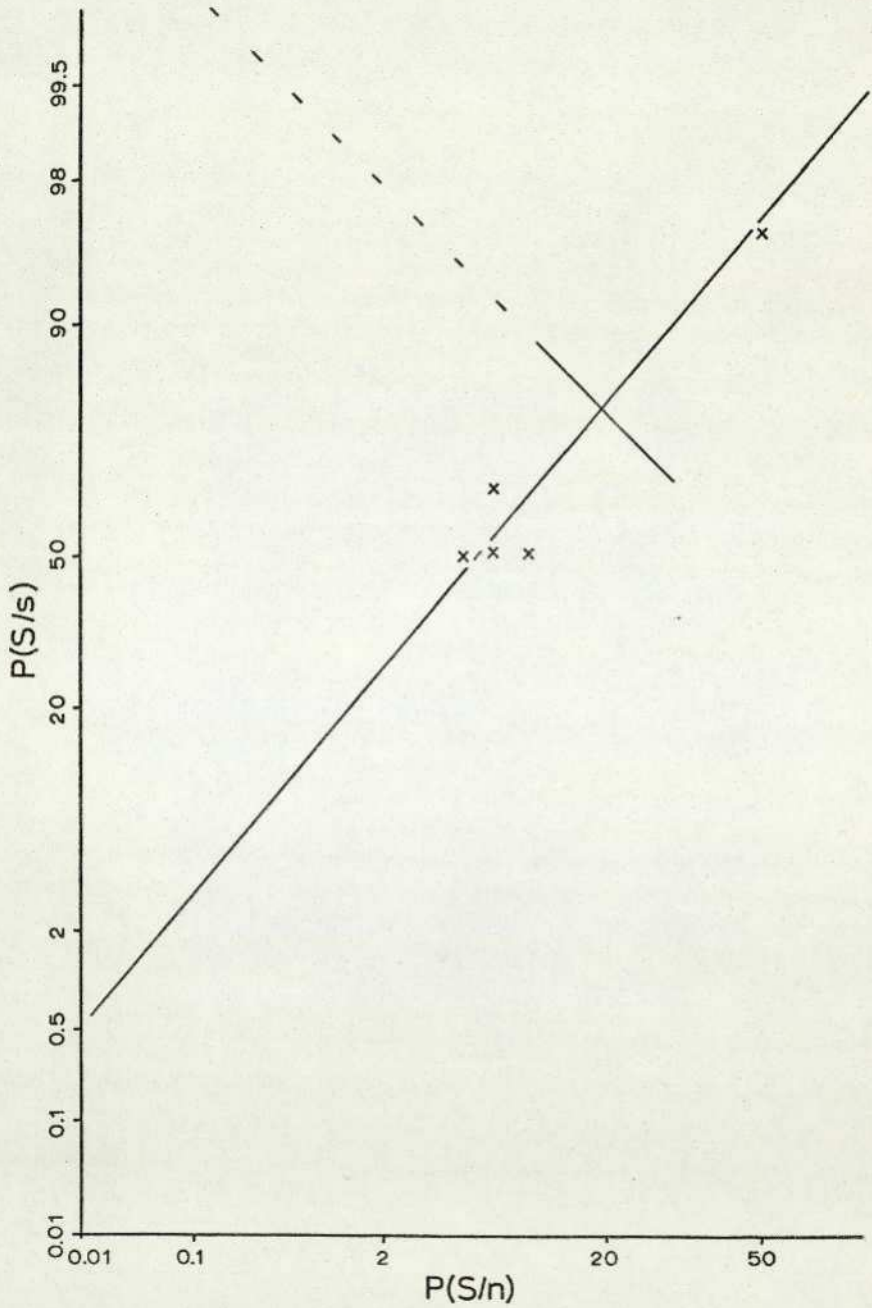
Graph 16 O.C. User BF



Graph 17 O.C. User BH

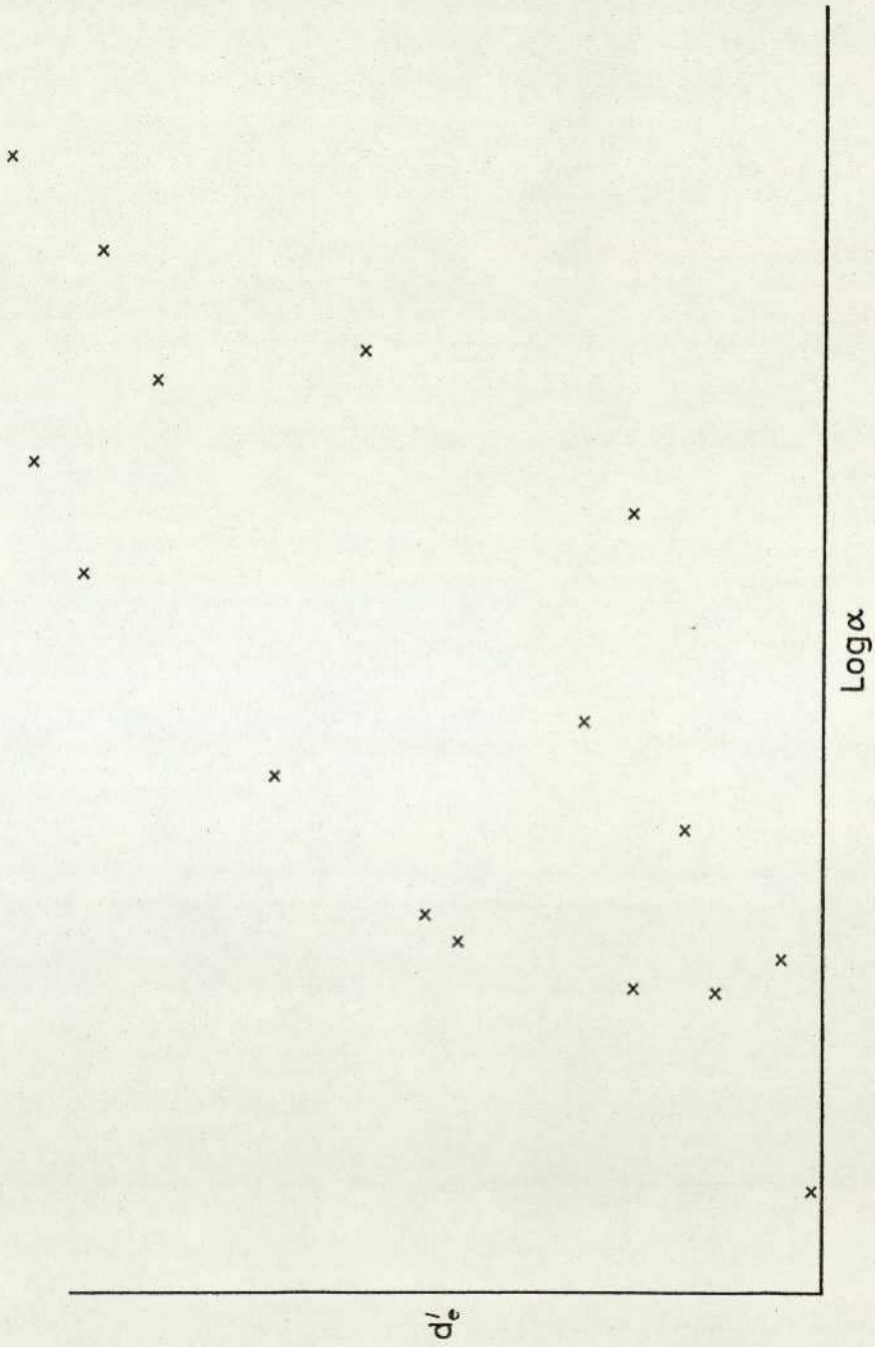


Graph 18 O.C. User DK

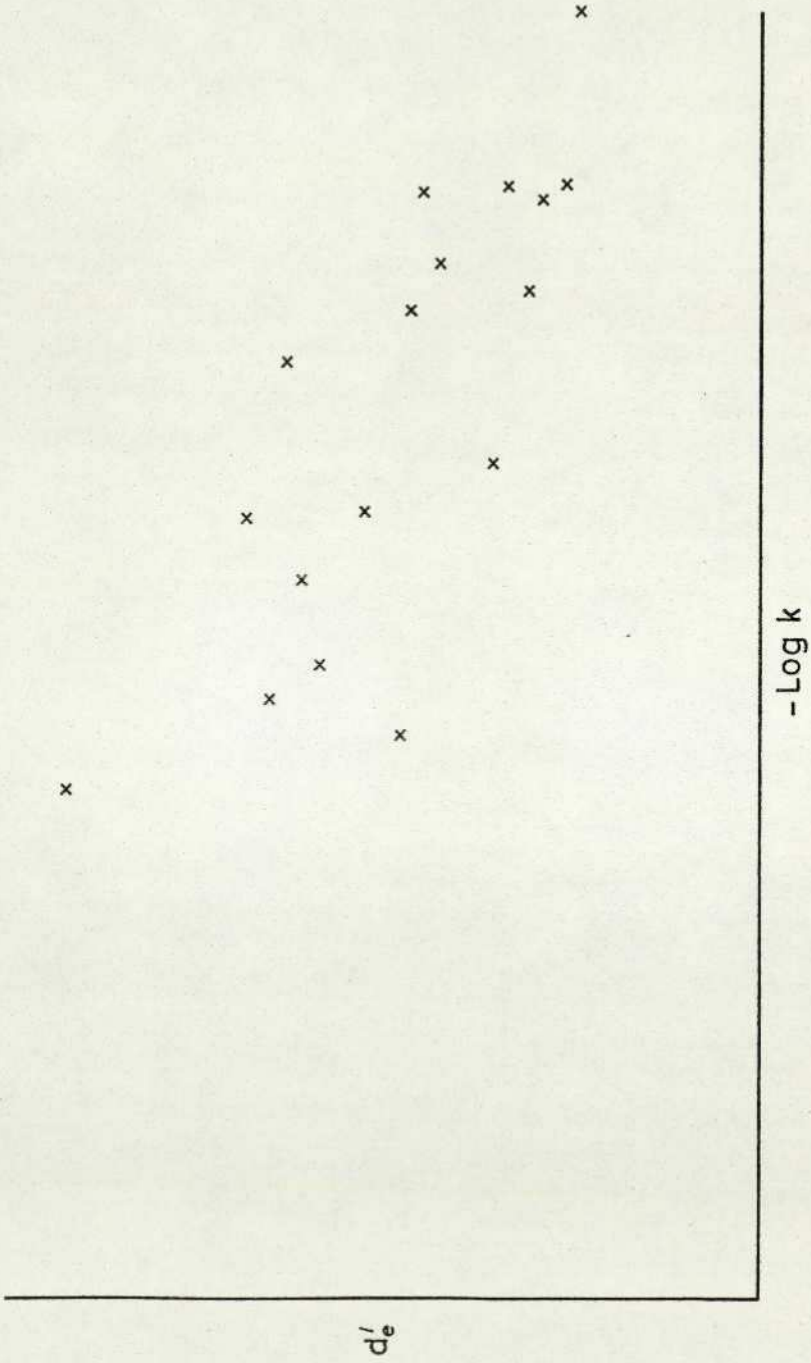


A5.2 Comparison of sensitivity measures

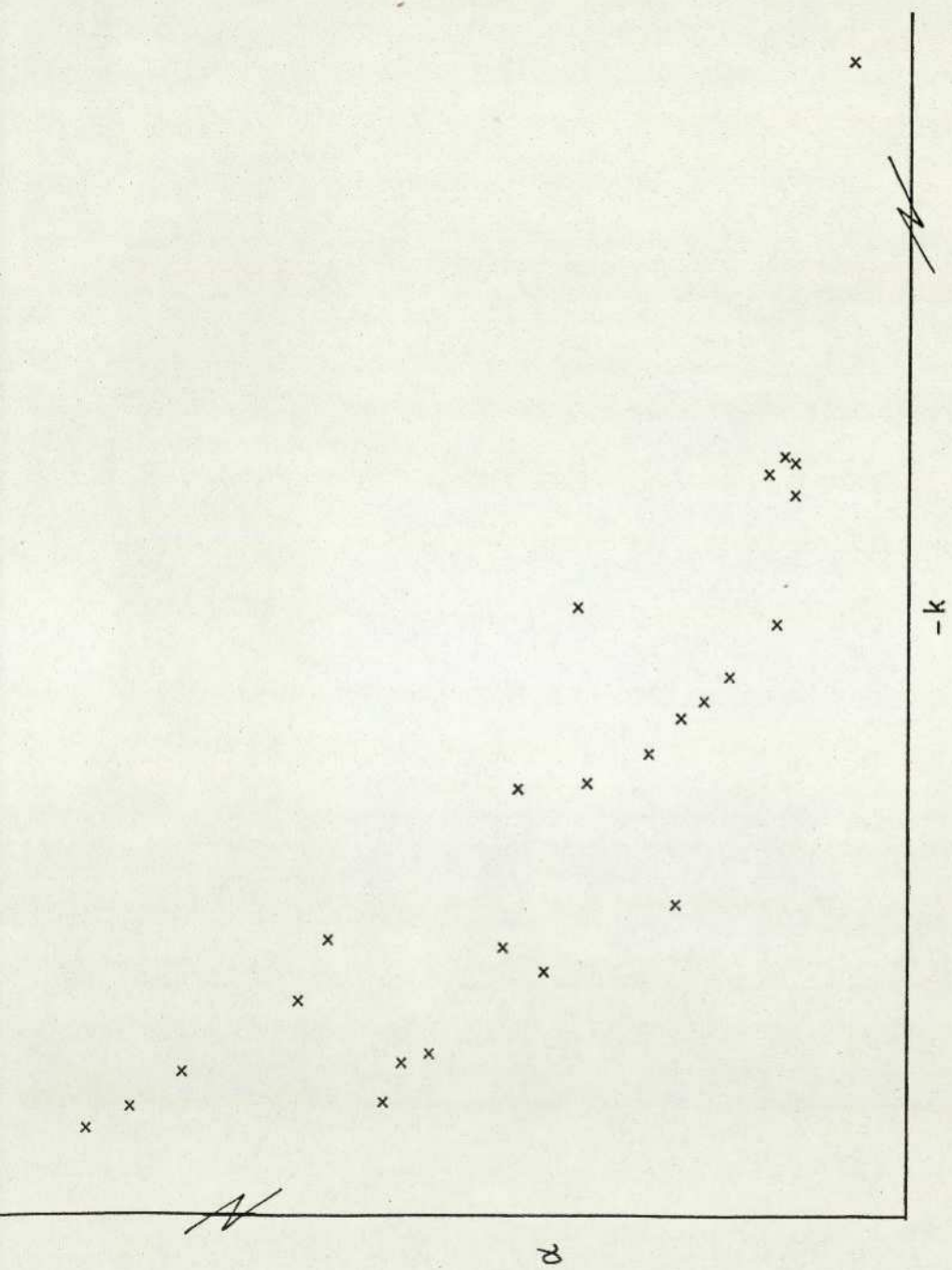
Graph 19 d'_e vs Logα



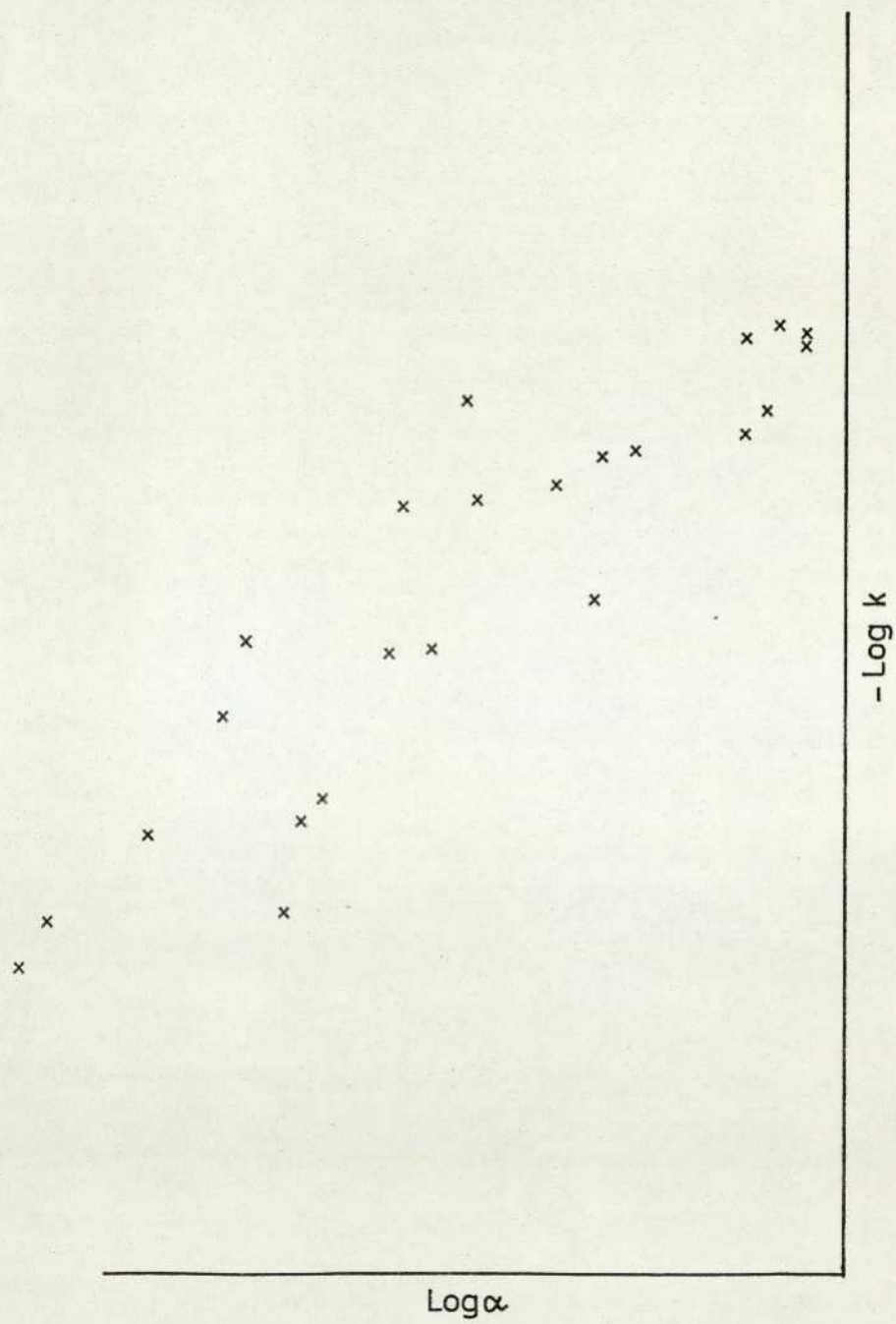
Graph 20 d'_e vs -Log k



Graph 21 α vs -k

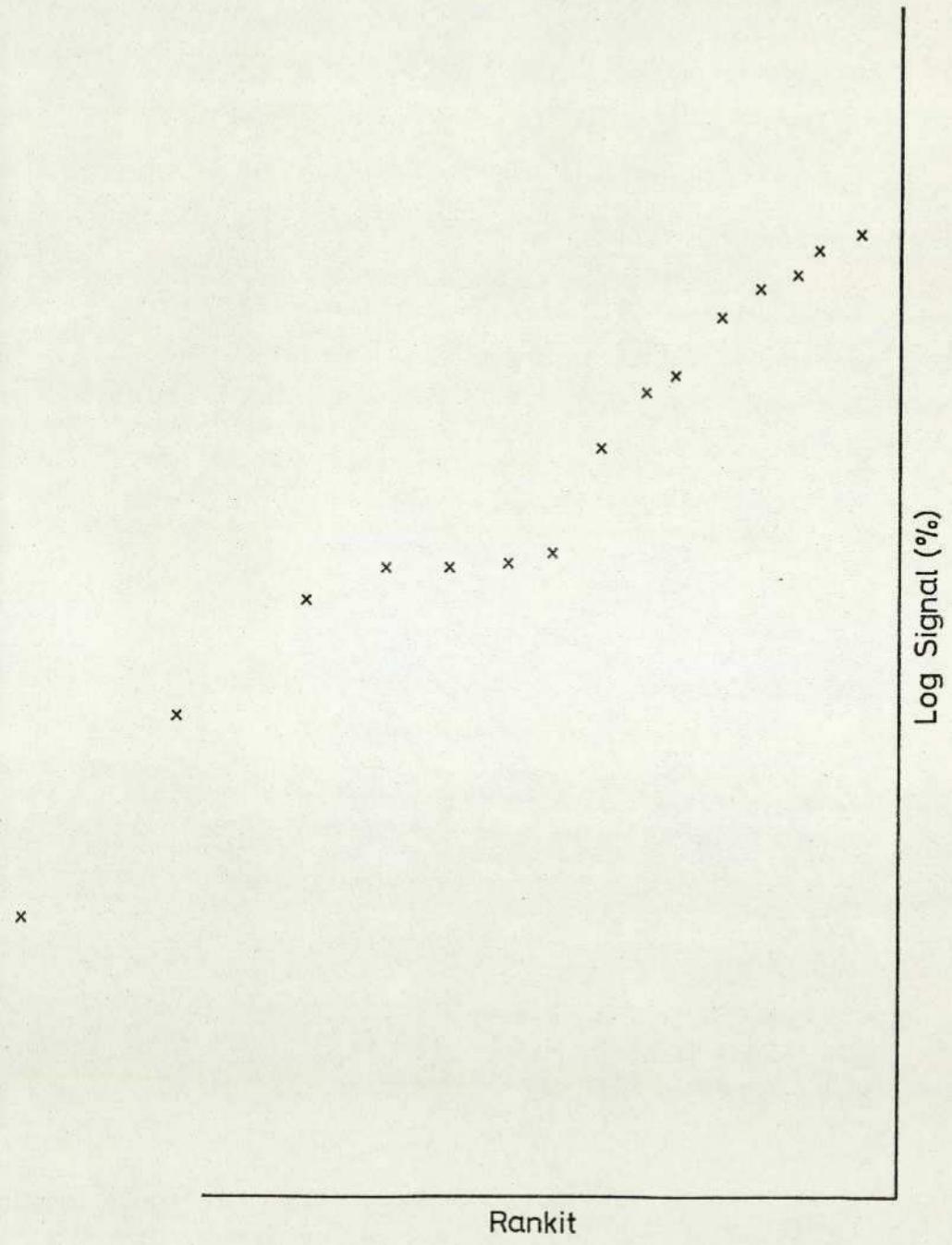


Graph 22 Log α vs -Log k

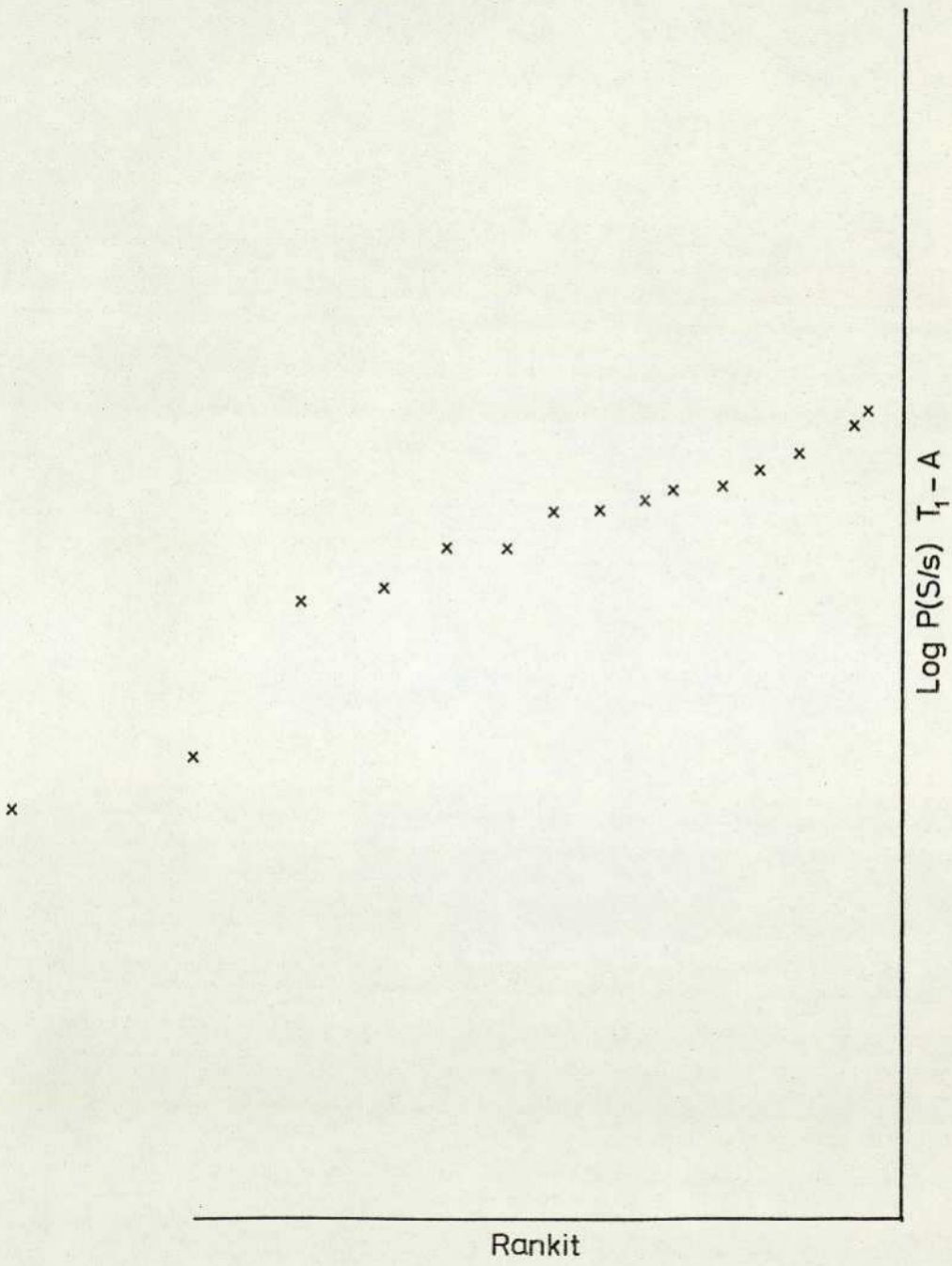


A5.3 Rankit plots

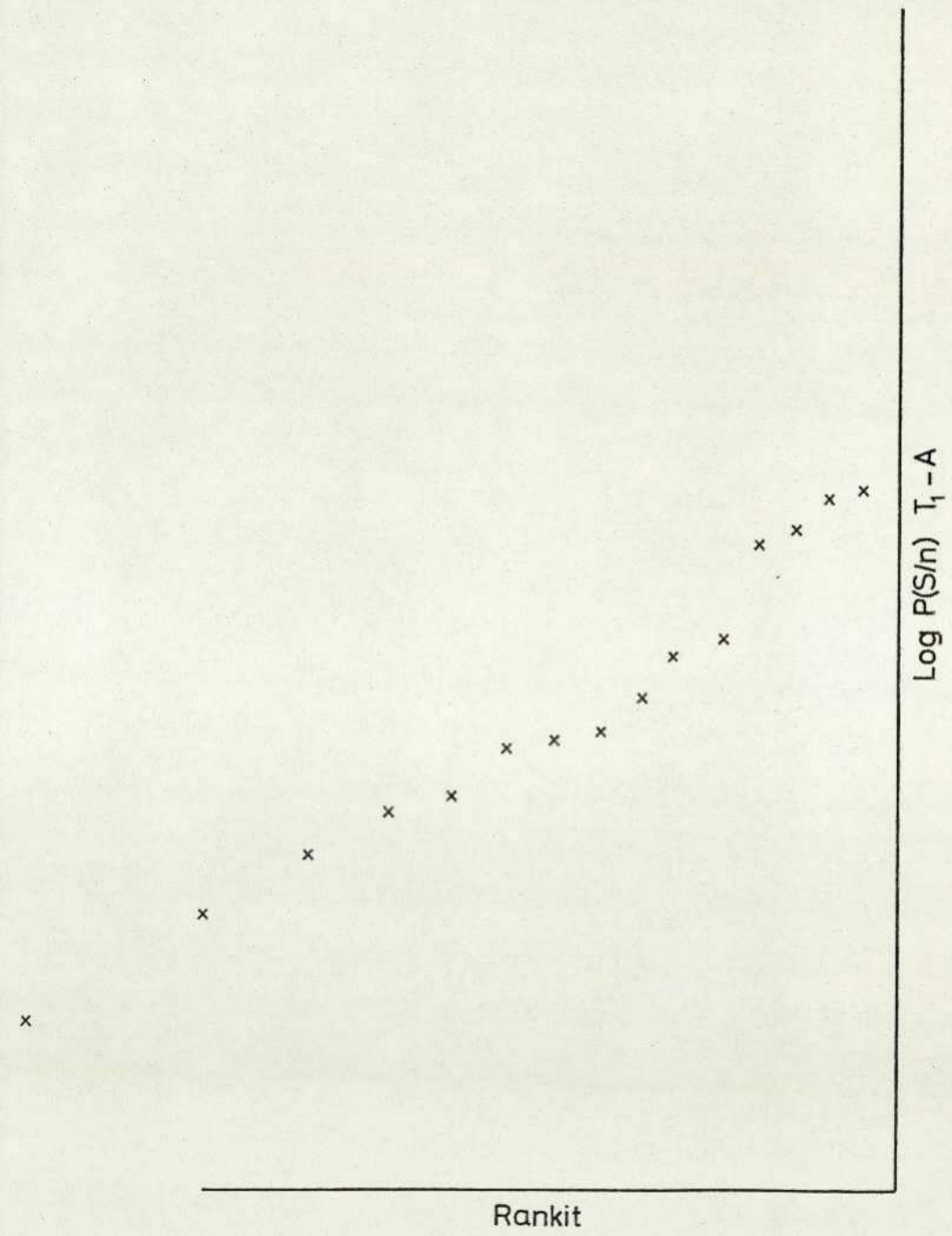
Graph 23 Rankit vs Log Signal (%)



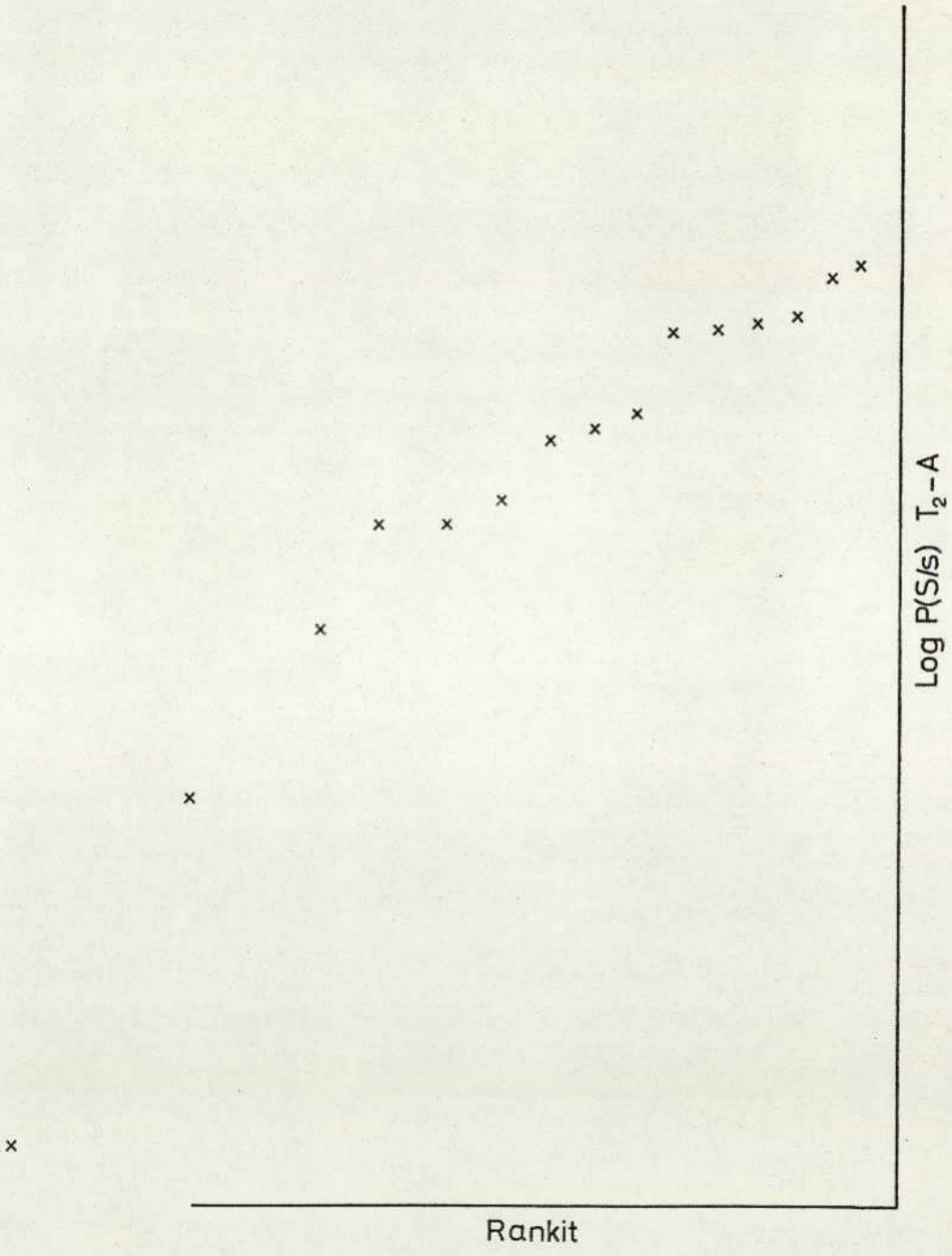
Graph 24 Rankit vs Log P(S/s) $T_1 - A$



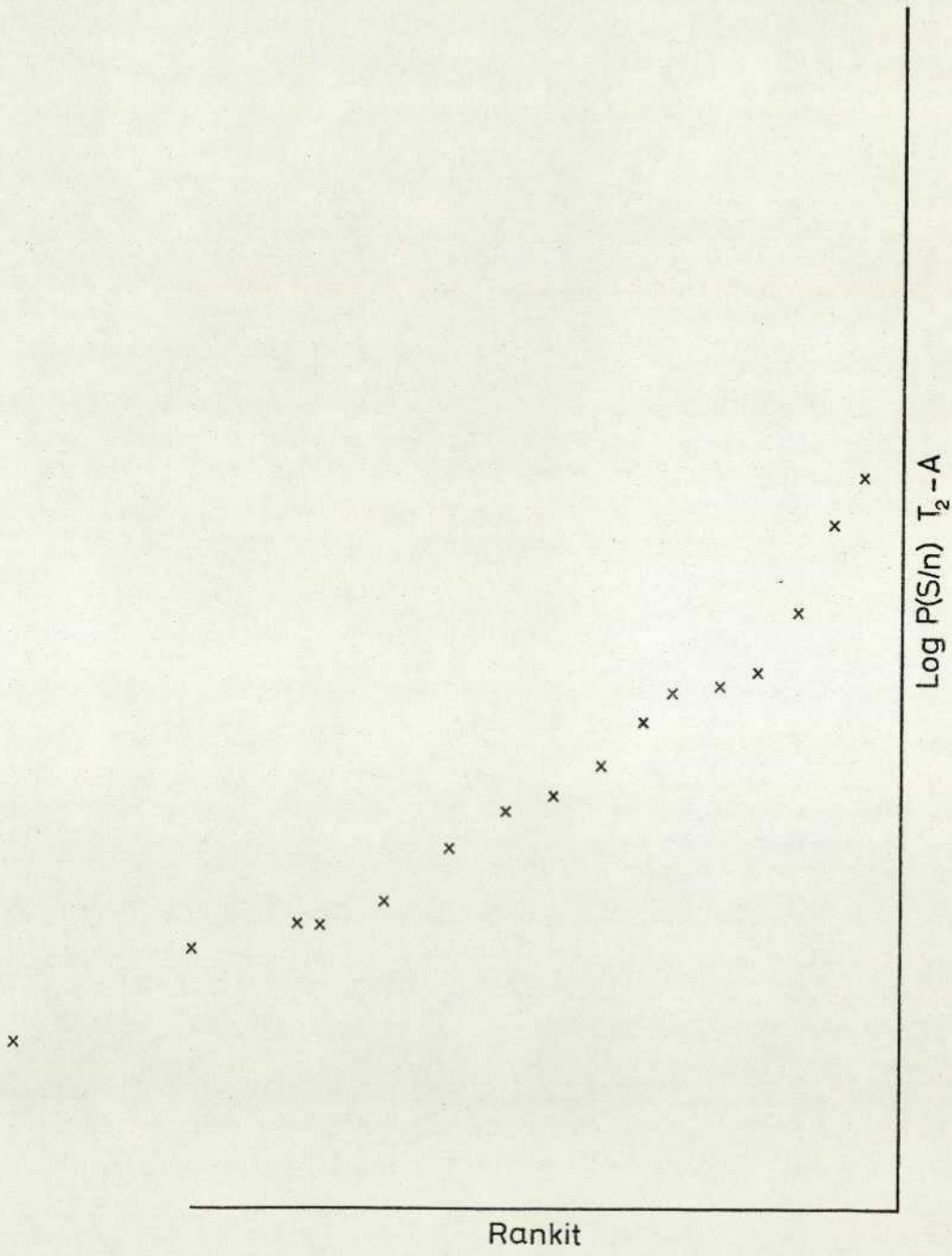
Graph 25 Rankit vs Log P(S/n) $T_1 - A$



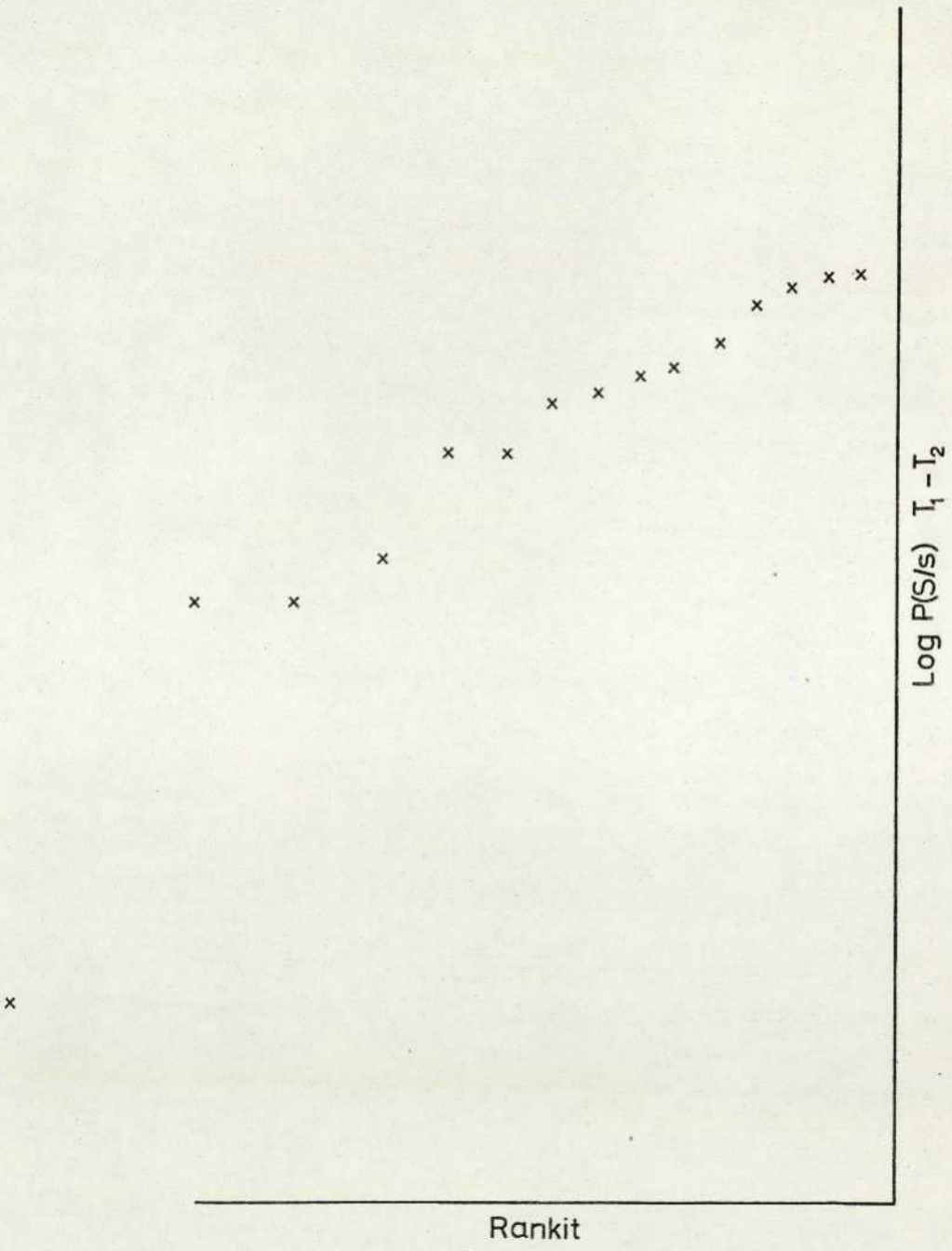
Graph 26 Rankit vs Log P(S/s) T_2 - A



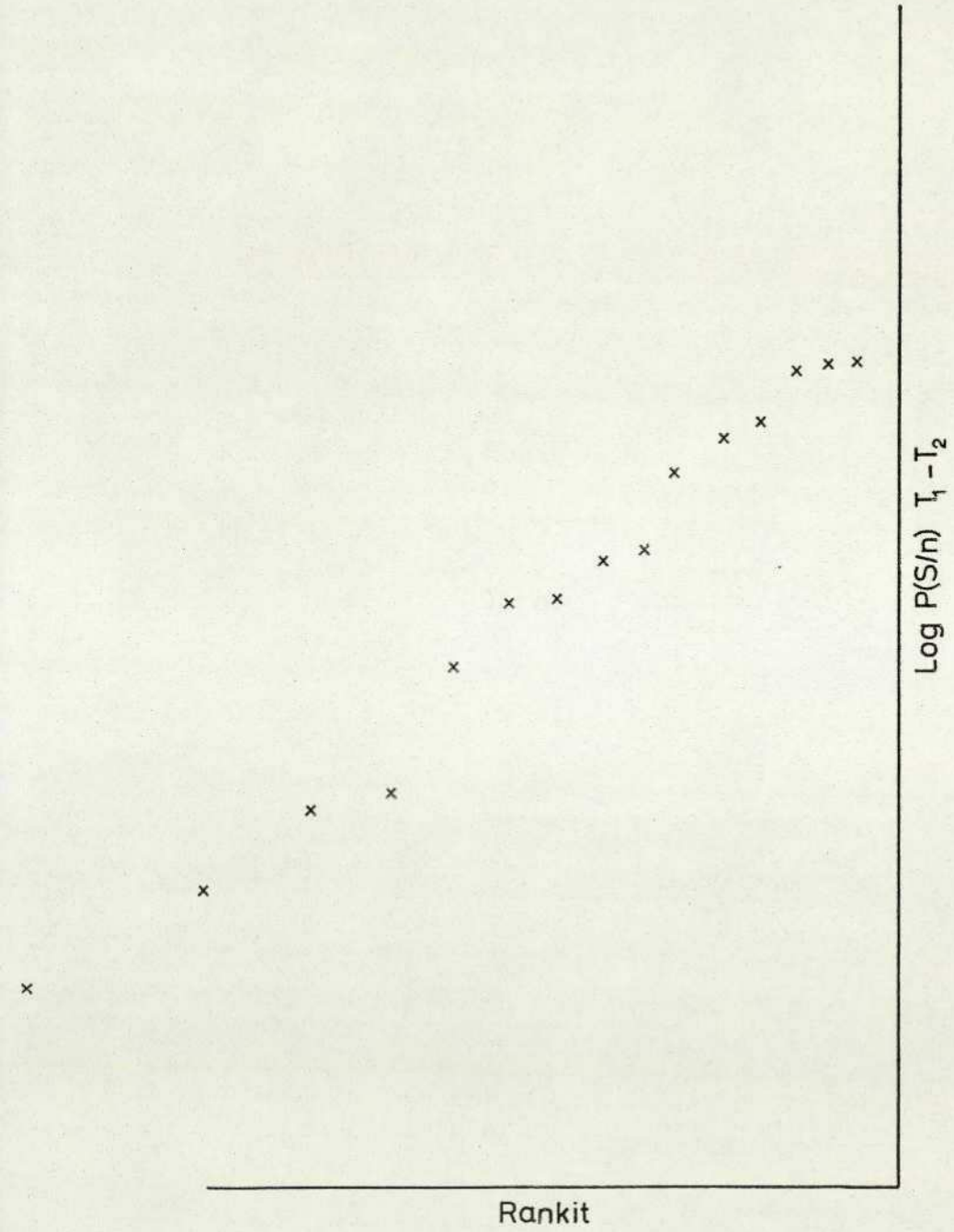
Graph 27 Rankit vs Log P(S/n) T_2-A



Graph 28 Rankit vs Log P(S/s) $T_1 - T_2$



Graph 29 Rankit vs Log P(S/n) $T_1 - T_2$



REFERENCES

1. E.M.Arnett (1970), Computer based chemical information services, Science,170,1370-76,1970
2. R.C.Atkinson and J.F.Juola (1974), Search and decision processes in recognition memory, In D.H.Krantz, R.C.Atkinson, R.D.Luce and P.Suppes (eds.), Contemporary Developments in Mathematical Psychology, Vol.1, W.H.Freeman and Co.,San Francisco,1974
3. G.C.Barhydt (1964), Comparison of relevance assessments by three types of evaluator, Proc. ADI,1,383-85,1964
4. G.C.Barhydt(1967) Effectiveness of non-user relevance assessments,J.Doc.,23,146-49,1967
5. F.H.Barker,A.K.Kent and D.C.Veal (1970), Report on the evaluation of an experimental computer-based current awareness service for chemists, UKCIS Research Report No. 1, The Chemical Society, London,1970
6. D.Bell(1973), A study of the use by scientists of photocopying and microforms,MSc Thesis,University of Strathclyde,1973
7. J.Belzer (1974), Information theory as a measure, In A.Debons (ed.), Information Science Search for Identity, Marcel Dekker,New York,1974
8. J.D.Bernal (1948),Preliminary analysis of a pilot questionnaire on the use of scientific literature, Proc. Royal Soc. Sci.Inf. Conf.,1948,Royal Society,London 1948
9. J.D.Bernal (1960), Scientific information and its users, Aslib Proc.,12,432-38,1960
10. R.T.Bottle and C.R.Seeley (1970), Information transfer limitations of titles of chemical documents,J.Chem. Doc.,10,256-59,1970
11. C.P.Bourne (1962), Review of the methodology of information systems design, In Information Systems Workshop, Spartan Books,Washington,1962
12. D.A.Broadbent (1971), Decision and Stress,Academic Press, London,1971
13. B.C.Brookes (1976), A new paradigm for information science? Inf.Sci.,10,103-11,1976
14. M.N.Carrington (1973), On the study of some variables in the measurement of relevance,MSc Thesis, City University, 1973

15. J.M.Carroll and J.M.Tague (1973), Use of automatic text analyser in the preparation of SDI profiles, J.Amer.Soc.Inf.Sci.,24,277-81,1973

16. D.Child (1970), The Essentials of Factor Analysis, Holt Rinehart and Winston,London,1970

17. P.Clague (1971), INSPEC SDI investigation, OSTI Report 5102, 1971

18. F.R.Clarke,T.G.Birdsall and W.P.Tanner (1959), Two types of ROC curves and definitions of parameters, J.Amer. Acoust.Soc.,31,629-30,1959

19. P.F.Cole (1966), Letter to editor,Aslib Proc.,18,325,1966

20. D.Colquhoun (1971), Lectures on Biostatistics,Clarendon Press,Oxford,1971

21. K.Cook (1975), A threshold model of relevance decisions, Inf. Processing Management,11,125-35,1975

22. W.S.Cooper (1971), A definition of relevance for information retrieval,Inf.Stor.Retr.,7,19-37,1971

23. W.S.Cooper (1973), On selecting a measure of retrieval effectiveness,Part 1,J.Amer.Soc.Inf.Sci.,24, 87-100,1973
Part 2,J.Amer.Soc.Inf.Sci.,24,413-24,1973

24. C.A.Cuadra (1964), On the utility of the relevance concept, Technical Report SP-1595,Systems Development Corp.,Santa Monica,California,1964

25. C.A.Cuadra and R.V.Katter (1967),Experimental studies of relevance judgements,TM-3520/001,002,003/00, 2 Vols,Systems Development Corp.,Santa Monica, California,1967

26. D.Davidson (1974), An examination of the effects of individual cognitive styles on the judgement of document relevance,PhD Thesis,Syracuse University,1974

27. L.B.Doyle (1963),Is relevance an adequate measure retrieval system evaluation? Proc.ADI,Part 2,199-200,1963

28. E.Dym (1967), Relevance predictability(1), In A.Kent et. al. (eds),Electronic Handling of Information, Testing and Evaluation, Thompson,Washington 1967

29. J.P.Egan and F.R.Clarke (1966), Psychophysics and signal detection,In J.B.Sidowski(ed),Experimental Methods and Instrumentation in Psychology,McGraw-Hill, New York,1966

30. D.Elwen (1972), Criteria for relevance decisions,MSc Thesis, City University,1972

31. R.A.Fairthorne (1963), Implications of test procedures, In Information Retrieval in Action,Western Reserve University Press,Cleveland,Ohio,1963

32. R.A.Fairthorne (1964), Basic parameters of retrieval tests, Proc.ADI,1,343-45,1964

33. D.J.Foskett (1972), A note on the concept of relevance,Inf. Stor.Retr.,8,77-78,1972

34. C.Friend (1970), A study of the interaction between enquirers and information staff in the answering of written enquiries,MSc Thesis,City University,1970

35. R.Fugmann (1973), On the role of subjectivity in establishing, using,operating and evaluating I.R. systems,Inf. Stor.Retr.,9,353-72,1973

36. W.D.Garvey,E.Tomita and P.Woolf (1974), The dynamic scientific information user,Inf.Stor.Retr.,10,115-31,1974

37. F.Gebhardt (1975), A simple probabilistic model for the relevance assessment of documents,Inf.Processing Management,11,59-65,1975

38. C.Gifford and G.J.Baumanis (1969), On understanding user choices: Textual correlates of relevance judgements, Amer.Doc.,20,21-26,1969

39. W.Goffman (1964), On relevance as a measure,Inf.Stor.Retr., 2,201-03,1964

40. W.Goffman (1968), An indirect method of information retrieval, Inf.Stor.Retr.,4,361-73,1968

41. W.Goffman and V.A.Newill (1966), A methodology for test and evaluation of I.R. systems,Inf.Stor.Retr.,3,19-25, 1966

42. D.M.Green and J.A.Swets (1966), Signal Detection Theory and Psychophysics,Wiley,New York,1966

43. J.G.Greeno and R.A.Bjork (1973), Mathematical learning theory and the new 'mental forestry',Ann.Rev.Psychol., 24,81-116,1973

44. J.P.Guilford (1954),Psychometric Methods,McGraw-Hill,New York,2nd Edition,1954

45. D.E.Gushee (1968), The reading behaviour of chemists,J.Chem. Doc.,8,191-94,1968

46. A.M.Hall,P.Clague and T.Aitchison (1972), The effects of the use of an SDI service on the information gathering habits of scientists and technologists, INSPEC Report R72/11,Institution of Electrical Engineers,London,1972

47. M.Hannah (1971), The basis of relevance decisions,MSc Thesis, City University,1971

48. C.W.Hanson (1964), Research on users needs:where is it getting us? Aslib Proc.,16,64-78,1964

49. H.H.Harman (1967), Modern Factor Analysis,Univ. Chicago Press,Chicago,1967

50. G.Harmon (1970), Information need transformation during enquiry,Amer.Soc.Inf.Sci.Proc.,7,41-43,1970

51. S.Herner (1962), Determination of user needs for the design of information systems,In Information Systems Workshop,Spartan Books,Washington,1962

52. S. and M.Herner (1967), Information needs and uses in science and technology,Ann.Rev.Inf.Sci.Technol.,2, 1-34,1967

53. O.D.Hibbert (1972), User needs in answering enquires,MSc Thesis,City University,1972

54. D.J.Hillman (1964), The notion of relevance,Amer.Doc.,15, 26-34,1964

55. J.D.Ingleby (1969), Decision making processes in human perception and memory,PhD Thesis,Cambridge University,1969

56. V.I.Ivankin (1975), A study of the semantic relationships between the text of a query and a relevant document,Nauch-Tekh.Inf.Series 2,21-26,1975

57. D.M.Jackson (1970), The construction of retrieval environments and pseudoclassifications based on external relevance,Inf.Stor.Retr.,6,187-219, 1970

58. G.Jahoda (1965), Information needs of science and technology: a background review,Proc.FID Conf.1965,Spartan Books,Washington,1965

59. R.V.Katter (1968), Influence of scale form on relevance judgements,Inf.Stor.Retr.,4,1-11,1968
60. J.Katzer (1972), Development of a semantic differential to assess users' attitudes towards an on-line interactive reference retrieval system,J.Amer. Soc.Inf.Sci.,23,122-27,1972
61. P.M.Kean (1973), The attitudes of scientists to information, MSc Thesis,City University,1973
62. D.A.Kemp (1974), Relevance, pertinence and information system development,Inf.Stor.Retr.,10,37-47,1974
63. A.Kent,J.Belzer,M.Kurfeerst,E.Dym,D.L.Shirey andA.Bose (1967), Relevance predictability in information retrieval systems,Methods Inf.Medicine,6,45-51,1967
64. D.W.King (1968), Design and evaluation of information systems,Ann.Rev.Inf.Sci.Technol.,3,61-103,1968
65. D.W.King and B.F.Bryant (1971), The Evaluation of Information Services and Products,Information Research Press,Washington,1971
66. D.W.King and V.E.Palmour (1974), User behaviour,In C. Fenichel (ed),Changing Patterns of Information Retrieval,Amer.Soc.Inf.Sci.,1974
67. M.Konigova (1970), Mathematical and statistical methods of noise evaluation in retrieval systems,Inf.Stor. Retr.,6,437-44,1970
68. D.Laming (1973), Mathematical Psychology,Academic Press, London,1973
69. D.Laming (1976), Personal Communication
70. F.W.Lancaster (1971), Aftermath of an evaluation,J.Doc., 27,1-10,1971
71. F.W.Lancaster and E.G.Fayen (1973), Information Retrieval On-line,Melville,Los Angeles,1973
72. F.W.Lancaster,R.L.Rapport and J.K.Penry (1972), Evaluating effectiveness of an on-line natural language retrieval system,Inf.Stor.Retr.,8,223-45,1972
73. P.Leggate (1971), Practical and theoretical problems of experimental design,OSTI Report 5111,1971
74. P.Leggate,B.Smith,J.Stow and M.I.Williams (1973a), The B.A.Previews project,OSTI Report 5139,1973

75. P.Leggate,B.N.Rossiter and J.B.F.Rowland (1973b), Evaluation of an SDI service based on the Index Chemicus Registry System,OSTI Report 5176,1973
76. P.Levy and D.Pugh (1969), Scaling and multivariate analyses in the study of organizational variables,Sociology,3,193-213,1969
77. M.Line (1974), Draft definitions,Aslib Proc.,26,87,1974
78. R.D.Luce (1963), Detection and recognition, In R.D.Luce and E.Galanter (eds),Handbook of Mathematical Psychology,Vol 1,Wiley,New York,1963
79. J.Marchlewska (1970), The information users and their categories,In Users of Documentation,F.I.D. Int.Congress on Documentation,F.I.D.,The Hague,1970
80. A.A.J.Marley (1971), Conditions for the representation of absolute judgement and pair comparison isosensitivity curves by cumulative distributions, J.Math.Psychol.,8,554-90,1971
81. M.E.Marons and J.L.Kuhns (1960), On relevance,probabalistic indexing and information retrieval,J.Assoc. Comp.Mach.,7,216-44,1960
82. M.W.Martin and R.L.Ackoff (1963), The dissemination and use of recorded scientific information,Management Science,9,322-36,1963
83. A.McAlpine,R.D.Whiteley and P.A.Frost (1972), The flow and use of scientific information in university research,OSTI Report 5138,1972
84. A.J.Meadows (1974), Communication in Science,Butterworths, London,1974
85. D.McNicol (1972), A Primer of Signal Detection Theory, George Allen and Unwin,Sydney,1972
86. F.Meister and D.J.Sullivan (1967), Evaluation of user reactions to a prototype on-line information retrieval system,Quoted in F.W.Lancaster and E.G.Fayen,Information Retrieval On-line, Melville,Los Angeles,1972

87. H.Menzel,L.Lieberman and J.Dulchin (1960), Reviews of studies in the flow of information among scientists, 2 Vols,Bureau of Applied Social Research, Columbia University,1960

88. N.Morikawa (1974), An experimental study of relevance judgements in literature searching processes, Abstract in Lib.Inf.Sci.Abs.,Absract No. 75/1839,1975 (Original in Japanese)

89. National Academy of Sciences (1959), Proc.Int.Conf. on Science Information,2 Vols,Nat.Acad.Sci., Washington,1959

90. National Science Foundation (1964), Summary of study conference on the evaluation of document searching systems and procedures,Nat.Sci. Found.,Washington,1964

91. Nan Lin and W.Garvey (1972), Information needs and users, Ann.Rev.Inf.Sci.Technol.,7,5-39,1972

92. N.H.Nie,C.H.Hull,J.G.Jenkins,K.Steinbrenner and D.H.Bent (1975), Statistical Package for the Social Sciences,2ndEd.,McGraw-Hill,New York,1975

93. J.O'Connor (1968), Some questions concerning information need,Amer.Doc.,19,200-03,1968

94. J.O'Connor (1969), Some independent agreements and resolved disagreements about answer providing documents, Amer.Doc.,20,311-19,1969

95. A.N.Oppenheim (1968), Questionnaire Design and Attitude Measurement,Heineman Educational Books, London,1968

96. R.H.Orr (1970), The scientist as an information processor, In C.E.Nelson and D.K.Pollock (eds), Communication Among Scientists and Engineers, Heath Lexington Books,Lexington,Mass.,1970

97. E.B.Parker (1966), The users place in an information system, Amer.Doc.,17,26-27,1966

98. E.B.Parker and W.Paisley (1966), Research for psychologists at the interface of the scientist and his information system,Amer.Psychol.,21,1061-71, 1966

99. K.Pawlik (1973), Right answers to the wrong questions? A re-examination of factor analytic personality research and its contribution to personality theory,In C.R.Royce (ed), Multivariate Analysis and Psychophysical Theory,Academic Press,London,1973

100. O.Persson (1974), Relevance judgements in different information sources,Abstract in Lib.Inf.Sci. Abs.,Abstract No. 75/1840,1975 (Original in Swedish)

101. G.J.Rath,A.Resnick and T.R.Savage (1961), Comparisons of four types of lexical indicat ors of content, Amer.Doc.,12,126-30,1961

102. A.M.Rees (1963), Information needs and patterns of usage, In Information Retrieval in Action,Western Reserve University Press,Cleveland,Ohio,1963

103. A.M.Rees (1965), The evaluation of retrieval systems, Comparative Systems Laboratory Report TR5, Western Reserve University,Cleveland,1965

104. A.M.Rees (1966), The relevance of relevance to testing and evaluation of document retrieval systems, Aslib Proc.,18,316-24,1966

105. A.M.Rees and T.Saracevic (1966), The measurability of relevance,Proc.Amer.Doc.Inst.,3,225-34,1966

106. A.M.Rees and D.G.Schultz (1967a), A field experimental approach to the study of relevance assessments in relation to document searching,2 Vols., Center for Documentation and Communication Research,School of Library Science,Western Reserve University,1967

107. A.M.Rees and D.G.Schultz (1967b), Psychology and information retrieval,In G.Schecter (ed),Information Retrieval: a Critical View,Thompson, Washington,1967

108. A.Resnick and T.R.Savage (1964), Consistency of human judgements of relevance,Amer.Doc.,15,93-95, 1964

109. S.E.Robertson (1975a), A theoretical model of the retrieval characteristics of information retrieval systems,PhD Thesis,University College,London, 1975

110. S.E.Robertson (1975b), Personal Communication
111. S.E.Robertson (1976), Personal Communication
112. D.G.Rowlands (1970), The Unilever research SDI system,Inf. Stor.Retr.,6,53-72,1970
113. G.Salton (1965), Definition of relevance and pertinence, Amer.Doc.,16,341,1965
114. T.Saracevic (1968), Effects of question analysis and search strategy on performance of retrieval systems, Comparative Systems Laboratory Report TR15, Western Reserve University,Cleveland,1968
115. T.Saracevic (1970a), The concept of relevance in information science,PhD Thesis,Western Reserve University, 1970
116. T.Saracevic (1970b), The concept of relevance in information science,In T.Saracevic (ed),Introduction to Information Science,Bowker,New York,1970
117. T.Saracevic (1975), Relevance:A review and a framework for the thinking on the notion in information science,J.Amer.Soc.Inf.Sci.,26,321-43,1975
118. M.I.Sastri (1968), A linguistic approach to relevance judgement,Methods Inf.Medicine,7,49-54,1968
119. D.L.Shirey and M.Kurfeerst (1967), Relevance predictability, In A.Kent et.al. (eds),Electronic Handling of Information: Testing and Evaluation,Thompson, Washington,1967
120. S.Siegel (1956), Non-Parametric Statistics for the Behavioural Sciences,McGraw-Hill,New York,1956
121. M.Slater and P.Fisher (1969), Use made of technical libraries, Aslib Occasional Publication No.2,Aslib,London, 1969
122. M.B.Snyder (1966), Methodology for test and evaluation of document retrieval systems. A critical review and recommendations,Report PB169572,Human Sciences Research Inc.,McLean,Virginia,1966
123. J.A.Swets (1963), Information retrieval systems,Science, 141,245-50,1963
124. J.Tague (1965), Matching of question and answer terminology in an education research file,Amer.Doc.,16, 26-32,1965

125. W.P.Tanner and J.A.Swets (1954), A decision making theory of visual detection,Psychol.Rev.,61,401-09, 1954

126. M.Taube (1965), Pseudo-mathematics of relevance,Amer.Doc., 16,69-72,1975

127. R.S.Taylor (1962), The process of asking questions,Amer. Doc.,13,391-96,1962

128. M.E.Thomas (1968), How scientists obtain their information, MSc Thesis,City University,1968

129. G.Weiler (1962), On relevance,Mind,71,487-93,1962

130. G.Weiler (1973), Relevance again,Inf.Stor.Retr.,9,121,1973

131. G.Wersig (1970), Communications theory and user analysis, In Users of Documentation,F.I.D. Int. Congress on Documentation,1970,F.I.D.,The Hague,1970

132. Westat Research Inc. (1968), Procedural guide for the evaluation of document retrieval systems, Report PB182711,1968

133. P.Wilson (1973), Situational relevance,Inf.Stor.Retr.,9, 457-71,1973

134. D.N.Wood (1967), Foreign language problems facing scientists and technologists in the U.K.,J.Doc.,23,117-30, 1967

135. D.N.Wood (1970), User studies. A review of the literature 1966-1970,Aslib Proc.,23,11-23,1970