CHAPTER 2

Test Environment

> Communication is the means which enables society to adjust itself to alterations of technology and education and other social changes. The scientific method can offer no grand vision, no global strategy, no panacea. It will never be possible to demonstrate that anything is absolutely right or even completely scientifically true.
>
> L.T. Wilkins: Social Deviance, page 28.

In the first volume were considered the general plan of the test design, the variables that were to be investigated and the methods to be used. In the course of the project, changes were made regarding certain details, and this chapter presents the environment in which the testing was actually done.

While an information retrieval system may be defined in its scope as 'all stages from the receipt of a document within a system, to the making of that document (or a representation of it) available to an enquirer', not all these stages have been included in the investigations in the present project. The central concern was the effect of index language devices on the operational performance, but in addition a number of other variables or factors have been included for various reasons. In order to clarify later discussions, a breakdown of an indexing system into four main groups is suggested, namely environmental factors, software factors, operational factors and hardware factors (see Fig. 2.1).

The environmental factors relate to the environment or conditions in which a given system has to operate. Four general factors are given, and, in the case of an operational system, they are all determined to a great extent by the needs of the user group which the system exists to serve. The subect field, the questions asked and the relevance needs directly depend on the users, while the collection size will be determined by the management largely in relation to user needs. However, for an

ENVIRONMENTAL FACTORS

SUBJECT FIELD; precision of terminology, overlap of terminology with other fields.
COLLECTION SIZE; Number of recognisable subject fields.
QUESTIONS ASKED; Broad survey, specific request, etc.
RELEVANCE NEEDS; graded decisions.

SOFTWARE FACTORS

CONCEPT INDEXING; level of exhaustivity.
INDEX LANGUAGE; hospitality for specificity and provision of devices.
SEARCH STRATEGY; flexibility to vary exhaustivity and specificity.

OPERATIONAL FACTORS

SUBJECT COVERAGE
TIME; indexing and searching.
EFFORT; intellectual and physical.
PERSONNEL; indexers and searchers, qualifications and performance.
CLERICAL ROUTINES
RETRIEVAL PERFORMANCE

HARDWARE FACTORS

TYPE OF STORE
INPUT
EXPANSION CAPACITY
UPDATING ABILITY
PHYSICAL FORM OF OUTPUT

FIGURE 2.1 MAIN FACTORS IN AN INFORMATION RETRIEVAL SYSTEM

experimental test a set of environmental conditions has to be created, and some of those are inevitably, to a greater or lesser extent artificial.

The software factors relate to the intellectual design of the storage and retrieval parts of an indexing system. The three main software factors are all the subject of management decisions in a given situation, and such decisions are always centred (although often unconsciously) on the twin parameters of exhaustivity and specificity (defined and discussed in Ref. 2).

The operational factors are concerned with the routine operation of a system, i.e. all the processes required to make documents available to enquirers when the system has been set up. The factors in Fig. 2.1 are not intended to be an exhaustive list, but are given to illustrate the range of operations involved. Any basic evaluation of such factors is complicated by an infinite number of possible compromises between the least effort and the best quality, with both effort and quality being subjective notions notoriously difficult to measure.

The hardware factors refer to the purely physical aspects of system operation that involve man-made entities. A brief and incomplete listing of five items is given.

If one considers Cranfield II within this framework, it can be seen basically to have investigated the software factors, in the context of a laboratory situation in which the environmental factors and operational factors have been strictly controlled. Hardware factors have been ignored because, in this investigation, the measurements are being made on those software factors which are quite unaffected by changes in hardware. The operational factor of retrieval performance is the main measurement made, and details of how this is done are given in Chapter 3. In the artificial environment created for the test it was found that a limited set of changes could be investigated; these included several sets of questions picked by different criteria, relevance judgements made in four different grades, collections of three different sizes and tests in two related but different subject fields.

Software factors

The software factors examined in the tests will be described and discussed first. A simplified table (Fig. 2.2) shows the variables that have been examined, listed under the three major factors of indexing, index languages and searching.

CONCEPT INDEXING

1. Manual indexing, at three levels of exhaustivity
2. Natural language abstracts and titles

INDEX LANGUAGES

1. Single terms
2. Simple concepts
3. Controlled terms
4. Abstracts and titles
5. Recall devices
   a. Single term indexing, eight languages
   b. Simple concept indexing, fifteen languages
   c. Controlled term indexing, six languages
   d. Abstracts and titles, four languages.
6. Precision devices
   a. Single term indexing, four types
   b. Simple concept indexing, one type
   c. Controlled term indexing, two types
   d. Abstracts and titles, one type

SEARCH RULES

1. Coordination levels, all possible levels
2. Combination rules, six types.

FIGURE 2.2 SOFTWARE FACTORS EXAMINED IN TEST

Concept-Indexing

The manual indexing carried out on the document collection is described in Chapter 4 of Volume 1, and this constituted the main body of data tested; of particular importance was the fact that three levels of exhaustivity of indexing were distinguished. The results of this variation in exhaustivity have been evaluated on the single term languages, but not on the simple concept or controlled term languages. In addition, Professor Salton prepared (with the SMART programme) a KWIC type index of the titles and abstracts of 200 documents (subset 1); in this connection abstracts and titles can be considered as variant forms of concept indexing, and the test searches which were made enabled direct comparison to be made with the manual indexing carried out by the project staff.

Data concerning the usage of terms in the single term language is given in Fig. 5.1 of Volume 1; some additional information on term usage is given in Fig. 2.3 in relation to the simple concept and controlled term languages, the average postings per document being 18 and 24 respectively. Fig. 2.4 gives similar data for the abstracts, with the average postings of key terms being 74. This latter figure is not strictly comparable, since the same word may be 'posted' several times for the same document.

SIMPLE CONCEPTS

Collection size	200	documents
Total terms in vocabulary	2,798	
Average posting per document	. 18	

CONTROLLED TERMS

Collection size	200	documents	350	documents
Total terms in vocabulary	816		985	
Terms in E.J.C. Thesaurus	694		827	
Additional terms	122		158	
Added lead-in vocabulary terms	1,285		1,514	
Average postings per document	24		24	

FIGURE 2.3 DATA CONCERNING USAGE OF TERMS IN SIMPLE CONCEPT AND CONTROLLED TERM INDEX LANGUAGES

COLLECTION SIZE	200	abstracts
Total postings of all words	33,042	
Total postings of words less those on restriction list	14,783	
Distinct words on restriction list	204	
Distinct words not on restriction list	3,123	
Average postings of all words per document	165	
Average postings of words not on restriction list per document	74	

First ten terms ranked by usage	FLOW
	NUMBER
	MACH
	PRESSURE
	RESULTS
	WING
	EFFECTS
	SHOCK
	BOUNDARY
	LAYER

FIGURE 2.4 DATA CONCERNING USAGE OF WORDS IN ABSTRACTS

Index languages

As described in Vol. I, Chapter 5, the languages tested fall into three main groups:

I Single Terms, with the base being the natural language concept indexing split into unit terms,
II Simple Concepts, with the base also being the natural language concept indexing, with some of the more complex pre-coordinated concepts split into simple concepts,
III Controlled Terms, with the base being the controlled vocabulary derived from the E.J.C. Thesaurus, and indexing performed by translating the natural language concepts into the controlled vocabulary.

In defining any particular index language, these three main types will be denoted by the Roman numerals I, II and III; the various sets of recall devices tested are denoted by Arabic numerals and the precision devices by lower case letters.

Recall devices

The starting point of each series of tests is the use of the basic terms as indexed. From this base, various recall and precision devices are added, both separately and in different aggregates. In the single term languages, four different recall devices were tested, namely control of synonyms, confounding of word forms, control of quasi-synonyms and control of clusters of terms by means of reduced vocabularies based on hierarchies. A total of eight aggregates was tested, and a tree diagram giving details of the eight languages is given in Fig. 2.5.

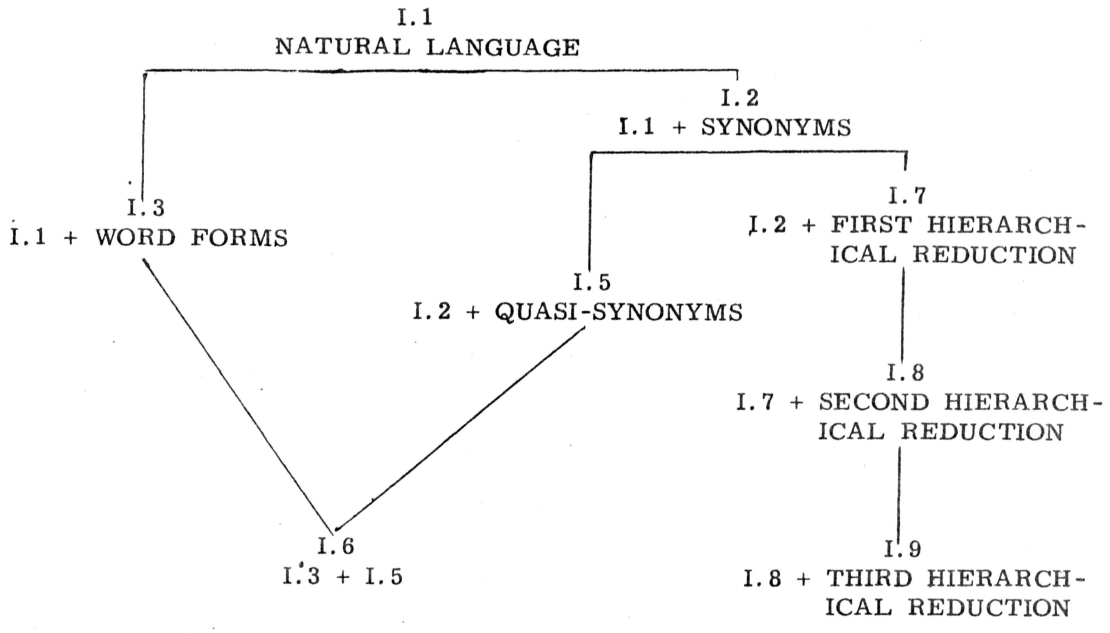

FIGURE 2.5 SINGLE TERM INDEX LANGUAGES

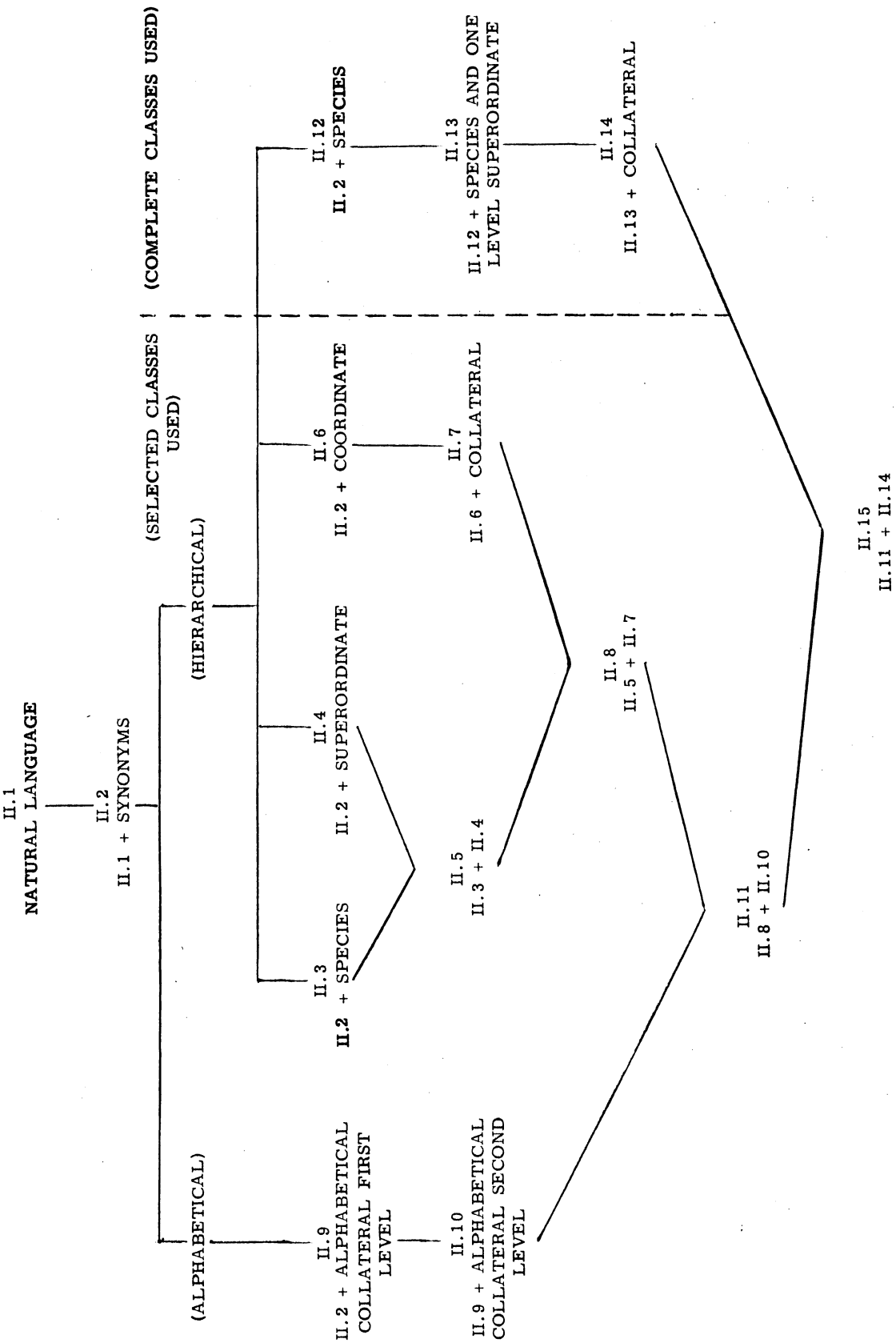

FIGURE 2.6 SIMPLE CONCEPT INDEX LANGUAGES

From this it can be seen that quasi-synonyms were tested together with synonyms, and that synonym control was also the base from which the three levels of reduction by hierarchy were tested.

The recall devices tested with the series of simple concept languages were the most comprehensive investigated. They involved one alphabetical and seven hierarchical devices, in fifteen different aggregates as shown in Fig. 2.6 (discussion on the hierarchies used and the rotated alphabetical list of concepts was given in Vol. 1, pages 74-83). It should be noted that recall devices 12, 13, and 14 of Fig. 2.6 involved the use of the complete classes of terms in the various hierarchical reductions, but, with the other languages, selections, based on intellectual decisions, were made from the various classes.

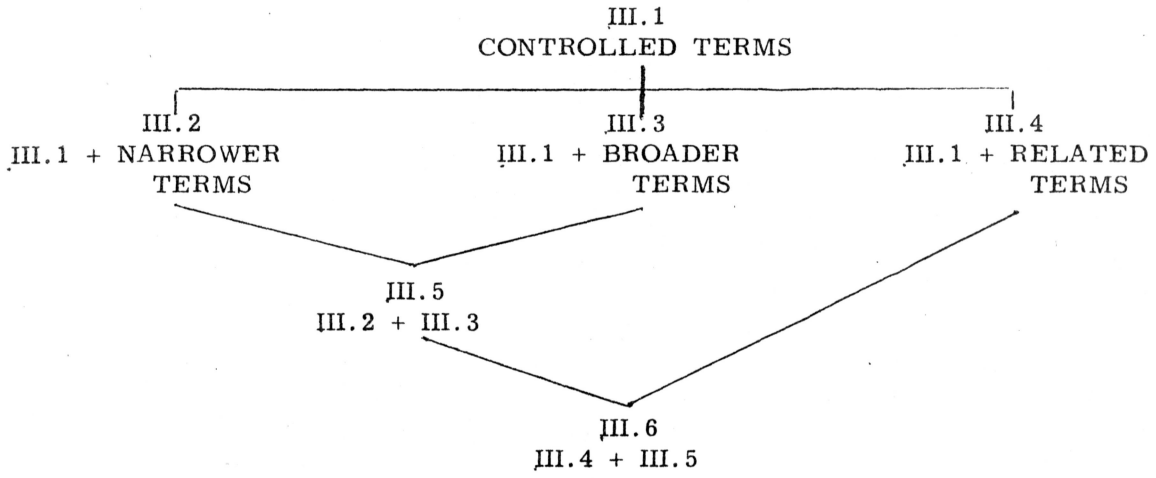

FIGURE 2.7 CONTROLLED TERM INDEX LANGUAGES

With the controlled terms, six index languages were tested. These consisted first of the basic terms, followed by the three classes of related terms as used in the E.J.C. Thesaurus (i.e. broader terms, narrower terms and related terms). In addition, two aggregates were tested; the six languages are listed in Fig. 2.7.

Precision devices

All the languages mentioned were tested for recall without any precision devices; this involved searches which accepted any one single term in the question. The fundamental precision device of coordination was also investigated in every test made, and all the basic tables of results in Chapter 4 show the coordination level in the rows of the tables. Two

additional precision devices were tested on the single term languages, namely partitioning and interfixing, as shown in Fig. 2.8.

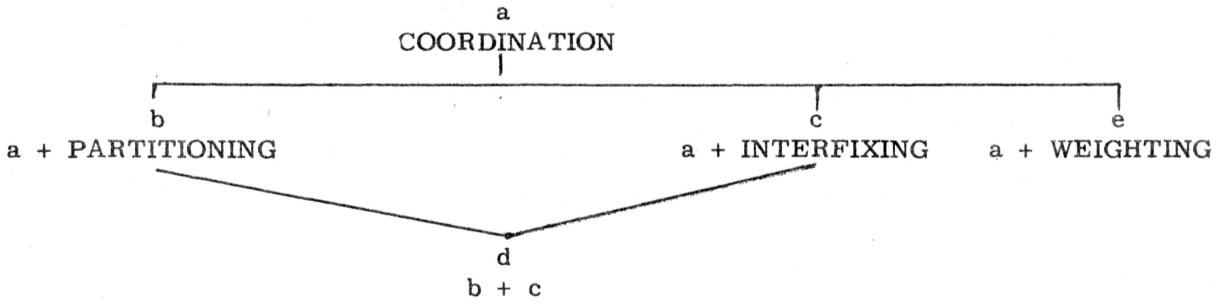

FIGURE 2.8. PRECISION DEVICES

No precision devices other than coordination were tested on the simple concept languages. The device of weighting was tested on the controlled terms. In this weights are assigned to the search term and a match sought with the weights assigned to the terms in indexing.

All the index languages tested may now be specified; for example II.2.a represents Simple Concept Index Language (II), with the recall device of Synonyms controlled (2), and coordination (a) as the precision device. The code for Single Term Index Language, with the recall device of Quasi-synonyms and the precision devices of partitioning and interfixing would be I.5.d.

Search Rules

In the search programmes for the questions tested an exhaustive extraction of all the possible notions contained in each question was made in the natural language of the questions as they were received. All these notions were included in the search prescription initially prepared for the three main index languages. After the basic question terms had been recorded, all the additional terms included in a logical sum relationship were pre-formulated by the very structure of the various languages already described. For example in Question 61 'Are there any papers dealing with acoustic wave propagation in reacting gases'. The terms underlined made up the search prescription, and these terms, as they are, were used for Index Language I.1. For Index Language I.2, Synonyms controlled, reference to Appendix 5.2 of Vol. I shows that the term Sound is now combined with Acoustic. For Index Language I.3 Word endings, the term Acoustically is combined with Acoustic; Waviness and Wavy are combined with Wave and there

are similar groupings for the other terms. For Index Language I.5, Quasi-synonyms, the term Sonic is combined with Acoustic, and Reaction now forms a group which includes the quasi-synonyms Energy, Force, Action, Behaviour, Kinetic, Response. With Index Language I.7, I.8 and I.9, the groups for each starting term are determined by the decisions taken in the compilation of the single term hierarchies as given in Appendix 5.3 of Volume I. There is nothing to add regarding the search prescription, for it was the search rules that were capable of variation; this could be achieved by varying the coordination level or by selecting acceptable combinations of the search terms.

As has been noted, all possible levels of coordination (logical product) were investigated at every stage, and therefore the effect of any rules that might be postulated concerning a minimum coordination level that would be acceptable can be seen from the tables of results. For instance, if a question had six terms, then the results would have been recorded for a search made with all six terms, then for a search with five terms, then with four terms and so on down to a single term search. No test was made in which the searches of a set of questions either commenced or were terminated by a subjective decision that varied from question to question.

The main variations introduced as search rules concerned the combinations of terms that were accepted. The six variations tested are given in Fig. 2.9.

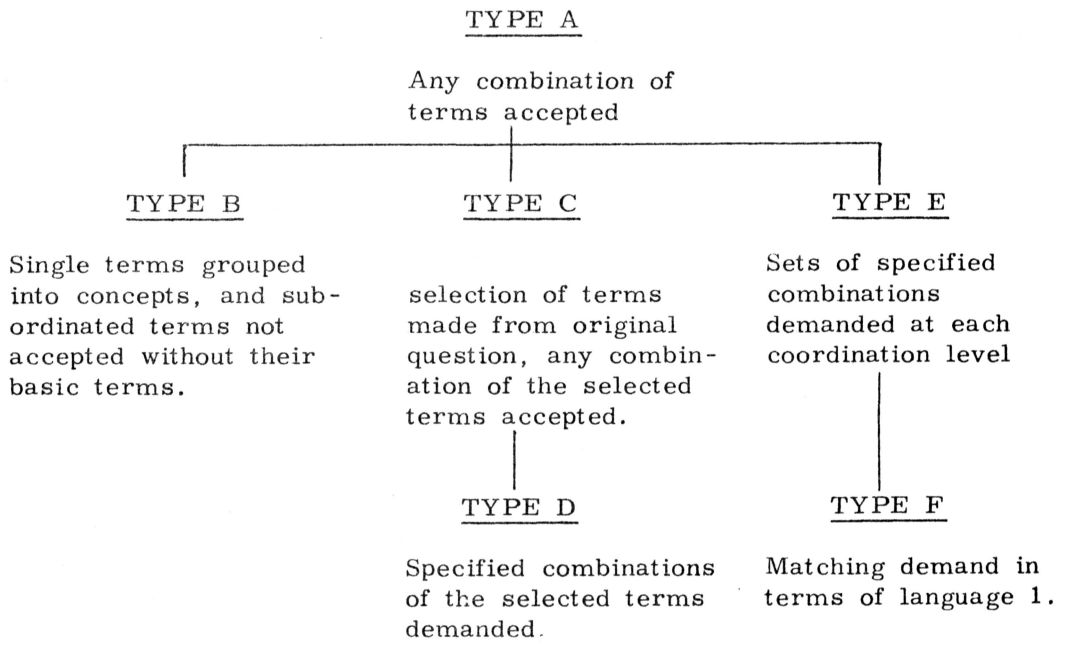

FIGURE 2.9 SUMMARY OF SEARCH RULES

In search Type A, any combination of terms was always accepted, without regard for the cases where some combinations accepted might be meaningless. For example, consider a question with the search terms Methods, Testing, Analysis, Investigating, Static, Dynamic, Stability, Characteristics, Re-entry, Body, Free, Flight, Tests. At, say, a coordination level of four, any combination of these search terms would be accepted, such as Methods, Static, Re-entry, Free. This is only one of many non-sensical combinations of the search terms at this level of coordination. The use of this search rule for investigating nearly every other variable was adopted, since it could be applied with equal consistency to all the different languages, with the exception of the tests of the precision devices of partitioning and inter-fixing on the single term languages. For these tests it was felt that a certain amount of intellect should be put into the search rules, and this consisted of a rule (Type B) which did not permit 'subordinate' terms to be accepted unless the associated 'basic' terms was present. The distinction between basic and subordinate terms became apparent when the single terms of the search questions were grouped into concepts, prior to the test of inter-fixing. In the example mentioned there are certain concepts that would emerge, such as Static stability characteristics, Re-entry body, Methods of testing. Basic terms in these concepts might be Stability, Body and Testing, for these terms are meaningful on their own in the context of the search question. Therefore Search Rule B would require, for instance, that Re-entry would not be accepted unless Body was also present, nor would Static be accepted unless Stability was present. The importance of adopting this rule before making a test of interfixing is that at least two terms from a concept must be present for interfixing to be tested. If, in the indexing of a document, the two single and separate terms Static and Stability appeared, and the demand for inter-fixing was added, if they were not interfixed then only one of the single terms could be accepted (which would have to be Stability to accord with Search Rule B). Without Search Rule B the single term Static could be accepted in this case, or in a case where Stability did not occur at all.

Searches C and D were carried out on the single term index languages, and represented an attempt to discover the effect of including more intelligence in searching. The first stage, Search C, involved making a selection of the original starting terms taken from the search question. This was to eliminate from the search prescriptions certain terms such as Problem, Applied, Variation, Influence, Solution, Comparison, Determination, Effect, etc. This search rule was tested on a set of twenty questions, all of which originally had seven starting terms; the selections made resulted in a range of from two to six of the terms, with the average being 4.1. In using these selected sets of search terms, any combination of these was still accepted, as in Search A.

Search D used the selected search terms of search C, and made strict

restrictions concerning the actual combination of terms that would be accepted at every coordination level, so as to eliminate the non-sensical combinations.

The most satisfactory and carefully applied search rules were applied to the controlled language tests, since it was thought that intelligence in searching would be best tested on an index language that also had an average degree of intelligence used in its formulation. This was Search E, where all the combinations of acceptable terms were individually selected for each coordination level. It was usual to accept a number of such combinations, with the object of retaining as many of the relevant documents as possible. This search rule was applied to the controlled term index languages (III.1 - III.6) both with and without the precision device of weighting. The sets of acceptable combinations were formulated on the basis of the starting terms of the question, and thus the use of Search E in testing languages other than III.1 (Basic terms) may have resulted in a poorer performance for the languages than is theoretically possible; the reason for this is that the grouping of a number of terms in the later languages might result in non-sensical combinations of terms.

One further additional rule designed to be used with the various recall languages was tried. This was Search Type F, also carried out on the controlled term index languages III.2 to III.6. The reasoning behind this search was that in all previous rules tested, the terms that actually made a match between a document and search prescription were all treated 'equally'. For example, if two documents had a match of five terms with a question using the controlled term index language III.5a (related terms), no distinction would be made between a document which actually had four starting terms, and only one related term, and a second document which was matched only by related terms, without a single starting term. The first document clearly represents a closer match with the search prescription, and it might generally be assumed that a starting term match is more desirable than any related term match. In Search F, a record was made of the number of starting terms that came up in a given match, and was done with the rules of Search E in use. This was used to make up sets of results with a given minimum match demanded, and results will be given for controlled term languages III.5 and III.6.

Document relevance

Before demonstrating the form of the results obtained when these variables are tested, a single environmental variable will be mentioned. This is the variation made in document relevance, resulting from the scale of four grades of relevance that was followed by the questioners in assessing the relevant documents (see Vol I, p.21). In finding the effect on retrieval performance of these decisions, four sets of results

were obtained, comparing first a set of questions when only relevance 1 documents were accepted as relevant, then with documents of relevance 1 or 2, next with documents of relevance 1 or 2 or 3, and finally with documents of relevance 1 or 2 or 3 or 4. Apart from the particular test to measure this variable, the broadest relevance decision, namely 1 - 4, was always used in other tests.

The Composite Table

Some idea of the volume, variety and complexity of the tests carried out can be seen from the composite table, (Fig. 2.10) which gives results for various combinations of six variables tested on the single term index languages I.1 to I.6. The basic set of questions used is subset 1, which has 35 questions, each having seven starting terms, but some of the results are based on two selections of these, namely 19 questions of subset 4 and 20 questions of subset 6. Four of the variables are listed at the head of the table, and the other two at the left side; the table divisions consist of the following factors:-

1. The coordination level varies from 1 to 7, which would result in seven main sections of the table. However, due to problems of presentation in this report, the table is truncated by the omission of the figures relating to the first three levels, so that it only presents four main sections covering the coordination levels of 4, 5, 6 and 7.
2. Four search rules (A,B, C and D) are next varied, and are applied in order of increasing intelligence within each coordination level.
3. The precision devices (a, b, c and d) are recorded next, with most results using no linking devices, apart from the three columns near the centre of each section.
4. The final factor at the head of the table is document relevance, with the three higher grades listed first, followed by the lowest grade used for all subsequent combinations (1, 1-2, 1-3, and 1-4).
5. The rows are first divided into five, representing the index languages I.1, I.2, I.3, I.5 and I.6.
6. The final variable is indexing exhaustivity, the three levels being repeated as divisions of each index language in turn.

The meaning of the codes used in this table has already been described earlier in this chapter.

The search results are shown as percentages for recall and precision.

Thus each set of recall and precision devices can be understood by examining the columns above, and the row to the left of a set of ratios, and then reading off the particular combination of variables being tested. For example, if the first section of the table as printed is examined

Co-ordination		4+													
Search Rules		A				B								C	D
Precision Device		a				b		c		d		a			
Document Relevance		1		1-2		1-3		1-4							
Recall Device	Exhaustivity	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
I.1	1	28	2	25	9	17	17	19	24												
	2	44	1	35	5	28	10	30	15												
	3	44	1	38	4	33	10	33	14	28	23	20	26	19	32	12	31	28	29	21	59
I.2	1	28	2	25	8	17	16	19	23												
	2	50	1	37	5	30	10	31	15												
	3	50	1	39	4	34	10	35	13	29	21	21	23	19	32	12	31	29	30		
I.3	1	39	3	30	9	20	16	21	23												
	2	56	1	41	5	33	10	33	13												
	3	56	1	43	4	37	8	36	11	33	15	24	24	22	24	15	29	33	24		
I.5	1	39	1	27	4	23	9	25	14												
	2	56	1	41	2	36	5	38	7												
	3	56	1	44	2	42	4	44	6	40	8	27	12	26	11	18	15	36	14		
I.6	1	44	1	32	4	24	8	26	12												
	2	56	1	42	2	37	4	40	6												
	3	56	1	47	1	44	4	45	5	44	7	30	11	29	11	21	15	40	11	27	54

FIGURE 2.10a THE COMPOSITE TABLE. COORDINATION LEVEL 4+
R = RECALL RATIO, P = PRECISION RATIO
(Performance figures are expressed as percentages)

Co-ordination		5+									
Search Rules		A				B				C	D
Precision Device		a				b	c	d		a	
Document Relevance		1	1-2	1-3	1 - 4						
Recall Device	Exhaus-tivity	R P	R P	R P	R P	R P	R P	R P	R P	R P	R P
I.1	1	11 5	9 17	7 37	8 54						
	2	28 4	19 11	14 22	15 31						
	3	28 3	20 9	16 18	16 26	12 64	7 64	6 100	5 100	16 47	16 64
I.2	1	17 7	9 16	7 35	8 51						
	2	33 4	22 12	15 22	16 32						
	3	33 3	23 9	17 18	18 25	13 65	7 64	6 100	5 100	19 47	
I.3	1	17 6	10 17	8 36	8 51						
	2	39 4	23 11	17 22	17 29						
	3	39 3	25 8	19 17	19 23	16 51	9 55	7 82	5 78	19 38	
I.5	1	11 2	11 9	8 16	10 27						
	2	39 2	25 5	17 10	19 15						
	3	44 2	30 5	21 8	23 12	21 23	11 48	11 38	7 50	19 33	
I.6	1	22 4	15 11	10 21	11 30						
	2	44 2	27 5	19 9	21 13						
	3	50 1	30 4	24 8	25 11	22 21	13 36	11 36	7 45	28 27	23 63

FIGURE 2.10b. THE COMPOSITE TABLE. COORDINATION LEVEL 5+

Coordination		6+									
Search Rules		A				B				C	D
Precision Device		a				b	c	d		a	
Document Relevance		1	1-2	1-3	1 - 4						
Recall Device	Exhaus-tivity	R P	R P	R P	R P	R P	R P	R P	R P	R P	R P
I.1	1	11 17	5 33	4 75	4 83						
	2	11 5	11 21	8 40	8 44						
	3	11 3	11 15	8 28	8 38	8 100	5 100	3 100	3 100	5 33	5 33
I.2	1	11 17	5 33	4 75	4 83						
	2	11 4	11 20	8 37	8 50						
	3	11 3	11 15	8 27	8 37	8 100	5 100	3 100	3 100	5 33	
I.3	1	11 17	5 33	4 75	4 83						
	2	17 6	13 21	9 40	9 52						
	3	17 4	13 14	9 28	9 36	8 100	5 100	3 100	3 100	5 33	
I.5	1	11 13	5 25	5 63	4 75						
	2	28 6	19 18	12 30	12 43						
	3	28 4	19 12	12 28	12 28	12 44	5 58	3 44	3 44	10 50	
I.6	1	11 12	5 24	5 59	4 71						
	2	33 8	19 15	12 26	13 37						
	3	33 4	22 12	13 19	13 26	13 37	6 62	3 44	3 44	10 50	5 33

FIGURE 2.10c. THE COMPOSITE TABLE. COORDINATION LEVEL 6+

Co-ordination		7+							
Search Rules		A				B			
Precision Device		a				b		c	d
Document Relevance		1	1-2	1-3	1 - 4				
Recall Device	Exhaustivity	R P	R P	R P	R P	R P	R P	R P	R P
I.1	1	11 29	4 43	2 71	2 71				
	2	11 13	4 19	3 38	3 50				
	3	11 13	4 19	3 38	3 50	3 100	3 100	2 100	2 100
I.2	1	11 29	4 43	2 71	2 71				
	2	11 13	4 19	3 38	3 50				
	3	11 13	4 19	3 38	3 50	3 100	3 100	2 100	2 100
I.3	1	11 29	4 43	2 71	2 71				
	2	11 13	4 19	3 38	3 50				
	3	11 13	4 19	3 38	3 50	3 100	3 100	2 100	2 100
I.5	1	11 25	4 38	3 75	2 75				
	2	11 9	4 14	3 32	4 45				
	3	11 7	5 14	4 28	4 38	5 100	3 100	2 100	2 100
I.6	1	11 25	4 38	3 75	2 75				
	2	17 12	6 20	4 36	4 48				
	3	17 9	8 19	5 31	5 41	6 53	4 100	2 100	2 100

FIGURE 2.10d. THE COMPOSITE TABLE. COORDINATION LEVEL 7+

(coordination level 4), and the ratios at the top left corner examined (28% recall, 2% precision), the following variables are shown to have produced that result: a search at coordination level of four terms; search rule A (any combination); precision device 'a' (no linking in the index language); relevant documents graded 1 only accepted; recall language 1 (natural language terms); and indexing exhaustivity 1 (low exhaustivity). After this, a move across this section of the table to the right will first alter the document relevance grades, then introduce a search rule, then include the three precision devices and finally test three more search rules. A move into the next section will increase the coordination level of the search, and in any section a move down the table will increase the indexing exhaustivity before a new recall language is brought in.

The position of these variables in the table is of no significance; the table could, for instance, first have been divided into the five recall languages, with the seven coordination levels repeated at each stage, etc. and hundreds of variations are possible. The actual combinations of different variables for which results have been presented in the complete composite table total 609, which is a choice of the most useful combinations out of the theoretical total of 6720 combinations possible.

Each set of recall and precision ratios is an average of results from the set of 35 questions and it is estimated that the composite table represents more than 16,000 individual results. When it is considered that the scope of the whole project extends to 221 questions, that there are some 28 other index languages which are not included in this table and that there are a number of other new variables, the individual results available are estimated to exceed 200,000.

Environmental Factors

The main environmental factors involved in the testing are listed in Fig. 2.11. For various reasons, as the test proceeded, different sets of questions and collections of different sizes were used. To consider first the sets of questions. Although 279 questions were available for use, the largest set for which results are presented numbers 221. The balance of 58 were multi-themed questions, that is they really consisted of more than one question, e.g. Question 3 'How can one describe the aerodynamic forces and the heating rates acting on high speed aircraft'. Four of these were used in some of the smaller question sets only. The first series of tests, on the recall devices of the single-term index languages, were made of the complete collection of 221 single-theme questions. The major problem that then arose was to find a satisfactory method of totalling the results of searches based on different numbers of starting terms (this matter is considered at length in Chapter 3). For this reason, we investigated the results on a set of 35 questions each of which had seven starting terms. The tests on interfixing and partitioning were particularly difficult to do, because of the painstaking clerical work necessary. These were therefore done on

two subsets which had 19 questions with 7 starting terms and 17 questions with 11 starting terms.

QUESTIONS

1. Relevance assessments, 4 grades
2. Differing number of starting terms and retrieving terms
3. Differing totals of relevant documents
4. Two sources of questions, 'basic' and 'supplementary'
5. Question sets of different sizes, picked according to different criteria, searched on collections of varying sizes.

COLLECTION SIZE

1. 1400 documents
2. 350 documents from the 1400 documents.
3. 200 documents from the 350 document subset.

SUBJECT TERMINOLOGY

1. Aerodynamics
2. Aircraft Structures.

FIGURE 2.11 SUMMARY OF MAIN ENVIRONMENTAL FACTORS

By the time we came to investigate the simple concept languages and the controlled term languages, the clerical effort involved in carrying out searches precluded the use of the full sets of questions, and accordingly a set of 42 questions was prepared, consisting entirely of questions in the field of aerodynamics. It is this set which is used for presenting the majority of the test results in Chapter 4. At a later stage, this subset was extended to 77 questions in the field of aerodynamics; finally an additional set of 42 questions in the field of structures was compiled for purposes of comparison, with the aerodynamic question set of similar sizes. The subsets of questions are all numbered, and details of these appear in Fig. 2.12. Lists of the question numbers for subsets 1, 2 and 3 were given in Vol. I, Appendix 3E; the remaining subsets are shown in Appendix 3.2 of this volume.

Reduced collection sizes were also used for reasons of the effort involved in testing. This was not only the clerical effort involved in the searching, but also the intellectual effort involved in compiling word lists for the various index languages. When it was decided to test simple concepts, a set of 200 documents was chosen, and the initial task involved re-formulating the indexed concepts from the original

Question Subsets	No. of Questions	No. of Documents in Collection Tested	No. of Relevant Documents	Generality Number
Subset 1 All have seven starting terms in single term languages, covering aerodynamics and structures.	35	1400	287	5.9
Subset 2 Starting terms vary, all questions aerodynamics only. 38 are drawn from Subset 3, and 4 from the 58 questions not used. The number of relevant in the 1400 collection is actually 201, but the three documents concerned (1329 in Q119, 2289 in Q145 and Q146) are deleted from the collection in results of this subset since they did not appear in collection Subset 1.	42	1400	198	3.4
	42	200 (subset 1)	198	23.6
	42	350 (subset 2)	198	13.5
Subset 3 Starting terms vary, covering aerodynamics and structures. The largest set of questions available, all single theme in single term languages.	221	1400	1590	5.1
Subset 4 Part of Subset 1, all having seven starting terms.	19	1400	131	4.9
Subset 5 All having eleven starting terms in single term languages, covering aerodynamics and structures.	17	1400	109	4.6
Subset 6 Part of Subset 1, all having seven starting terms.	20	1400	147	5.3
Subset 7 Includes all the questions in Subset 2, aerodynamics only.	77	350 (subset 2)	454	16.8
Subset 8 Structures only.	42	1400	255	4.3

FIGURE 2.12

indexing records. The choice of the subset of 200 documents extracted from the 1400 was governed by:-

1. Use of a set of aerodynamic questions and documents.
2. Choice of the largest set of questions that could be tested on a subset of 200 documents.
3. Choice of questions restricted to those not having any relevant documents in the range of Numbers 1001 - 1299 (because of the different weighting method used at that stage of the indexing).

The third rule restricted the choice of questions quite considerably, and the second rule was modified by not allowing 'similar' questions - mainly two or more questions having an overlapping set of identical relevant documents asked by the same questioner. The 42 questions finally used had 198 relevant documents in the subset of 200 documents. None of the base documents to the 42 questions were included.

The second subset chosen was one of 350 documents (subset 2), and this subset included the 200 documents of subset 1. Subset 2 also consisted entirely of aerodynamic documents, with an additional 35 questions having all their relevant documents in the subset. These, together with the 42 questions for subset 1, resulted in the 77-question subset (subset 7) which was used for the tests on the controlled vocabulary.

In presenting the test results, the majority of results are based on these smaller subsets of documents and questions. The first tests were made with 221 questions on the 1400 collection, and these tests were repeated on smaller documents and question subsets in order to validate the use of such subsets. It will be shown in the next chapter how the difference in performance can be adequately accounted for, and the use of smaller subsets does not, we believe, impare either the validity or, to any appreciable extent, the accuracy of the results and findings.

The use of these subsets enabled the various environmental factors involved to be investigated. For example, the effect of the change in collection size from 1400 to 200 documents with a fixed set of questions was investigated. Comparison was also possible between the 350 and 200 collections.

In the case of the questions, different subsets were made up and results obtained when environmental factors such as those listed in Fig. 2.11 were being investigated. The four grades of document relevance are included in the main test results (Chapter 4) and the effect of the other factors that are listed is considered in Chapter 6.

An attempt was made to compare the two distinct subject fields that existed in the 1400 document collection. Many of the question sets contained both aerodynamics and structures questions, but the direct

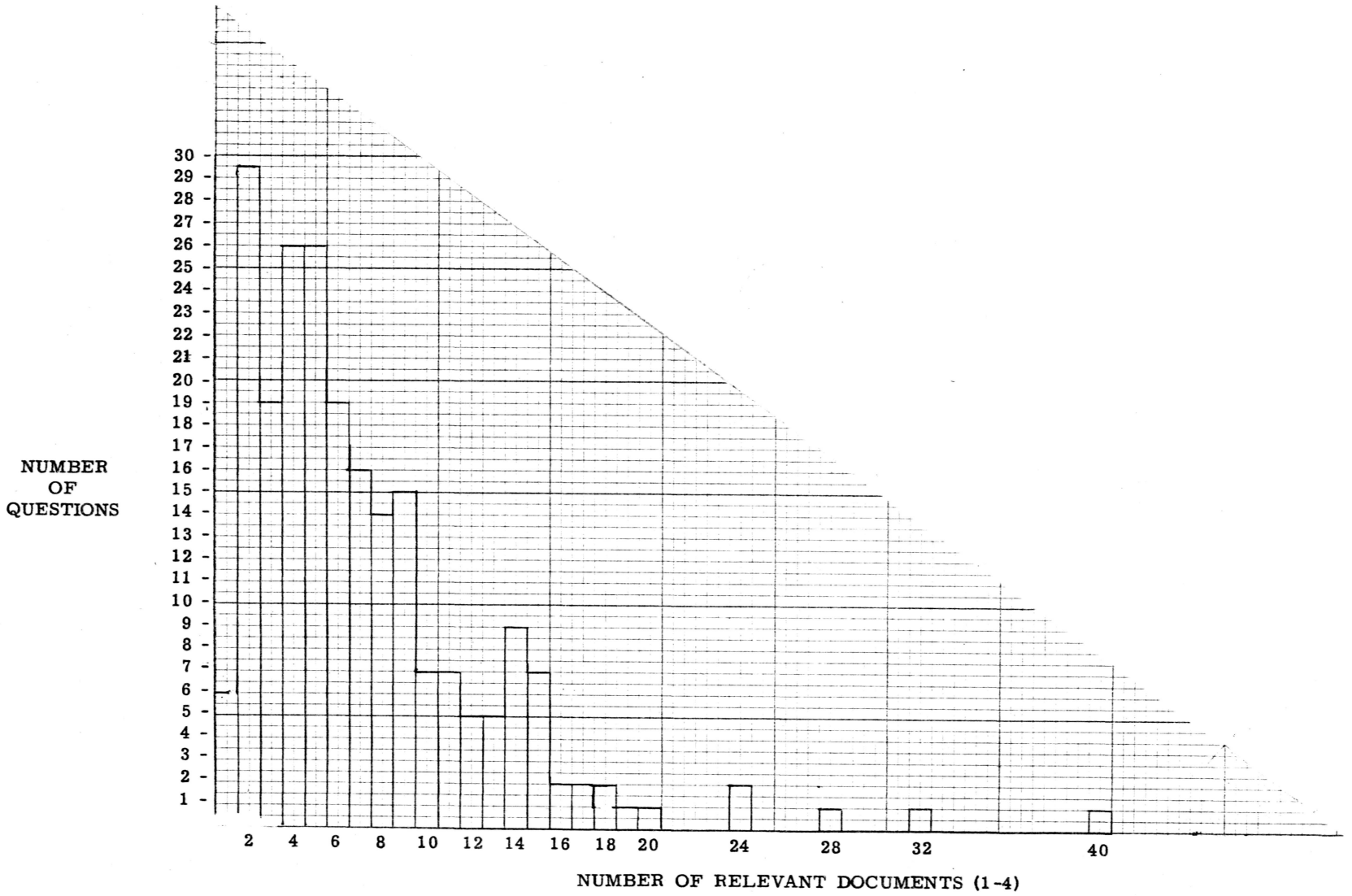

FIGURE 2.13 DISTRIBUTION OF THE RELEVANT DOCUMENTS IN THE 221 QUESTIONS

comparison was made of 42 questions on aerodynamics and 42 questions on structures.

Fig. 2.13 is a chart showing the distribution of the number of relevant documents throughout the 221 questions, from which it can be seen that the range is from six questions each having only one relevant document to one question which had forty relevant documents.

Sample Precision Results

At low precision ratios, the clerical work involved in obtaining correct figures was so great that in some cases it did not appear to be justified. This was due to the large number of non-relevant documents which would be retrieved and therefore had to be recorded. With index language I.1a, results were obtained down to the single term level but with other index languages the decision was taken that, with the searches in the 1400-document collection, no attempt would be made to obtain precision figures below 5%. This, however, introduced a variation between questions, since for a question having six starting terms, a precision figure lower than 5% might not be reached until the coordination level was down to two terms. However, with a ten starting term question, this figure might be reached by the coordination of four terms.

In the presentation of the test results, note has been taken of this point and also the additional point regarding the number of questions capable of giving results, this being dependent on the number of starting terms which each question had. This can be best illustrated by referring to Fig. 2.14, which presents condensed results for 221 questions on the 1400-document collection with Index Language I.2a. The column headed 'z' presents the figures for the number of questions that were potentially capable of giving results, and it can be seen that at a coordination level of 2, every question came in this category. However, at a coordination level of 3, the total has dropped to 220, this indicating that there is one question which had only two starting terms. At a coordination level of 4, the total drops to 212, showing that there are eight questions with only three starting terms. As the coordination level rises, so the number of questions drops until, at a level of 15, it is seen that only three questions have this number of starting terms.

The column headed 'y' shows the number of questions which actually contributed figures for the calculation of the precision and fallout ratios - not, it should be noted, for the recall ratio which was always checked down to single term level. In Fig. 2.14, y is equal to z from a coordination level of15 down to a coordination level of 7, and therefore the precision and fallout ratios can be based on complete data. However, at a coordination level of 6, only 161 questions were searched, and the precision and fallout ratios have been calculated on the basis of the non-relevant documents retrieved in these 161 searches. To indicate this, an asterisk

Index Language I. 2. a (S. T. Synonyms. Coordination)

Exhaustivity of Indexing 3
Search Rule A
Document Relevance 1 - 4

Number of Documents in Collection 1,400
Number of Questions 221 (Subset 3)
Number of Relevant Documents 1,590
Generality Number 5.1

Coord-ination Level	Documents Retrieved Rel.	Non-rel.	Recall Ratio a/a+c	Precision Ratio a/a+b	Fallout Ratio b/b+d	x	y	z
1	1,514	(-)	95.2%	(-)	(-)	221	0	221
2	1,313	59,734*	82.6%	2.2%*	19.406%*	221	44*	221
3	981	23,654*	61.7%	4.0%*	7.680%*	216	109*	220
4	644	8,850*	40.5%	6.8%*	2.873%*	192	142*	212
5	355	2,946*	22.3%	10.4%*	0.957%*	139	177*	197
6	169	928*	10.6%	15.4%*	0.301%*	92	161*	164
7	80	254	5.0%	24.0%	0.083%	55	140	140
8	24	59	1.5%	28.9%	0.019%	23	105	105
9	8	8	0.5%	50.0%	0.003%	8	78	78
10	1	0	0.1%	100.0%	0.000%	1	52	52
11	0	0				0	32	32
12	0	0				0	15	15
13	0	0				0	8	8
14	0	0				0	4	4
15	0	0				0	3	3

FIGURE 2.14 SAMPLE TABLE OF TEST RESULTS

is always given against any figures which have been calculated on a reduced set. At a single term level, it can be seen that no searches were made, and therefore no figures can be estimated for precision or fallout ratios.

There were various possible procedures for estimating these figures, and these can be illustrated by reference to Fig. 2.15, which deals with the 35 questions subset searched on 1400 documents by Index Language I.5.a. Since all the questions had seven starting terms, z remains constant throughout. However, at a coordination level of 2, it is shown in column y that only 23 questions were searched. It was found that, with these 23 questions 8,565 non-relevant documents were retrieved together with 157 relevant documents. The simplest way of estimating the total non-relevant for the complete subset of 35 questions would be to scale up the above figure of 8,565 in the ratio of $\frac{35}{23}$, which would give a total of 13,033 non-relevant documents. On the basis of this figure the precision and fallout ratios* could now be calculated. A second method is first to determine the precision ratio for the 23 questions searched; in this case it works out at 1.8%. It is known that the 35 questions retrieved 253 relevant documents; to maintain the precision ratio of 1.8% the total of non-relevant is scaled up by $\frac{253}{157}$, namely the totals of relevant documents retrieved in the full set and in the subset. This gives a figure of 13,803 and from this the fallout ratio can be calculated.

The accuracy of these scaled up results will depend on whether the sample of questions that were searched is typical of the whole set. It is unlikely that this was the case; as stated earlier, questions were not searched when they would retrieve an excessive number of non-relevant documents, so conversely the questions which were searched, and which are therefore in the sample, were those which had fewer non-relevant documents. Scaling-up from the sample could therefore be expected to give a somewhat higher precision figure than was really the case.

To check on this, we can consider the actual situation in regard to the same set of questions with Index Language I.1.a, on which, as previously mentioned, searches were made down to the single-term level.

In this language, at a coordination level of 2, the 23 questions retrieved 3871 documents. By the methods already suggested, the estimated figures would have been 6043 and 6476 respectively. In fact,, the correct figure is 8086, and bears out the expectation expressed in the previous paragraph. This was also checked at the coordination level of 3, and again it was found that the remaining 12 searches retrieved

---

*The method of calculating these ratios is discussed in Chapter 3.

Index Language I. 5. a (S. T. Synonyms, Quasi-synonyms. Coordination)

Exhaustivity of Indexing 3
Search Rule A
Document Relevance 1 - 4

Number of Documents in Collection 1.400
Number of Questions 35 (subset 1)
Number of Relevant Documents 287
Generality Number 5.9

Coord-ination Level	Documents Retrieved Rel.	Non-rel.	Recall Ratio a/a+c	Precision Ratio a/a+b	Fallout Ratio b/b+d	x	y	z
1	280	(-)	97.6%	(-)	(-)	35	0	35
2	253	17,130*	88.2%	1.5%*	34.959%*	35	23*	35
3	194	7,472	67.6%	2.5%	15.339%	35	35	35
4	125	2,086	43.6%	5.6%	4.282%	34	35	35
5	65	463	22.7%	12.3%	0.950%	30	35	35
6	35	88	12.2%	28.4%	0.181%	16	35	35
7	11	18	3.9%	38.0%	0.037%	5	35	35

FIGURE 2.15 SAMPLE TABLE OF TEST RESULTS

approximately the same number of relevant documents as the original 23 searches. For this group of results, therefore, the figures at the coordination level of 2 have been estimated by doubling the total obtained for the 23 questions. Similar procedures have been used in other cases.

This can certainly be considered somewhat unsatisfactory, and it could be argued that it would have been preferable not to have attempted to obtain figures by such a dubious method. However, it is felt that they do have some value; in every case where any such action is taken, an asterisk is placed against the figure or the ratio, and, if the reader feels so inclined, these results can be ignored.

As can be seen from the example of Figs. 2.14 and 2.15, each table of results contains details of the environment in which the test was carried out. This includes the particular index language, the level of exhaustivity, the search rule, the level of relevance, the number of documents and questions, the number of relevant documents, and the generality number. The latter, and the meaning of x in the tables, is considered in Chapter 3.