CHAPTER 8

CONCLUSIONS

> The first step in testing a theory (qua theory) is to examine it to see what deductions can be made from it - to set up postulates which may be tested either experimentally or by observations of the 'real-life' situation. That is to say, the first step in testing a theory is to state the practical consequences of it. If the deduced practical consequences (operational definitions) are proved to be unsustained, the theory is discredited. No theory can ever be proved to be true; it is held for so long as no better theory can be found.
>
> L.T. Wilkins: Social Deviance Page 36

Although the results presented in this volume inevitably represent only a condensation of the tens of thousands of individual results which have been obtained, it is hoped that they are in sufficient detail for anyone interested to make their own interpretation. It might, therefore, be argued that much of this final chapter is redundant, and that it would be better to leave readers to reach their own conclusions. However, the following comments are offered as a personal contribution, with the hope - and expectation - that others will feel free to deduce and argue.

The results have been presented in three main ways. Firstly, there are the details of the search results for the various index languages, recall and precision devices and search rules as obtained with the conventional coordination level cut-off. Secondly, some of these results have been regrouped to illustrate various aspects of the test and thirdly, many of the test results have been re-calculated by the document output cut-off method based on simulated ranking. While the opinions presented in this chapter may be illustrated by referring to a particular set of figures, they are not usually based on a single result.

Within the definition as given in Chapter 2 of Volume 1, every set of figures supports the original hypothesis of an inverse relationship between recall and precision. It is immaterial which variable is changed to give a new system; it may be the coordination level (e.g. Fig. 4.100T), the exhaustivity of indexing (e.g. Fig.4.912P), the recall devices (e.g. Fig. 6.10T), the precision devices (e.g. Fig.6.17T), the search programmes (e.g. Fig.4.850T), or the relevance decisions (e.g. Fig. 6.3P); it has been impossible to find any exception to what can be claimed as a basic rule.

Quite the most astonishing and seemingly inexplicable conclusion that arises from the project is that the single term index languages are superior to any other type. This is mainly evidenced by the results based on the normalised recall ratios of Fig. 5.15T, but also, although less obviously, by the comparison of different systems using the conventional coordination level cut-off (see Fig. 6.2P). This conclusion is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used to obtain these results, and our own first reaction was to doubt the evidence. A complete recheck has failed to reveal any discrepancies, and unless one is prepared to say that the whole test conception is so much at fault that the results are completely distorted, then there is no other course except to attempt to explain the results which seem to offend against every canon on which we were trained as librarians.

ORDER	NORMALISED RECALL		INDEXING LANGUAGE
1	65.82	I-3	Single terms. Word forms
2	65.23	I-2	Single terms. Synonyms
3	65.00	I-1	Single terms. Natural Language
4	64.47	I-6	Single terms. Synonyms, word forms, quasi-synonyms
5	64.41	I-8	Single terms. Hierarchy second stage
6	64.05	I-7	Single terms. Hierarchy first stage
7=	63.05	I-5	Single terms. Synonyms. Quasi-synonyms
7=	63.05	II-11	Simple concepts. Hierarchical and alphabetical selection
9	62.88	II-10	Simple concepts. Alphabetical second stage selection
10=	61.76	III-1	Controlled terms. Basic terms
10=	61.76	III-2	Controlled terms. Narrower terms
12	61.17	I-9	Single terms. Hierarchy third stage
13	60.94	IV-3	Abstracts. Natural language
14	60.82	IV-4	Abstracts. Word forms
15	60.11	III-3	Controlled terms. Broader terms
16	59.76	IV-2	Titles. Word forms
17	59.70	III-4	Controlled terms. Related terms
18	59.58	III-5	Controlled terms. Narrower and broader terms
19	59.17	III-6	Controlled terms. Narrower, broader and related terms
20	58.94	IV-1	Titles. Natural language
21	57.41	II-15	Simple concepts. Complete combination
22	57.11	II-9	Simple concepts. Alphabetical first stage selection
23	55.88	II-13	Simple concepts. Complete species and superordinate
24	55.76	II-8	Simple concepts. Hierarchical selection
25	55.41	II-12	Simple concepts. Complete species
26	55.05	II-5	Simple concepts. Selected species and superordinate
27	53.88	II-7	Simple concepts. Selected coordinate and collateral
28	53.52	II-3	Simple concepts. Selected species
29	52.47	II-14	Simple concepts. Complete collateral
30	52.05	II-4	Simple concepts. Superordinate
31	51.82	II-6	Simple concepts. Selected coordinate
32	47.41	II-2	Simple concepts. Synonyms
33	44.64	II-1	Simple concepts. Natural language

FIGURE 8.1T ORDER OF EFFECTIVENESS BASED ON NORMALISED RECALL FOR 33 CRANFIELD INDEX LANGUAGES (AVERAGE OF NUMBERS)

Before considering some of the particularly striking aspects of the ranked order of effectiveness as given in Fig. 5.15T, there are certain points to be noted about this table. The normalised recall ratios range from 65.82% to 44.64% and this range encompasses some 33 different index languages plus 14 languages (or options) of the SMART system. It is impossible to state here what is a significant difference; most people who have been consulted agree that anything less than 1% is probably of doubtful significance, but that a difference of 3% or 4% almost certainly represents a significant change in performance. Rather than try to postulate on this point, we would prefer to rely on the consistency with which certain actions have certain effects.

For convenience of discussion, the normalised recall table, with the SMART results deleted, is reprinted as Fig. 8.1. It can be seen that the Single Term index languages rank 1, 2, 3, 4, 5, 6, 7= and 12 with the normalised recall ratio ranging from 65.82% and 61.17%. Starting from the base of natural language (with a score of 65.00%), the use of synonyms and word forms shows a slight improvement, whereas an enlargement of the classes by quasi-synonyms and hierarchical grouping detracts from the performance.

Of the six Controlled Term index languages, that using only the basic terms gave the best performance, with a ranking of 10= and a normalised recall ratio of 61.76%, this being a slight improvement on the lowest score with a Single Term index language. As narrower, broader and related terms are brought in, ranking orders for the other five Controlled Term index languages are 10=, 15, 17, 18 and 19, with the lowest score being 59.17%.

The searches on abstracts and titles gave four languages which ranked 13, 14, 16 and 20, the range being from 60.94% to 58.94%. The abstracts (which included titles) seem to be marginally better than the titles on their own. It is interesting that, with the abstracts, the confounding of word forms results in a slightly lower score, whereas the reverse is true with the titles.

The highest rank of the Simple Concept index languages is 7=, with a normalised recall ratio of 63.05%. Another language in this group is ranked 9, but the other thirteen Simple Concept index languages occupy the final ranks from 21 to 33. The two Simple Concept index languages which perform reasonably well are - surprisingly - those where the selection of additional related terms is based not on the classification schedules but on the rotated alphabetical index (see Vol. 1, Appendix 5.5).

In Fig. 8.1 it is significant that Single Term Natural Language I.1.a has a score of 65.00%, while Simple Concept Natural Language II.1.a has the lowest score of 44.64%. There is only one difference between these two index languages. In the former, the single terms are free; in the latter exactly the same single terms are interfixed into concepts. Index Language II.1.a represents the concept taken directly from the terminology of the document, e.g. 'conical afterbody', 'centrifugal compressor'; Index Language I.1.a uses exactly the same words, but they are broken down to the single terms, i.e. 'conical', 'afterbody', 'centrifugal', 'compressor'. It would therefore seem that interfixing is such a powerful device that it can severely depress the performance when calculated by the normalised recall ratio. Even when one considers the performance by coordination level cut-off, it can be seen from Fig. 4.700T and from the composite graph in Fig. 4.715P, that the Simple Concept Natural Language II.1.a has a very low maximum recall ratio, which is not compensated for by a particularly good precision ratio. Because it is so relatively inefficient, one finds that, for the Simple Concept index languages, the broadening of

classes by the use of various recall devices results in a considerable improvement in performance, which is contrary to the effect observed with the Single Term index languages. This leads to the following conclusions.

There was in this test an optimum level of specificity in the terms which were used. The conceptual terms of the Simple Concept index languages were over-specific when used in natural language, this high level of specificity being related to the strength of interfixing between the single terms of the natural language. Because of this, the broadening of the natural language concepts into more general classes resulted in a significant improvement in performance, in that it helped to overcome the high specificity. On the other hand, the Single Terms in natural language appear to have been near to the correct level of specificity; only to the relatively small extent of grouping true synonyms and word forms could any improvement in performance be obtained. Contrary to the experience of Simple Concepts, the broadening of the classes by the use of quasi-synonyms or hierarchical grouping resulted in a significant loss of performance. In between these two extremes of Single Term and Simple Concepts came the Controlled Terms. Less specific than the Concepts but more specific than the Single Terms, the effect of broadening the classes from the Controlled Terms Basic Terms (Index Language III.1.a) was to depress the performance, although not to the same extent as single terms.

While the evidence is not so easy to interpret from the tables and plots of the main test results as given in Chapter 4, it is quite obvious that within the various groups of index languages - where a direct comparison can be made - there is a difference between systems, and that these substantiate the rankings which are given in Chapter 5.

To restate the main conclusions more precisely

1. In the environment of this test, it was shown that the best performance was obtained by the use of Single Term index languages.

2. With these Single Term index languages, the formation of groups of terms or classes beyond the stage of true synonyms or word forms resulted in a drop of performance.

3. The use of precision devices such as interfixing and partitioning was not as effective as the basic precision device of coordination.

In the light of these unexpected conclusions, it is necessary to consider very carefully the test environment and to see whether there is any factor which could have distorted the results.

The subject field is a matter on which it is difficult to argue. There has in the past been a tendency to assume that, with an imprecise (mushy) subject language, where the same notion can be expressed in several different ways, there is the necessity for broad grouping of terms in the index language. Yet it seems possible that this imprecision is such that it is virtually impossible to make any logical practical grouping or class which can improve overall performance. To form a single class of two vague, imprecise terms may merely add confusion to confusion, so that any resulting improvement in the retrieval of relevant documents is more than outweighed by the increase in the retrieval of non-relevant documents.

In Chapter 6, the results were given for a set of questions dealing with aircraft structures, where, it has been earlier suggested, the subject language is less mushy. The results are not easy to interpret, but it appears probable that the assumption that aerodynamics represents the mushier language was unjustified. In the final chapter of Volume 1, we said that "It would seem, that next to the question of relevance assessments, the determination of the effect of subject language precision is the most important problem to be tackled". This opinion still holds, and we find it impossible to say categorically that the subject area of the test collection did not have an influence on the comparative test results.

Undoubtedly the size of the test collection (on which the normalised recall ratios are based) is smaller than one would have liked. The test results presented in Chapter 4, Section 1, show that the smaller sets of documents and questions were representative of the complete document collection and question set, but these tests were only concerned with the Single Term index languages, and it will be necessary to await confirmation on this point from the tests being carried out using the complete collection with the SMART system. However, there appears to be no justification for suggesting that the size of the test collection could have significantly affected the comparison between systems.

A matter that has already been raised in reviews of Volume 1 (e.g. Ref.14), and will undoubtedly be argued again is the matter of relevance decisions used in this test. It was in fact considered in the earlier volume, and the reader is referred in particular to the table on page 14 of Volume I. However, since that section was written, the matter of relevance has become the object of research and investigation in its own right, and it may be worth reopening and expanding the argument in the hope that some of the complexities introduced by psychological overtones might be clarified.

Consider first the matter of the evaluation of an operational information retrieval system, which we have earlier described as covering all stages from the first receipt of an enquiry to the stage of supplying the requester with the references to the set of documents (or, if the system is so designed, to an actual set of documents) which represent the system's answer to his enquiry. It is particularly stressed that the process starts with the first receipt of an enquiry. This enquiry is expressed in the form of a "stated requirement"; anyone with practical experience of information work will know that quite often the stated requirement is far removed from the real needs of the questioner. The greater the expertise of the information staff concerned, the greater the probability that it will be possible, before commencing a search, to reduce the gap between the real and stated needs of the enquirer.

However, in such a situation, namely the evaluation of an operational system, it is essential that the relevance assessments should be based on the real needs of the questioner; it therefore follows that the questioner must make the relevance judgements. Only if this is done can it be found whether there are any errors (i.e. the retrieval of non-relevant documents, or the non-retrieval of relevant documents) which are due to a failure to bridge the gap between the real and the stated needs. At the same time, however, it is necessary to determine the relevance of documents in relation to the stated needs. With these two sets of relevance judgements, it is possible to pinpoint the reasons for the failures in the complete system.

These two types of relevance are called "user relevance" and "stated relevance". The former can only be decided by the questioner himself, but "stated relevance" can be determined (as has been argued in the table on

page 14 of Volume 1) by anybody with reasonable knowledge of the subject field.

On the other hand, if the evaluation is only intended to cover a sub-system of the complete operational system, such as the index language, then there is not the same necessity of having "user relevance" decisions; in fact, such decisions could introduce an additional variable which might mitigate against the interpretation of the test results, and a set of "stated relevance" decisions could be more satisfactory.

So far the argument has been concerned with the evaluation of operational systems. All the tests of experimental systems have been or are being conducted in artificial, created environments. Under such circumstances, "user relevance" decisions cannot be obtained, and in the few tests so far carried out, "stated relevance" decisions of one kind or another have been used. However, in this particular project, as explained in the first Volume (pages 21 - 23) an endeavour was made to simulate "user relevance" decisions. At the same time (and contrary to what was done in Cranfield I), we deliberately eschewed any effort to interpret the stated needs; in all cases the search terms were based solely on the terminology of the question. Whether the original decision to simulate user relevance decisions was correct has already been considered (Vol. 1, page 114) and tentatively the conclusion was there reached that it might have assisted the interpretation of the test results if, instead, stated relevance decisions had been used. On the whole, this is a view to which we would still subscribe but for one fact. If stated relevance decisions had been used, and assuming the test results had shown the similar superiority of Single Term Natural Language, then it would have been virtually impossible to refute an argument that the results were unduly influenced by the relevance decisions.

In the artificial situation, a person - or a group of persons - is presented with a search question (which may have been devised by someone else) and a set of documents (or their surrogates in the form of titles or abstracts) and told to make a series of decisions as to which documents are relevant. He can be given specific instructions, such as the type of person that he is supposed to be or the purpose for which he is supposed to require the information. Whatever such instructions he may receive, he is ultimately faced with a sequence of words which make up the question, and other sequence of words which make up the documents, and by the intensity with which the words and the meaning of the question appear to match the words and the meaning of a document, he must decide that a given document is or is not relevant to a given question. In this artificial situation it seems reasonable to assume - and such experimental evidence as is available bears out the assumption - that there will be a closer direct match between the actual words of a question and a relevant document, than is the case in the natural situation of a questioner making user relevance decisions. Conversely, and just as important, there will, in the artificial situation, be a lower match between the question and a non-relevant document than will often be the case with user relevance judgements.

Under such circumstances, it is highly probable that system performance will be better with stated relevance decisions, than with user relevance decisions, since a source of possible error in the complete system has been eliminated. This is not an important factor in the present investigation, since the objective is not to obtain maximum performance per se, but is concerned with the comparison between the performance of different index languages. The important point is that stated relevance decisions which can

only be based on a match between words in the document and the question, might be expected to favour systems using precise natural language, while user relevance decisions might logically be expected to favour systems which bring in groups of related terms. The conclusion is therefore reached that the method of obtaining relevance decisions in this test could not have been responsible for the unexpected results, since any influence it might have had would have tended to work in the opposite direction.

Without going against the above argument, Vickery (Ref.15) rightly points out that "There are still verbal links between source document and question; the questions supplied by the author - some time after doing the research - were formulated after the cited papers had been read and possibly influenced the wording of his question." This raises two separate questions; firstly, is it very much different to what happens in a real life situation, and secondly, is the effect serious enough to distort the test results? To consider the first point, experience in the evaluation test of Medlars at the National Library of Medicine has shown that the majority of questioners are already aware of certain relevant documents before asking for a search to be carried out. It therefore seems likely that, in real life, search questions must often be influenced by the terminology of relevant documents, and therefore the procedure which was adopted in this test for obtaining questions is not far removed from what normally happens. If, however, the actions of those who prepared the search questions were significantly different from what happens in real life, then it is necessary to consider whether the results are likely to have been distorted. To determine whether this is so would require a far deeper analysis of the individual searches than has so far been done. Our own opinion is that if such an analysis were made, it would show that in the large majority of cases there had been no serious distortion, and it is difficult to believe that the few cases where it might have occurred would have been sufficient to produce the significant - and consistent - variations in performance.

The concept indexing was done by selecting from each document those concepts which appeared to be of importance. This being an intellectual task it is not possible to argue that it was done correctly. Readers of the reports on Cranfield I will recollect that the errors of the indexers were the cause of a significant number of failures to retrieve relevant documents, but that considered as a percentage of total indexing, it represented a very low "error rate". Usually in that test the errors were errors of omission. The higher level of indexing exhaustivity, and the longer time devoted to indexing each document made it less likely that these would occur in this project, and some analysis of the failures to retrieve relevant documents has not revealed any significant errors in this respect. Certainly it does not seem plausible to suggest that any such errors could have influenced the comparative results.

While the complete indexing was more exhaustive than would normally be the case, the assignment of an indexing weight to each concept permitted the testing of various levels of indexing exhaustivity. The test results are given in Chapter 4, Section 4, and again show that whatever the level of indexing exhaustivity might be, the effect of moving from Index Language I.1.a to Index Language I.6.a is consistent, and there is no evidence to suggest that the exhaustivity of indexing affected the comparison between different index languages.

Concerning this level of exhaustivity of indexing, it again becomes obvious that there was an optimum in regard to this particular document/question set. The lowest level of exhaustivity of indexing investigated was the search on titles only; the highest level of exhaustivity occurred with the

search on abstracts. Intermediary were the three levels of indexing done by the project staff. Figure 8.2T shows the normalised recall ratios obtained in these five cases, all using natural language terms.

Index Language	Average No. of Terms	Normalised Recall Ratio
Titles	7	59.76%
Level 1 Single Term Natural Language	14	62.88%
" 2 Single Term Natural Language	22	63.57%
" 3 Single Term Natural Language	33	65.00%
Abstracts	Approx 60	60.94%

FIGURE 8.2T NORMALISED RECALL RATIOS FOR FIVE LEVELS OF EXHAUSTIVITY

There is the possibility that the selection of terms by the indexer was more descriptive of the document content than those terms used for the titles and the abstracts, but the main variable in these five results concerns the level of indexing exhaustivity. It would seem that while the titles were at too low a level of exhaustivity, the gradual increase in the level, up to an average of 33 terms, brought about an improvement in performance. However, the higher level of exhaustivity represented by the abstracts (probably about 60 terms per document) was too high, resulting in the retrieval of large numbers of additional non-relevant documents, so that the performance only represented a slight improvement on that obtained with titles. This hypothesis is supported by the effect with titles and abstracts of enlarging the classes by the use of word forms. With titles, where it has been shown that the level of exhaustivity is too low, the use of word forms improves the normalised recall ratio from 58.94% to 59.76%. With abstracts, however, no such improvement is noted; already there are too many terms and the use of word forms results in a fall from 60.94% to 60.82%. Admittedly this in itself cannot be considered a significant change, but taken in the context of the other results, appears to be of some importance.

The compilation of the dictionaries or schedules was done, in the main, by Mr. Jack Mills. Although there can be few people more competent in such work, there can obviously be no guarantee but that different classes in the Single Term index languages might have given an improved performance as compared to natural language. However, it seems unlikely that the classes prepared for the Simple Concept index languages could have been solely responsible for the relatively poor performance as compared to the Single Term index languages. With the Controlled Term index languages, the classes of terms were formed on the basis of groupings given in the Thesaurus of Engineering Terms of the Engineers Joint Council, yet the use of any groupings except Narrower Terms (Index Language III.2.a) resulted in a loss of performance.

In Chapter 3, the statement was made that for any given question, the total number of postings of the search terms of that question must be equal to the total number of retrievals at the various coordination levels. To explain this point with a simple example, assume the search programme is

made up of four terms A, B, C and D, each of which have been used five times in the indexing of a set of documents as follows (x represents any other term or terms also used in indexing the documents):

Document Number	Index Terms
1	ADx
2	x
3	ACx
4	x
5	BCDx
6	x
7	Bx
8	x
9	ABCDx
10	x
11	BDx
12	x
13	Ax
14	x
15	BCDx
16	x
17	Cx
18	x
19	Ax
20	x

Searches for any combination of A, B, C and D would result in retrieval at various coordination levels as follows:

Coordination Level	No. of Documents Retrieved
4	1 (Document 9)
3	3 (Document 9, 5, 15)
2	6 (Document 9, 5, 15, 1, 3, 11)
1	10 (Document 9, 5, 15, 1, 3, 11 7, 13, 17, 19)

Thus the sum of the retrievals (1+3+6+10=20) is the same as the total number of postings for the four terms.

The particular significance of this point is the effect on retrieval performance of enlarging the classes. Assume that the search terms are broadened by being grouped with a related term, A_1, B_1, C_1 or D_1, and that these related terms have also each been used five times in the same set of 20 documents, the indexing being as follows:

Document Number	Index Terms
1	ADx
2	$A_1B_1D_1x$
3	ACD_1x
4	B_1x
5	$BCDA_1x$
6	C_1x
7	BC_1x
8	D_1x
9	$ABCDx$
10	A_1x
11	BDA_1x
12	C_1x
13	AB_1C_1x
14	D_1x
15	$BCDx$
16	D_1x
17	CA_1B_1x
18	C_1x
19	Ax
20	B_1x

Assuming the search now is for any coordination of $(A + A_1)$, $(B + B_1)$, $(C + C_1)$ and $(D + D_1)$, the retrieval at different coordination levels will be as follows

Coordination Level	No. of Documents Retrieved
4	2 (Document 5, 9)
3	8 (Document 5, 9, 2, 3, 11, 13, 15, 17)
2	10 (Document 5, 9, 2, 3, 11, 13, 15, 17, 1, 7)
1	20 (Document 5, 9, 2, 3, 11, 13, 15, 17, 1, 7, 4, 6, 8, 10, 12, 14, 16, 18, 20)

Again it is shown that the sum of the retrievals (40) equals the total postings for the four groups of terms. Assume now that there were four relevant documents, numbers 3, 7, 9 and 15. The performance in the two cases would then be as follows

Coordination Level	Case A				Case B			
	R	N-R	Recall Ratio	Precision Ratio	R	N-R	Recall	Precision Ratio
4	1	0	25%	100%	1	1	25%	50%
3	2	1	50%	66%	3	5	75%	38%
2	3	3	75%	50%	4	6	100%	40%
1	4	6	100%	40%	4	16	100%	20%

While this particular result has obviously been prepared to illustrate the point, it would seem that this is an example of what has been consistently happening in the test searches with the Single Term and Controlled Term index languages. Whereas the broadening of the term classes has increased the recall of relevant documents at higher coordination levels, the effect of doing this has been more than offset by the increased number of non-relevant documents. Only when the index terms being used are too precise, as in the case of the Simple Concept Natural Language, can the formation of broad classes of terms bring about an improvement.

Finally, it is necessary to consider the measures which have been used in this test, and to ask whether it is possible that some other measures would have brought about a change in the comparative results. Obviously suspect is the normalised recall ratio, based on a simulated rank output. While at first it might seem that such a measure is likely to weigh in favour of systems having high recall ratios, it is in fact mainly influenced by the first two ranked documents. At this stage, the recall ratios, as can be seen from Figures 5.11T - 5.14T, are as follows

Recall Ratio at Document Output Cut-off of 2	Index Language
23%	I.2, I.3
22%	I.1
21%	I.6, I.7
20%	
19%	I.5, I.8, II.9, III.2
18%	II.3, II.12, IV.3, IV.4
17%	II.10, III.1, IV.1, IV.2
16%	I.9, II.11, III.3, III.4
15%	II.5
14%	II.13
13%	II.2, II.8, III.5
12%	II.1, II.4, II.6, III.6
11%	II.7, II.15
10%	
9%	II.14

It will be seen that with the exception of Index Language II.3, which (at 18%) rises from 28 to 10=, there is a strong correlation between this ordering and the final ordering as given in Table 8.1. With the document output cut-off method, recall and precision are, as we explained earlier, completely interdependent, and therefore it would appear to be a measure that is quite impartial as between recall and precision. It is known that others are investigating different measures, and most of those that have been proposed have already been considered in Chapter 3. Now that the results of this test are available, it is to be hoped that proponents of new measures will be able to demonstrate any superiority over those used in this report. Until such time, there appears to be no reason to suggest that the measures have affected the comparative results.

With the possible doubtful exception of the subject field, there appears to be nothing in the test environment which could be held responsible for serious distortion of the results as between one system and another. Therefore it is necessary to proceed on the assumption that the results are

correct, and attempt to find the reasons why they should be as they are.

It would be quite incorrect to suggest that no-one has previously argued in favour of single terms, natural language and coordination, for these were the bedrock of the Uniterm System of coordinate indexing as originally propounded by the late Dr. Taube in 1951. But while the device of coordination - or, as we would now term it, post-coordination - continues in favour, there are few who now accept (for Information Retrieval Systems) uncontrolled vocabularies, and some who insist additionally on the use of links and roles. Even Dr. Taube himself was, within a couple of years of the inception of the Uniterm System, to start devising associated maps, and there is no indication, in the writings at that time of the group at Documentation Inc., of any awareness that the resultant increased recall would be more than offset by the lower precision.

There are doubtless indexes in existence which follow the original Uniterm principles, but one of the few persons who has consistently, in print, advocated the use of natural language and coordination is Mr. Th. te Nuyl with his L'Unité System (Ref. 16). Even so, for most people L'Unité System will be associated mainly with the ingenious coding system rather than the use of natural language. It is of interest to note that the clustering of the natural language terms into broad alphabetical groups (as in L'Unité) brings about the confounding of word forms, so, possibly unintentionally, te Nuyl did adopt a coding device which was, it would appear from the results of this test, the only way to improve performance over natural language.

Then there are, of course, permuted title indexes, which use the natural language of the title, but these can hardly be considered in the same light, since they do not have the facilities of post-coordination.

Therefore it is against these few that are ranged, for instance, the activities over the last fifty years of the Universal Decimal Classification, which is probably now more widely used than ever before. At the same time, a large number of national and international organisations are engaged in constructing thesauri, while many groups in the research field are endeavouring to develop computer methods for the formation of classes of terms (e.g. Ref. 17).

The effort that is put into these activities, by whichever process the classes may be formed, is presumably influenced by the widely held belief that it is only by such means that a high recall ratio is obtainable. Yet even in Cranfield I we reported that a recall ratio of 97% was possible merely by using the words in the titles. There was no way of knowing in that experiment the corresponding precision ratio, but it was not only assumed (correctly) that it would be very low, but it was also assumed that it would be lower than would have been the case if such a recall ratio had been obtained with a conventional index language.

As far as this test is concerned, the latter assumption would be unjustified; is it now reasonable to assume that the grouping of natural language terms to form controlled vocabularies, or the broadening of search strategies, must inevitably result in a loss in overall performance?

We would certainly not make such a statement on the basis of this single test; however, it would be surprising if the comparative test results were peculiar to the particular environment of this test, and it does seem

that the results are sufficiently convincing to justify a fresh look at firmly held beliefs.

The present position is that, in the very large majority of cases, the manager of an information retrieval system employs indexers who apply their intelligence to the documents which are to be entered into the system. The indexers select the important concepts which they then translate into the terms of a controlled vocabulary (e.g. a thesaurus or classification schedule). This has possibly involved a considerable amount of intelligence in its compilation, and requires more intelligence for its maintenance. At the stage of a question being received, the search staff will apply their intelligence to deciding the exact meaning of the question and to preparing a suitable search programme, using the terminology of the system. In doing this they will take advantage of the intelligence that has been applied to denoting the relationships between the index terms, either in the arrangement of a classification schedule or by the visual display of a thesaurus. Normally the search is then made, and the questioner receives the output.

It would appear to be a reasonable assumption that the more intelligence that is applied to any of these three stages (i.e. the indexing, the compilation and maintenance of a controlled vocabulary, and the determination of the search strategy) the better the result should be in terms of recall and precision. For example, the most direct measurement (which can be isolated) of the effect of using intelligence in this project is given in the series of results presented in Figures 4.840P - 4.845P, and again in Figure 5.21T. From the latter it can be seen that Search E (where intelligence was used in deciding the acceptable combinations of search terms) resulted in a 1% - 2% increase in normalised recall ratio as compared to Search A (where any combination of terms was accepted).

However, the mere use of intelligence is not enough, for in all cases it is necessary that the intelligence should be applied intelligently in relation to the needs of the system. An example of this relates to the level of exhaustivity in indexing. One cannot say categorically that the selection of seven terms to index a document indicates more or less intelligence than the selection of sixty terms, for it could be argued that, while the latter certainly requires more clerical effort, the former requires more intelligence in selecting the most important terms. However, in the environment of this test, the results show that intelligence was more effectively applied in selecting an average of some thirty-three terms for indexing. (It should be emphasised that this is in no way intended to imply that this level of exhaustivity would be the optimum in a different environment.)

Intelligence is a valuable - and relatively expensive - commodity, and should be used in the most efficient manner. An interpretation which could be placed on the results of this test is that there may well be operational situations in which one should take advantage of the intelligence that has already been applied by the author of a paper by accepting as index terms the key-words in the title or abstract. There is the folk-lore that titles do not represent a correct indication of the content of the document or that authors cannot write reasonable abstracts, and everyone can quote examples where this is the case. Such examples are, however, comparatively rare; for instance, of the many thousands of research papers issued by the National Aeronautics and Space Administration, it would be very difficult to find a single paper where the title did not present an adequate representation of the main subject matter or the summary did not cover all the items of importance. We would therefore argue that, in many operational systems,

a case could be made out for dispensing with indexers within the system and for using the persons thus displaced to screen the search output. The indications are that information staff, merely on the basis of reading titles, can quickly and reliably screen out between 50% and 80% of the non-relevant documents which are retrieved in an average search, without any loss of relevant documents. Such a process would, in many cases, result in a far better service to the user, by giving an operational performance higher than that now obtained.

Additionally it can be argued that there are situations where the intellectual effort involved in the construction and maintenance of controlled vocabularies is unjustified. It is with very strict qualifications that this viewpoint is advanced; in certain subject fields it is almost certainly not true. One critical factor (there are certainly others) could be the occurrence of real synonyms as opposed to quasi-synonyms, or near-synonyms. To illustrate the difference between subject fields, it has been said that there are twenty-one synonyms for the term 'aspirin' (apart from trade names), any of which may be found in the literature, whereas in the subject field used in this project the number of true synonyms (in contrast to quasi-synonyms) was relatively small, the improvement in performance by grouping synonyms was equally small and was hardly sufficient to justify moving from natural language terms. It is difficult to believe that a controlled vocabulary should be less efficient than natural language, even though the evidence of this test points to such a conclusion. Apart from the theoretical reasons already advanced for this being so, there could be a more fundamental reason, and the answer may again lie in the intelligent application of intelligence. No one could deny that a large number of highly intelligent people have given a considerable amount of time to the maintenance of the Universal Decimal Classification or to the preparation of the Thesaurus of Engineering Terms of the Engineers Joint Council. It can, however, legitimately be asked whether these activities represent intelligent applications of intelligence. It may, in fact, be not possible to generate an efficient controlled vocabulary without the applied and close attention, over a relatively long period, of the professional staff of the operating group.

This test has shown that natural language, with the slight modifications of confounding synonyms and word forms, combined with simple coordination, can give a reasonable performance. This means that, based on such practice, a norm could be established for operational performance in any subject field, and it would then be for those who proposed new thesauri, new relational groups, links or roles, to show how the use of their techniques would improve on the norm. The availability of a computer programme, such as a simplified version of the SMART programme of Professor Salton, would make this relatively inexpensive.

Every quotation that has been taken from the book by Professor Wilkins is relevant to our final argument. We make no forcasts that a coordinate system will break down when it reaches a certain size, or any other speculations of this kind, for there is nothing that has been done in this project - or in any other experimental project recently completed or under way - which can justify categorical statements of this nature. As Cranfield I gave indications of the situation over the general field of information retrieval systems, so this project has shown, in a more specialised area, some of the basic problems which beset any and every operator of an information retrieval system. The results can be taken as an indication of what might be done to improve efficiency, but the application of the results to any given situation can only be on the basis of an evaluation of the operational system concerned.

In conclusion, we make no apology for repeating the quotation given at the beginning of this chapter

> "The first step in testing a theory (qua theory) is to examine it to see what deductions can be made from it - to set up postulates which may be tested either experimentally or by observations of the 'real-life' situation. This is to say, the first step in testing a theory is to state the practical consequences of it. If the deduced practical consequences (operational definitions) are proved to be unsustained, the theory is discredited. No theory can ever be proved to be true; it is held for so long as no better theory can be found."

To put it more colloquially "the proof of the pudding is in the eating". It must remain so with the results and conclusions of this project.