CHAPTER 3

Methods for Presentation of Results

Lord Kelvin is often credited with remarking, 'When you can measure what you are speaking about and express it in numbers you know something about it, but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind'. The problem of validity - how closely do the figures relate to the 'thing we are talking about' must be separated from the problem of reliability - how accurate are the figures themselves.

L.T. Wilkins: Social Deviance, page 147

In Cranfield I, the results of the main project did no more than record what is now generally known as the Recall Ratio, which was calculated on the basis of $\frac{100R}{C}$ where R equals the number of relevant documents retrieved and C equals the total number of documents in the collection which are relevant to the questions. In the subsequent test of the Western Reserve University Index, (Ref. 2) measurement was carried to the stage where, by making relevance assessments of all the retrieved documents, it was also possible to calculate what was originally called the Relevance Ratio, but which is now generally known as the Precision Ratio, namely $\frac{100R}{L}$, where L equals the total number of documents retrieved in the series of searches. In the course of this evaluation of the W.R.U. Index, the effect of varying the exhaustivity of indexing was measured, and allowed the production of the first - and, incidentally, so far the only - performance curve from the Cranfield project. It is reproduced in Fig. 3.1P and showed two interesting characteristics. The first was the inverse relationship between recall and precision, which has been considered at some length in Volume 1 of this report. The second point was that, when documents of lower relevance, were accepted, there was at any given level of indexing exhaustivity, a lower recall ratio but an improved precision ratio. It was tentatively suggested that this latter point was connected with a variation in the average number of relevant documents for each question, and that, for any given situation, it would be necessary to state also what was to be later termed the Generality Number*, expressing it as $\frac{1000C}{N}$, where N equals the total number of documents in the collection.

*In the earlier volume of this report, this was called Generality Ratio.

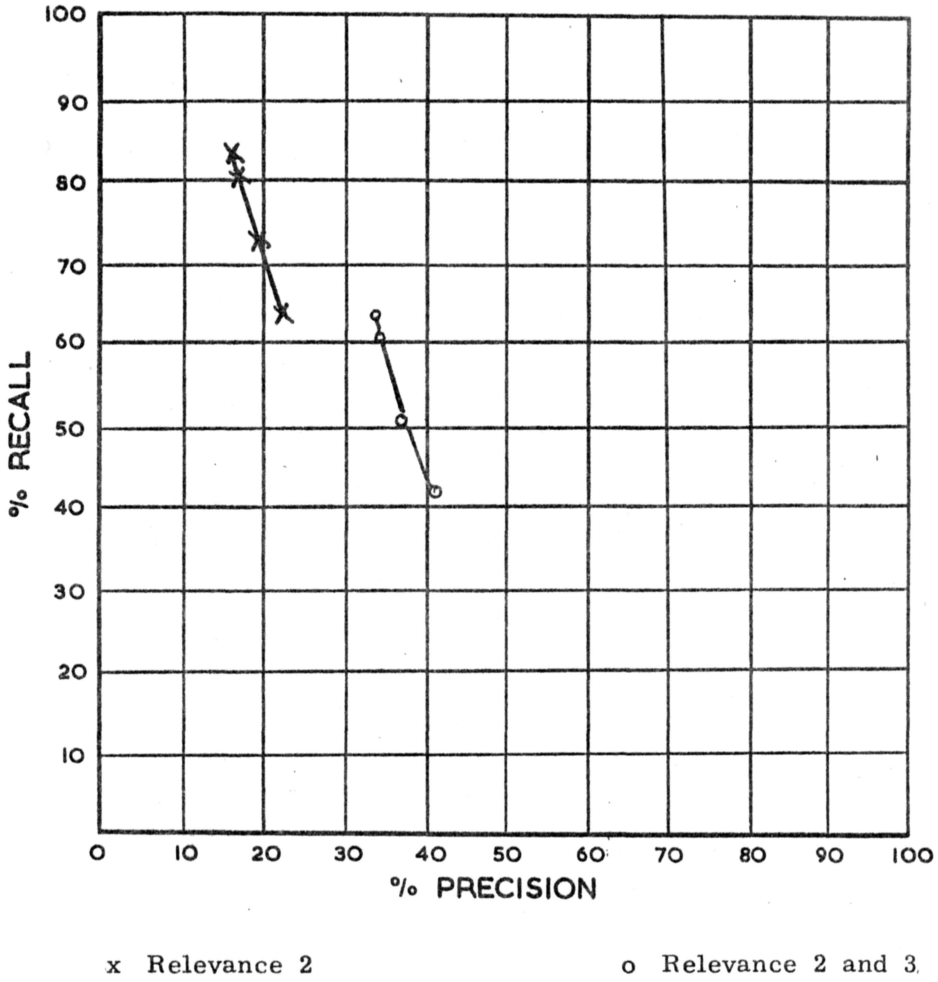


FIGURE 3.1P PERFORMANCE CURVE OBTAINED WITH FACET INDEX IN W.R.U. TEST

While the recall and precision ratios have been generally accepted as performance measures for information retrieval systems, they have also aroused some criticism. No serious attempt has been made to answer this criticism, partly because it was mostly trivial and never supported by experimental data, but mainly because an intention of Cranfield II was to investigate the performance measures which could or should be used. For this, sets of performance data were required and it was known that for every set of figures in Cranfield I, there would be hundreds of sets in Cranfield II, and it was obvious that the decisions regarding the measures to be used and the methods of presenting the test results would be of major importance. The programme of work which this aspect of the project has involved has been considerable, with many sets of results being calculated in a number of different ways. Based on this work, which has taken up a significant part of the effort during the last eighteen months of the project, the decision was finally reached that the most satisfactory method of calculating results involves three measures, namely Recall Ratio and Precision Ratio with, additionally, the new measure of Fallout Ratio. For the presentation of results on a plot, it is believed that, in the large majority of cases, the most straightforward and most meaningful method is the Recall/Precision curve. These are, in general, the measures used in this report, although to illustrate certain points various other measures and methods of presentation are used.

A detailed account of the Cranfield work on performance measures has been presented in a thesis by M. Keen, but the following is a resumé of the more important points which led to the decisions; other matters relevant to the presentation of results in this volume are also considered.

In tests of experimental systems, it is essential that measures should be used that accurately reflect the changes in the particular component being tested, which primarily, in this particular test, was a range of index language devices. In addition, there is the strong desirability, if not the absolute necessity, that it should be possible to make direct comparison between different sets of test results.

Measures of retrieval performance may be used in experimental tests of information retrieval systems when the following requirements are met:-

1. A document collection of known size to be used in the test;
2. A set of questions, together with decisions as to exactly which documents are relevant to each question;
3. A set of results of searches made in the test; these usually give the numbers of documents retrieved in the searches, divided into the relevant and non-relevant documents.

The successive dichotomies of the total collection have been displayed by B.C. Vickery (Ref. 13, page 174) by the following table:-

<table>
<tr><td colspan="4">TOTAL COLLECTION</td></tr>
<tr><td colspan="2">RELEVANT</td><td colspan="2">NON-RELEVANT</td></tr>
<tr><td>NOT RETRIEVED</td><td colspan="2">RETRIEVED</td><td>NOT RETRIEVED</td></tr>
<tr><td>(c)</td><td>(a)</td><td>(b)</td><td>(d)</td></tr>
</table>

The more usual way to present the categories is in the form of a 2 x 2 contingency table as shown in Fig. 3.2. The notation given in this figure will be used throughout the remainder of this report.

	RELEVANT	NON-RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d = N (Total Collection)

FIGURE 3.2 2 x 2 CONTINGENCY TABLE

Whether it is correct to regard the values that result from retrieval tests as components of a 2 x 2 table in the statistical sense, and thus apply the principles and tests that have been developed for this situation in statistics, is an unanswered question, and at this stage, therefore, the use of this table is purely for convenience.

As mentioned earlier, there is the necessity of being able to make a comparison between several sets of results obtained in different conditions. This can only be done when it is known exactly which variables are altered in the different situations; two such situations are considered.

Assuming N (the total collection) remains constant, a, b, c and d can each vary, while a + b (total retrieved) and c + d (total not retrieved) remain constant. More common is the situation where all the above six values change, but a + c (total relevant) and b + d (total non-relevant) do not alter. This is to say that the numbers of relevant and non-relevant documents remain the same, but the numbers of retrieved and not retrieved, together with the four categories making up these groups, all vary. In such cases the change could be due to the 'cut-off' applied, that is the point in the search where the rules do not allow any further documents to be examined. At this stage the search is stopped and a record made of all the documents retrieved, both a (relevant) and b (non-relevant). A different cut-off results in a different set of values for a and b, thereby changing c and d, but without in any way affecting a + c or b + d. Alternatively, the change could be due to different indexing decisions or to different search strategies.

The second point to consider is the variables that affect a + c, b + d and N. If the decision as to what is relevant (a + c) is altered, then it must also result in a change for the total of non-relevant (b + d); if the collection size (N) is changed, other values in the table may change. Although significant changes of this nature occur rarely in operational retrieval system tests, it is necessary to consider the matter in experimental tests. Either type of change, i.e. altering the number of relevant documents or altering the collection size, can vary the number of relevant documents in relation to the collection size. Examples of the two types of situations can be taken from these tests. Relevance decisions were based on four levels of relevance; if we consider Relevant 1 documents, there are 12 such documents relevant to the 42 questions of subset 2. Relevance 1 and 2 documents come to 57, Relevance 1, 2 and 3 documents total 154 and Relevance 1-4 documents come to 198. It can be seen that changing the decision as to the relevant documents (a + c) materially alters the proportion of relevant documents in the complete collection (N).

On the other hand, the collection size can be changed. Originally there were 1400 documents in the collection. A subset of the collection was formed which consisted of 200 documents; a characteristic of this subset was that it retained all of the 198 documents that were relevant to the 42 questions of subset 2.

While the number of relevant documents is now held constant, the proportion changes because of the reduction in the document collection from 1400 to 200 documents. It is convenient to express this variation as a parameter, and this is the aforementioned Generality number i.e. $\frac{1000\,(a + c)}{N}$, the total relevant documents divided by the collection size, with a constant. This parameter is not a measure of retrieval performance, but one which reflects the environment of the relevance decisions made; e.g. if the generality number for a set of questions is 5, this means that there are, for each question, an average of five relevant documents for every thousand documents in the collection, irrespective of what the actual size of the collection might be. For the example given above, the change from the larger to the smaller collection size (bearing in mind that there are 42 questions) changes the generality number from $\frac{1000 \times 198}{42 \times 1400} = 3.4$ to $\frac{1000 \times 198}{42 \times \cancel{1400}\ 200} = 23.6$. Therefore, as far as retrieval performance is concerned, the significance of a change in either the relevance decisions or the collection size is that in both cases it is the generality number which alters.

The single performance measures that can be used can be listed as follows:-

$\frac{a}{a + c}$ usually known as Recall Ratio; at Western Reserve University it is called 'Sensitivity', and has also been called 'Hit Rate'.

$\frac{c}{a + c}$ complementary to recall ratio. Called by Fairthorne, 'Snobbery Ratio'.

$\frac{a}{a + b}$ now generally known as Precision Ratio, formerly called by Cranfield 'Relevance Ratio'. Also described as 'Pertinency Factor' or 'Acceptance Rate'.

$\frac{b}{a + b}$ complementary to precision ratio. Called by Perry, 'Noise Factor'.

$\frac{b}{b + d}$ here called Fallout Ratio.

$\frac{d}{b + d}$ complementary to fallout ratio. Called by Western Reserve University, 'Specificity'.

Use of any of these single measures, either reflecting the retrieval of the relevant items or the retrieval of non-relevant items, is inadequate to reflect the performance of a system. High recall can mean very low precision, or vice versa, and the mere statement that the recall ratio is 99% means little, for it might only be achieved by retrieving more than half of the total collection.

While many different combinations of single measures have been proposed, they fall into two groups: 'twin variable measures' and 'composite measures'.

For the former, one of each of the single measures is taken and a comparison made between them by observing the relative changes in the two values, but retaining each value as a separate entity. The two major pairs of single measures are recall with precision and recall with fallout.

Examples of recall/precision ratios are given in Figs. 3.3 and 3.4. Fig. 3.3T illustrates the situation for a set of 20 searches where the variable being tested is the search coordination level, that is the number of search terms which must be matched with the index terms. At each different level, a cut-off is applied and the number of documents retrieved, relevant and non-relevant, is recorded. Since the total number of relevant documents is known, the recall and precision ratios can be calculated, as shown in the table. Alternatively these ratios can be plotted as on the graph (Fig. 3.3P) with the five performance points connected to make a recall/precision curve. In Fig. 3.4T are given the results of a series of searches with the same set of questions but with different search requirements. The particular change is incidental to the present discussion, but in fact whereas search X accepted any combination of terms, search Y would not accept certain terms unless some other given term was also present. (This matter of search strategy was discussed in Chapter 2). The result of this change was a different set of performance figures at the five coordination levels. The contrast between search X and search Y can be seen by comparing the tables or from the graph (Fig. 3.4P), which shows clearly that the maximum recall figure has fallen sharply in search Y, but on the other hand at any given recall ratio of 65% or less, search Y will give a higher precision ratio than search X.

Coordination Level	Documents Retrieved		Recall Ratio	Precision Ratio
	Rel (a)	Non-Rel (b)	a/a + c	a/a + b
1	133	16,492	95.0%	0.8%
2	108	6,642	77.1%	1.6%
3	80	1,825	57.1%	4.2%
4	57	430	40.7%	11.7%
5	39	94	27.9%	29.3%

Relevant Documents (a + c) = 140

Generality number 5.

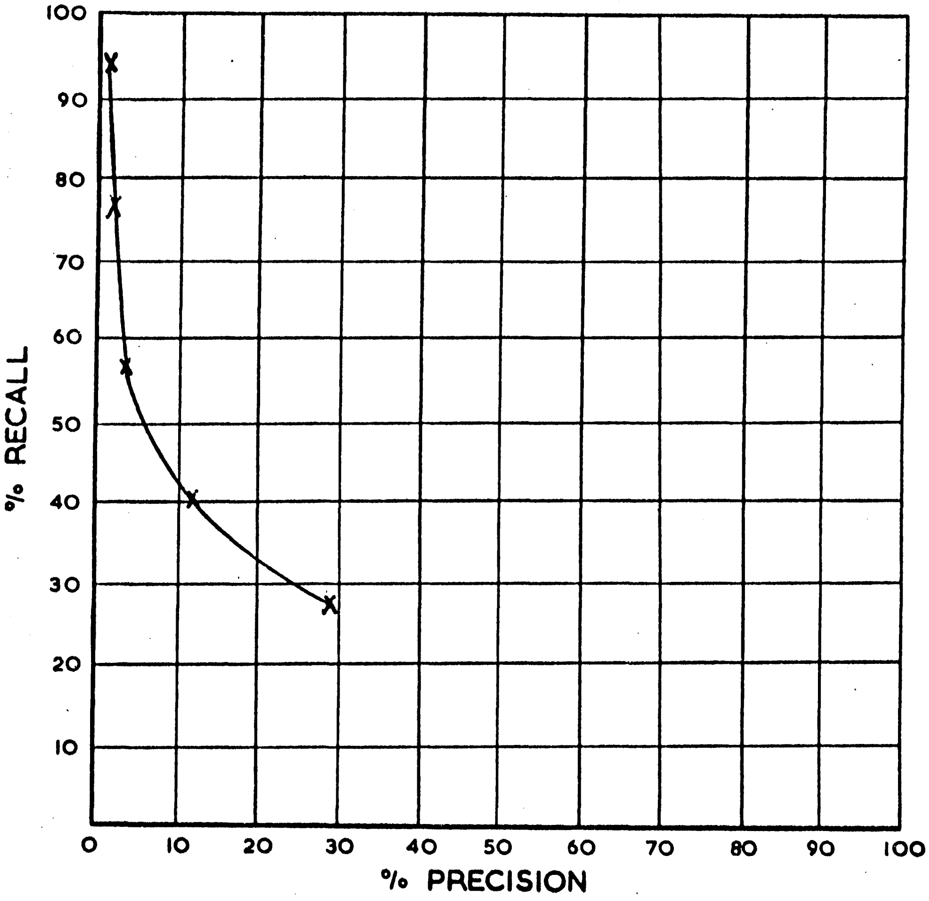


FIGURE 3.3TP TABLE AND PLOT OF TEST RESULTS FOR 20 SEARCHES SHOWING RECALL AND PRECISION RATIOS FOR SEARCH X.

Coordination Level	Documents Retrieved Rel (a)	Documents Retrieved Non-Rel (b)	Recall Ratio a/a + c	Precision Ratio a/a + b
1	97	2,674	69.3%	3.5%
2	77	788	55.0%	8.9%
3	56	220	40.0%	20.3%
4	37	30	26.4%	55.2%
5	33	17	23.6%	66.0%

Relevant documents (a + c) = 140

Generality number 5

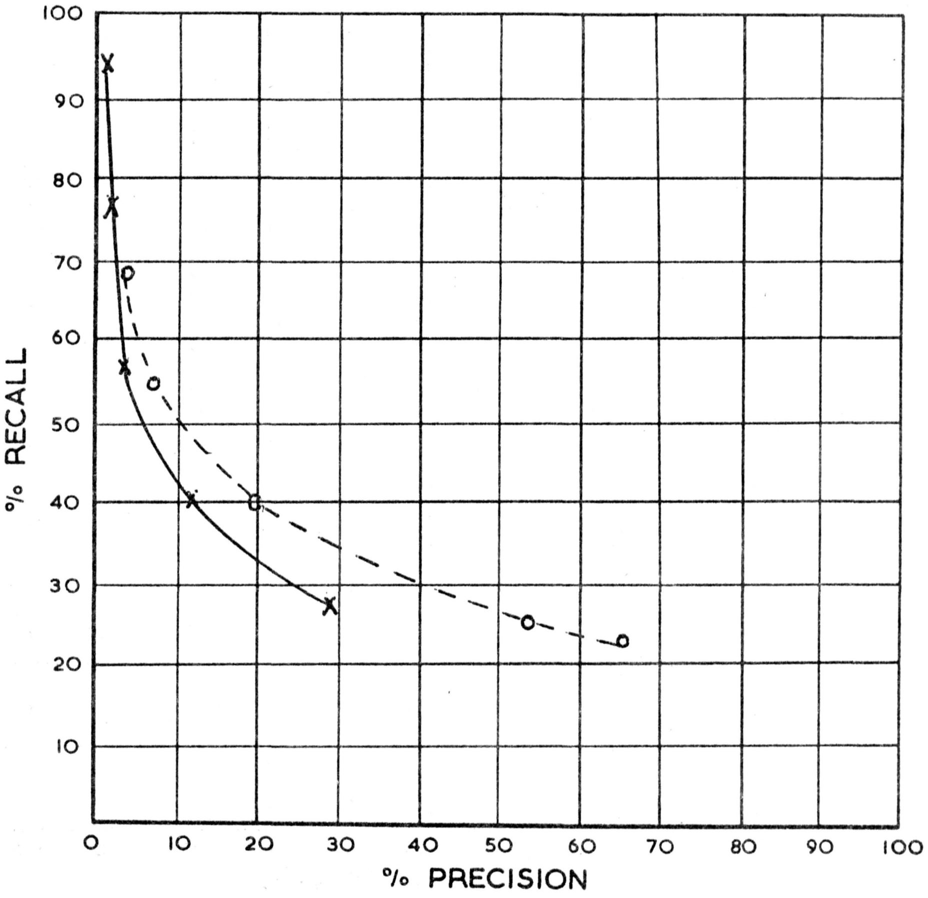


FIGURE 3.4TP TABLE AND PLOT OF TEST RESULTS FOR 20 QUESTIONS SHOWING RECALL AND PRECISION RATIOS FOR SEARCH Y (BROKEN LINE) (SEARCH X CONTINUOUS LINE)

A comparison of the recall ratio with fallout ratio can be made in the same way. We are not aware of any previous occasions when the fallout ratio has been used for presenting test results, although Swets (Ref. 4) has discussed its possible use. In that it measures the ratio of the non-relevant retrieved to the total non-relevant in the collection $\frac{b}{b + d}$, it is very sensitive to N, the total number of documents in the collection. While it might not be found to be particularly satisfactory for tests on operational systems, it has an attraction in experimental testing where collections of different but known size are being tested, since it automatically compensates for the changes in size. Fig. 3.5T takes the figures of Fig. 3.3T and Fig. 3.4T and replaces the precision ratio by fallout ratio. A characteristic of fallout ratios is that they tend to be concentrated at low numbers; for this reason the figures are taken to three places of decimals and the resultant plot of recall ratio against fallout ratio is clearer if made on a semi-log scale, as in Fig. 3.5P. In this case the better performance is obtained when the curve is nearer the top left hand corner, whereas the recall precision curve is optimised towards the top right hand corner. Therefore, as in Fig. 3.4P, search Y is shown to give a generally improved performance over search X.

Either of these twin measures is satisfactory for presenting the performance of systems where the generality number is held constant, although the argument has been advanced that a plot of recall/precision is not valid since both ratios contain a (relevant retrieved). It has been incorrectly argued that in plotting $\frac{a}{a + c}$ against $\frac{a}{a + b}$, all the a's cancel out, with the result that the factors being plotted are c against b. Fairthorne (Ref. 5) has said that a more reliable precision ratio is given by what he calls the 'distillation ratio' which is $\frac{a}{a + b} - \frac{c}{d}$. However, he agrees that when the correction factor of $\frac{c}{d}$ is negligible compared with the precision ratio, the latter is a valid measure. In fact, in the results presented in Fig. 3.3T, the correction factor at the coordination level of five terms is 0.0038, which can definitely be considered negligible.

Rees (Ref. 6) argues against precision ratio in favour of a measure that is complementary to fallout, namely $\frac{d}{b + d}$, on the grounds that it takes into account one of the vital parameters in a retrieval system - size of file. To some extent this is true, but it is a matter which has to be approached very carefully. The difficulty lies in determining exactly what is the correct value of N, that is to say how many documents can validly be considered to form the total collection in regard to any question. This matter is considered in more detail later in this chapter. It is true that the same difficulty arises in calculating the generality number, but if N is known, then it is just as easy to calculate the generality number as to

Coordination Level	Recall Ratio		Fallout Ratio	
	X	Y	X	Y
1	95. 0%	69. 3%	59. 196%	9. 598%
2	77. 1%	55. 0%	23. 841%	2. 828%
3	57. 1%	40. 0%	6. 551%	0. 790%
4	40. 7%	26. 4%	1. 543%	0. 108%
5	27. 9%	23. 6%	0. 337%	0. 061%

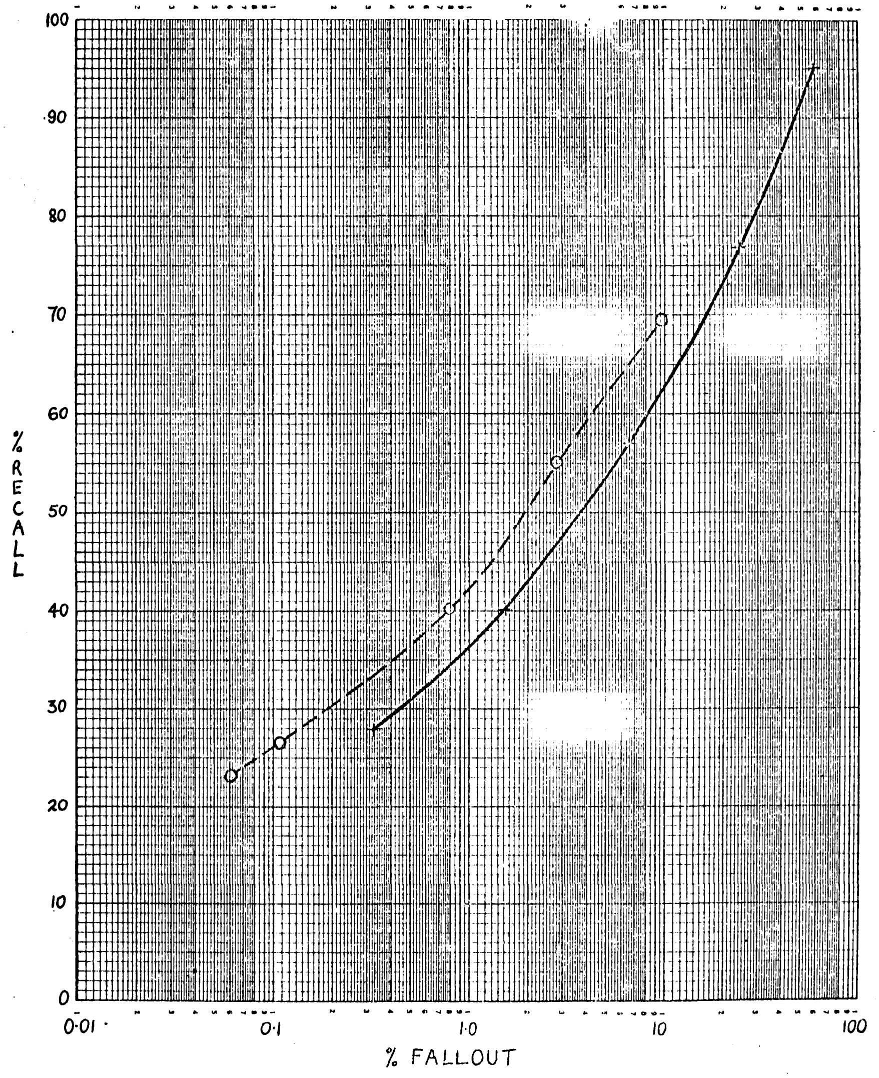


FIGURE 3. 5TP TABLE AND PLOT OF FALLOUT RATIOS DERIVED FROM FIGURES 3. 3T and 3. 4T FOR SEARCH X (CONTINUOUS LINE) AND SEARCH Y (BROKEN LINE)

calculate fallout.

A possible solution to making a full presentation of performance on a single plot is shown in Figs. 3.6P and 3.7P. Before considering these it is necessary to consider the relationship between the individual ratios of recall, fallout and precision, together with the generality number. These four ratios or parameters completely describe a given set of performance results in a retrieval table in terms of the measurements most likely to be of importance in presenting retrieval performance. However, it is only necessary to obtain any three of these in a given situation, since the fourth is then mathematically determined and can be written in terms of the other three. The four equations are:-

(1) R (Recall Ratio) = $$\frac{\left(\frac{F(1000 - G)}{1 - P}\right) \times P}{G}$$

(2) F (Fallout Ratio) = $$\frac{\left(\frac{R \times G}{P}\right) - (R \times G)}{1000 - G}$$

(3) P (Precision Ratio) = $$\frac{R \times G}{(R \times G) + F(1000 - G)}$$

(4) G (Generality Number) = $$\frac{1000}{\left(\frac{\frac{R}{P} - R}{F}\right) + 1}$$

where R (Recall Ratio) = $\frac{a}{a + c}$

F (Fallout Ratio) = $\frac{b}{b + d}$

P (Precision Ratio) = $\frac{a}{a + b}$

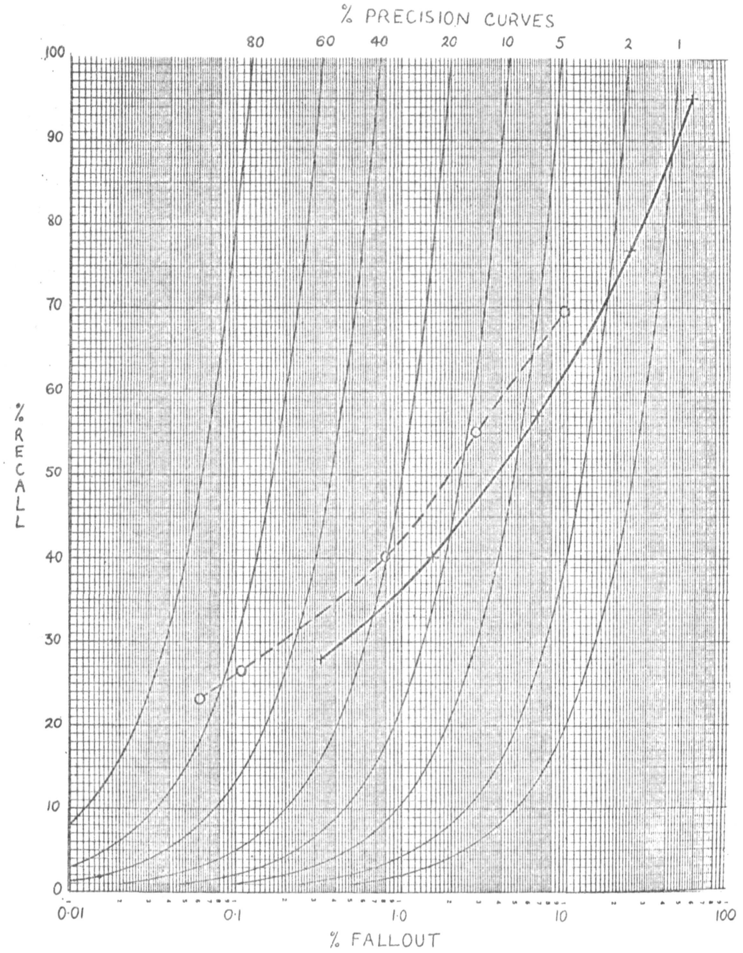


FIGURE 3.6P PLOT OF RECALL AND FALLOUT RATIOS AS FIGURE 3.5P SHOWING THE PRECISION RATIO CURVES

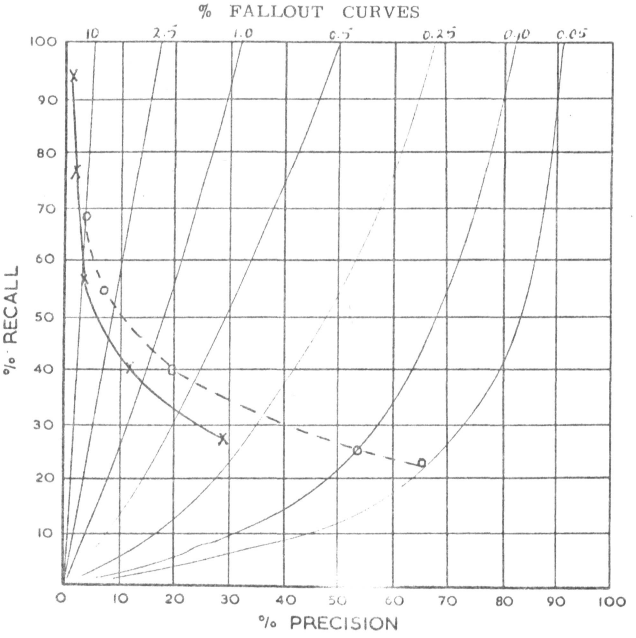


FIGURE 3.7P PLOT OF RECALL AND PRECISION RATIOS AS FIGURE 3.4P SHOWING THE FALLOUT RATIO CURVES

$$G \text{ (Generality Number)} = \frac{1000(a + c)}{N}$$

Thus equation (1) shows how, given the fallout and precision ratios together with the generality number, the recall ratio can be determined by calculation, and the other three equations show the other combinations possible. Because of this relationship, it has been possible to prepare, by computer, the figures for a series of situations where the generality number ranges from 1 - 50, recall from 5% to 100% and precision from 0.5% to 100%. In Appendix 3.3 is given this full set of tables for F (fallout) at varying generality numbers. From this set of tables, it is possible to plot on a recall/precision graph, the curves for fallout, or on a recall/fallout plot the curves for precision at all levels for any given generality number. For the example being considered, Fig. 3.6P shows the former, while Fig. 3.7P shows the precision curves on a recall/fallout graph. From either of these graphs it can be seen, for instance, that for search Y (the dotted line) at a recall ratio of 40%, precision ratio was 20% and the fallout ratio 0.8%. As the generality number for this set of searches is 5, the above figures can be confirmed from the sheet in Appendix 3.3A for generality number 5. In the column for recall of 40% and in the line for precision of 20%, fallout is 0.803%.

In a large number of situations arising in this test, comparison is made between various systems where everything is being held constant with one exception such as, for instance, the index language. In these circumstances the generality number remains constant and therefore the fallout measure does not contribute to the presentation of the results. In spite of the fact that there are some situations where comparative results are presented when the testing has been done on collections of different sizes, (with therefore, different generality numbers), the decision has been taken, as previously stated, to present the main sets of results on recall/precision graphs. The positive reason for doing this is that discussions with a number of people have led to the conclusion that such a graph can be more readily understood than a recall/fallout graph in that it more closely reflects the required performance aspects of a system. This may, of course, be due to the fact that recall/fallout graphs are unfamiliar compared with recall/precision graphs, and our decision is certainly not intended to imply that the latter are, in experimental work, basically superior to recall/fallout graphs.

In the course of this project, we have also considered a number of 'composite' measures which have been suggested. Swets (Ref. 4) argued that twin variable measures (e.g. recall/precision) were 'an unnecessarily weak procedure', but qualified this by assuming that a real retrieval system has a constant effectiveness, independent of the various forms of queries it will handle. He admitted that such an assumption is open to question, and it is clearly incorrect in an experimental situation where major variables are being changed with the result that new systems are being formed. In such tests, the twin variables are necessary to see the changes that are taking place over the whole range of performance and even then need the additional environmental control of generality. It is difficult to understand the use of the term 'weak', since all composite measures can only present some compressed and simplified combination of the whole range of values shown by twin variable measures.

The composite measures can themselves be evaluated by recording their scale or range of values on the two twin variable plots. Any composite measure must indicate perfect retrieval in a situation of 100% recall at 100% precision at 0% fallout, and must indicate the worst retrieval in a situation of zero recall and zero precision at 100% fallout. Thus all composite measures have some scale of values between those two extremes, which can be plotted for visual examination on both recall/fallout and recall/precision plots.

Some of the measures proposed may be described as linear composite measures, when their values vary in some linear way if either the recall alters, or the precision (or fallout) alters. Perhaps the simplest composite measure suggested is the sum of the recall and precision ratios. Fig. 3.8P shows an example of this, using the simple sum of the recall and precision percentages, resulting in a range of values from 0 to 200. As can be seen, a performance of 70% recall at 10% precision would be given a value of 80, and be regarded as a better performance than 45% recall at 30% precision, or worse than a performance of 80% recall at 1% precision. The limitations of such a measure are fairly obvious, since a 70% recall at 10% precision will be rated the same as a performance as 10% recall at 70% precision or 40% recall at 40% precision, and many other different levels along the diagonal line. A simple weighting can alter the slope of the lines, e.g. if the recall ratio is weighted 1, and the precision ratio 2, the lines are more steeply positioned (Fig. 3.9P). The performance curves from Fig. 3.4P plotted on both tables are seen to have composite values which generally indicate the superior performance of search Y but, of course, the detailed differences at the cut-off points and the loss of maximum recall with search Y as against search X cannot be indicated by any composite measures.

Of measures of this type that have been suggested, J.D. Sinnett, in his thesis describing a test of role indicators, (Ref. 7), uses an effectiveness measure 'R', originally suggested by H. Borko, being $R = 100 \left(\frac{a}{a + c} - \frac{b}{a + b}\right)$ which is the recall ratio minus the noise factor (the complement of precision). The resulting values are positioned as 45 degree diagonals on a recall/precision plot similar to Fig. 3.8P but having the range of values from -100 to +100, with the centre diagonal being 0. A second measure, put forward by Western Reserve University, is the measure 'Effectiveness', being the sum of sensitivity and specificity (Ref. 8), and appears as straight lines on a plot which reverses recall/fallout. This is shown in Fig. 3.10P, which, it should be noted, is not a semi-log plot as are the previous examples of recall/fallout.

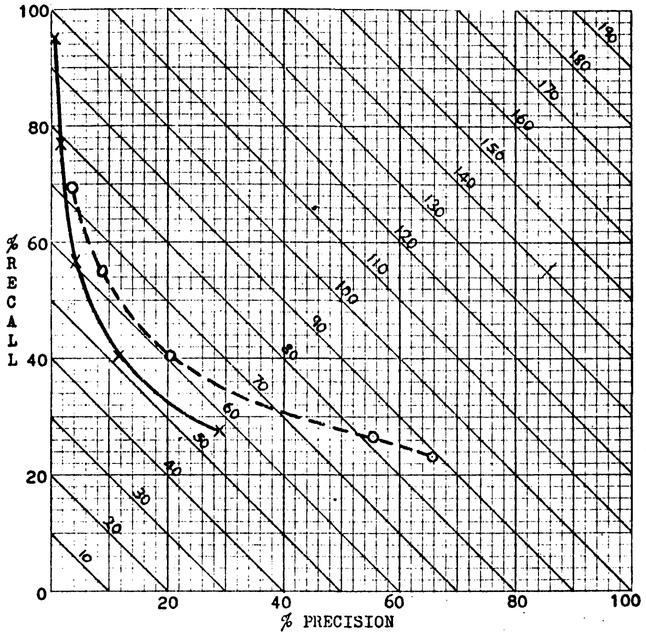


FIGURE 3. 8P PLOT OF RECALL AND PRECISION AS FIGURE 3.4P SHOWING THE 'RECALL + PRECISION' LINES

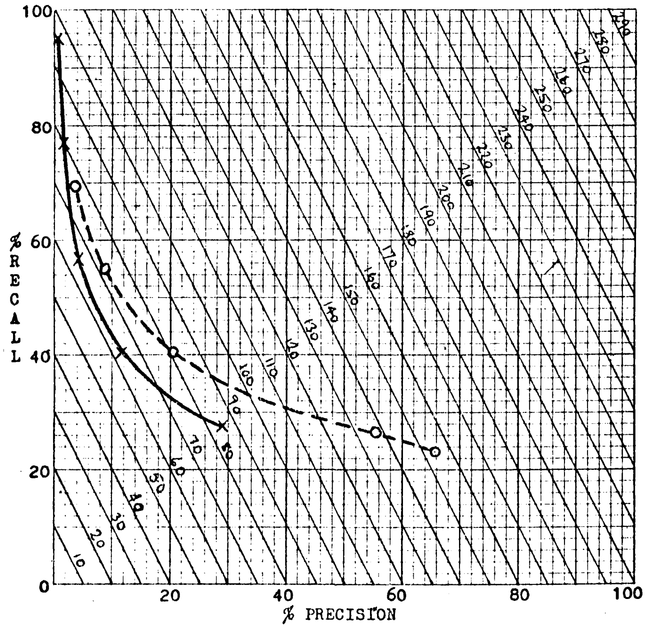


FIGURE 3. 9P PLOT OF RECALL AND PRECISION AS FIGURE 3.4P SHOWING THE 'RECALL + PRECISION X2' LINES

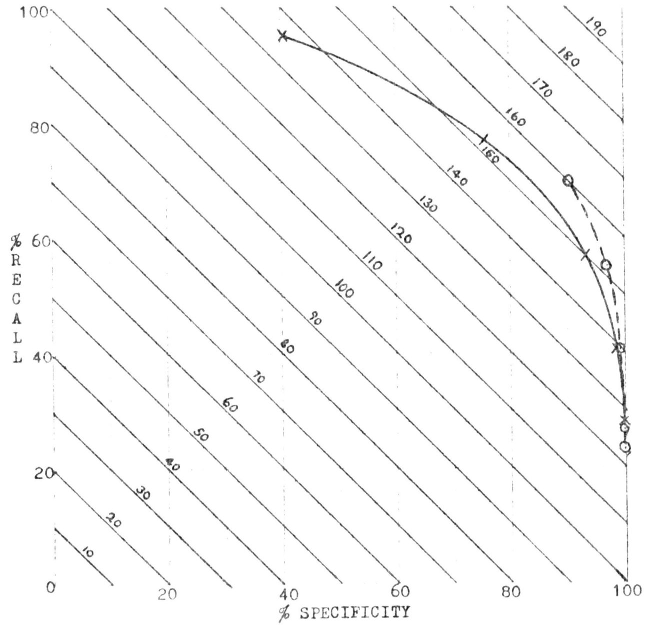


FIGURE 3.10P PLOT OF RECALL AND SPECIFICITY (ON A LINEAR SCALE) SHOWING THE 'EFFECTIVENESS' LINES

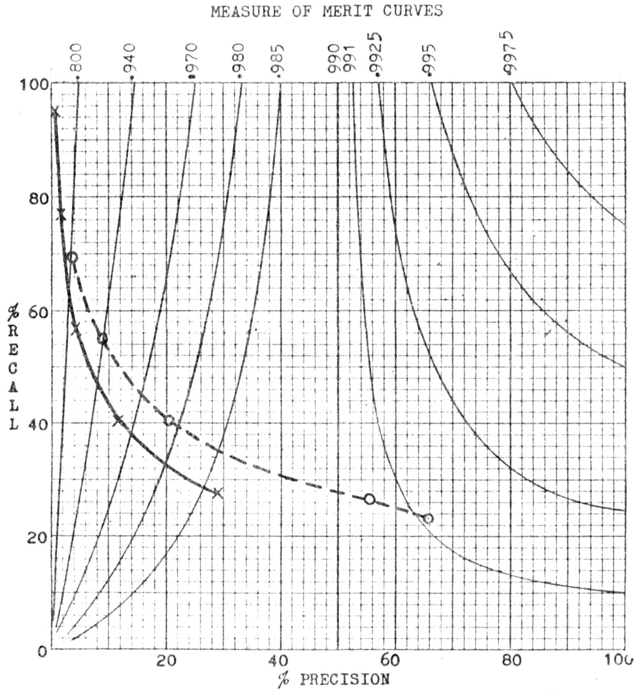


FIGURE 3.11P PLOT OF RECALL AND PRECISION AS FIGURE 3.4P SHOWING 'MEASURE OF MERIT' CURVES

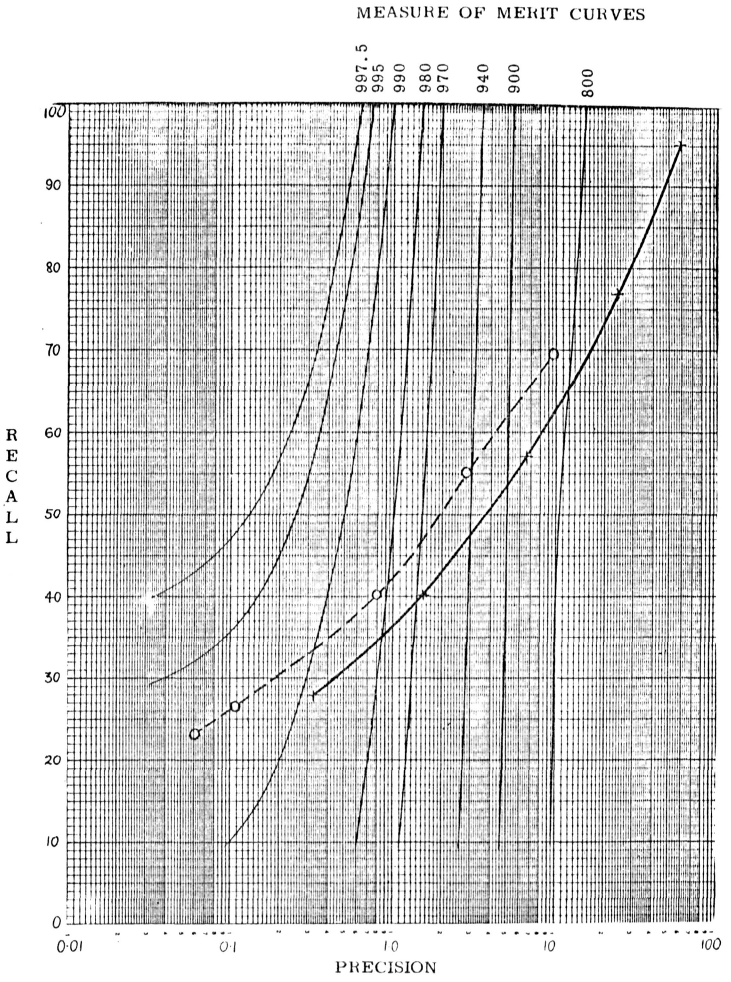


FIGURE 3.12P PLOT OF RECALL AND FALLOUT AS FIGURE 3.5P SHOWING 'MEASURE OF MERIT" CURVES

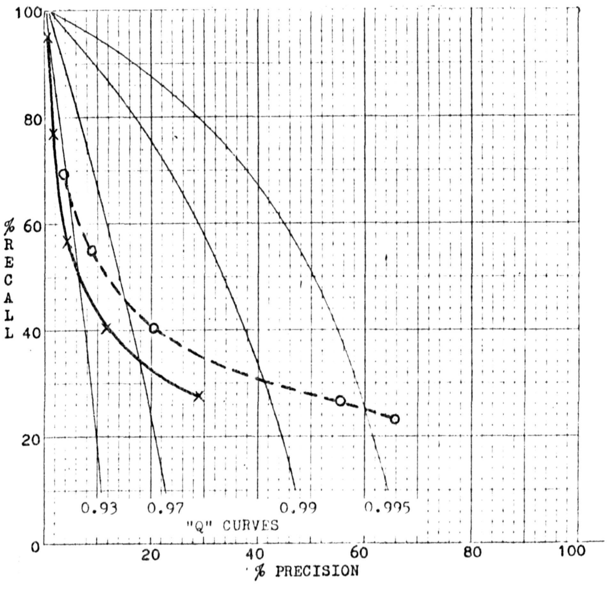


FIGURE 3.13P PLOT OF RECALL AND PRECISION AS FIGURE 3.4P SHOWING 'Q' CURVES

Other composite measures proposed can be described as non-linear composite measures, since their scale of values varies in a non-linear fashion when recall, precision, or fallout are varied, and the display of their values on the twin variable plots results in curves rather than straight lines. When a measure of this type includes d (non-relevant not retrieved) in its equation, the values and curves of the measure will be affected by the generality number. For Figs. 3.11 to 3.15 a generality of 5.0 is used in drawing the curves for the measures involved, since the performance results of searches X and Y that are plotted were obtained in a situation of that generality. The values of a composite measure of this type have been calculated in a manner similar to that adopted in making the two combined plots of recall, precision and fallout, Figs. 3.7P and 3.8P. In this case various sets of recall and fallout ratios, and also recall and precision ratios (at a generality of 5.0) were selected in advance and the resulting value of the composite measure calculated. This was done for different ratios to obtain curves of the measure that give a general indication of the range in its values.

The first of these non-linear composite measures which we consider is that proposed by J. Verhoeff and others, which is described as a 'Measure of Merit' (Ref. 9), with the basic equation:

$$M = a - b - c + d$$

This can also be written as $M = (a + d) - (b + c)$ which is really the sum of the 'successes' minus the sum of the 'failures'. The values are shown in the two twin variable plots, Figs. 3.11P and 3.12P, with the equations divided by 'N' to obtain a range of values between 0 and 1, and it can be seen how high values of the measure occur at high recall with high precision or, to say the same thing in a different way, high recall with low fallout. The measure was intended to be used with various weights associated with the four component values, and any of the composite measures being described could incorporate this if in a given situation a meaningful set of weights can be devised. One might, for instance, hypothesise 'cost values' of failing to retrieve a relevant document or retrieving a non-relevant document. Any such weighting would alter the position of the measure's curves on the plots.

A more complex version of this is the Q factor, which has been suggested by Farradane as suitable for use in retrieval tests. This is a statistical coefficient of association proposed by Yule (Ref.10). The formula is $Q = \dfrac{ad - bc}{ad + bc}$, which can be described as the product of the successes minus the product of the failures divided by the sum of the same two products. Figs. 3.13P and 3.14P show the two graphs with Q curves plotted, with the performance curves. It has not been shown that Q curves have any significance in retrieval tests, and there does not appear to be any reason why they should.

A measure put forward in discussion by Vickery at the NATO Advanced Study Institute on Evaluation, held at The Hague, July 1965, uses the values of a, b and c from the retrieval table. He suggested that the measure should reflect the ability of the system to maximise a relative to b and c, described as the selectivity of the system. The proposed measure F, uses a normalisation factor S, where S = a + b + c, and

$$F = \frac{100 \frac{a}{S}}{\frac{b}{S} + \frac{c}{S} + 1}$$

F varies from 0 to 100, and is plotted on a recall/precision plot in Fig. 3.15P. The curves are symmetrical about the diagonal from the bottom left corner to the top right corner, and alter in shape as they approach the top right side.

All the composite measures described have an apparently reasonable scale of values ranging from the case of worst performance to that of best possible performance, but none of these measures can show the very large differences that occur between these two points, in the different positions at which systems actually operate. The curves in Figs. 3.4P and 3.5P are indicators of retrieval performance when a component of a system is varied to give results over the largest possible operating range, but the composite measures can only reflect one, or sometimes two, points of such curves. It is unfortunate that, in examples investigated so far, the point on the curves which determines the highest value assigned to that test by a given composite measure is usually either the point of maximum recall, or of maximum precision, neither of which may be the best points to use. It is a reasonable conclusion that for experimental tests where changes of the variables in systems are examined, the composite measures so far proposed are inadequate, although for tests where a single cut-off point is chosen, or a single cut-off is applied to two systems in a comparable manner, some of the composite measures may be useful. In experimental tests it is suggested that an 'area measure' is required; a possible solution is put forward in Chapter 5.

Having examined the main suggested performance measures, it may be asked whether any theoretical objective methods are known which could be used to evaluate the proposed measures, or whether tests and experience of actual results will be the only arbiter.

The only theoretical basis suggested so far is the use of the 2 x 2 contingency table, as already mentioned. Although the retrieval situation obviously fits the case in the sense that the resulting values of a retrieval test perfectly fit the nine categories in the table, no reasons have been advanced to show that figures from retrieval tests can benefit from the statistical tests commonly used. The retrieval situation is very different from the simple statistical one. For example, a typical 2 x 2 table taken from a popular textbook on statistics by M.J. Moroney (Ref. 11, page 264)

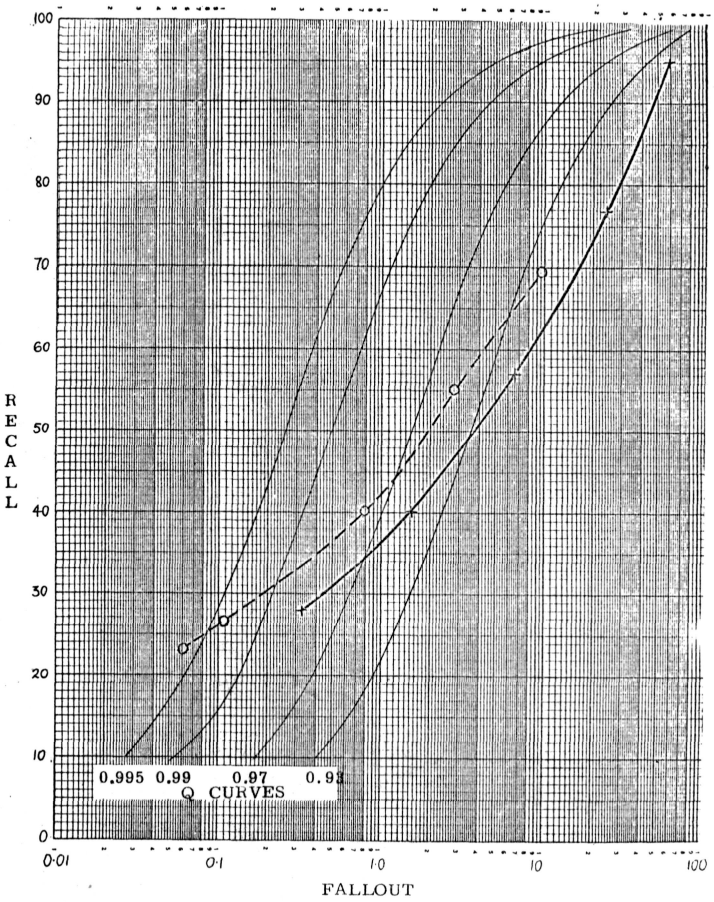


FIGURE 3.14P PLOT OF RECALL AND FALLOUT AS FIGURE 3.5P SHOWING 'Q' CURVES

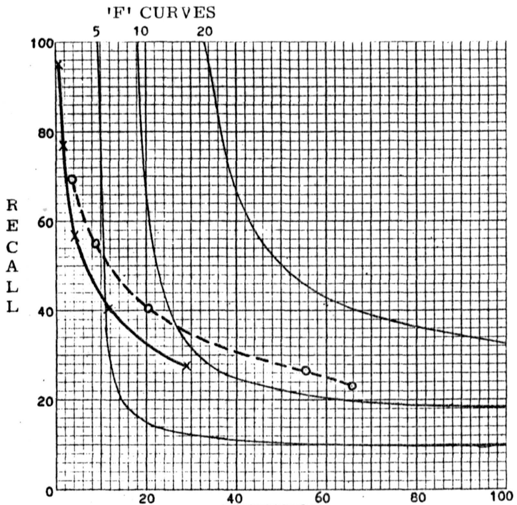


FIGURE 3.15P PLOT OF RECALL AND PRECISION AS FIGURE 3.4P SHOWING 'F' CURVES

gives data on a population of 77 people, showing the numbers that were both inoculated and not inoculated, and the numbers that were infected and not infected. The usual purpose of such a table is to ask a question of the kind, 'Is there really some degree of association between the events?', or in this particular case, 'Is the proportion of people that were not inoculated and became infected significantly different from the proportion of people that were inoculated and were infected?' In this situation, certain tests for the reality or existence of the association can be used (e.g. the chi square test), and other tests to determine the intensity of the association (e.g. the Q formula) can be applied. The form in which the question is posed, and the tests of the reality of association do not fit the retrieval case. Any question such as 'Is the proportion of relevant documents in the retrieved set significantly different from the proportion in the set not retrieved' does not make any sense in the retrieval situation. In the retrieval situation it is two sets of ratios from the table that are to be compared with one another by observing the relative changes in the ratios as conditions are changed. The actual comparative proportions do not need any test of significance. The tests of intensity of association do reflect the situation when the retrieval case is perfect, and when it is at its worst, and therefore provide one scale between the two extremes. But the deficiencies of the composite measures have been noted, and no assistance or confirmation of the twin variable measures being used seems to be given. The conclusion is that statistics does not help at all at this point.

Averaging sets of results

To present reliable results of performance, the figures from a set of questions must be averaged in some way. The size of the question set required in order to give reliable results will not be considered here, since there are many standard statistical tests to use in order to determine the significance level of a set of results. It is obvious that the results of individual questions will vary considerably, and some idea of the magnitude of this variation may be gained from Figs. 3.16P and 3.17P. In these plots of recall/precision, the individual results from a set of questions are plotted, where single term natural language indexing is being tested. Fig. 3.16P shows the points that result when any three out of a possible total of seven of the search terms in each of thirty-one questions are demanded in 'logical product' coordination. Fig. 3.17P shows points from thirty-five questions when the level of search terms demanded in coordination is varied from two to seven, and the scatter is quite wide, ranging from 11% recall at 1% precision in the bottom left corner, to 100% recall at 100% precision at the top right corner. However, a trend is clearly present down the left side of the plot and at the bottom right corner, with a tendency for results at a high coordination level to give high precision and low recall, and with lower coordination levels resulting in an inverse change. Two different methods of averaging these results, at each of the 'coordination levels', may be used.

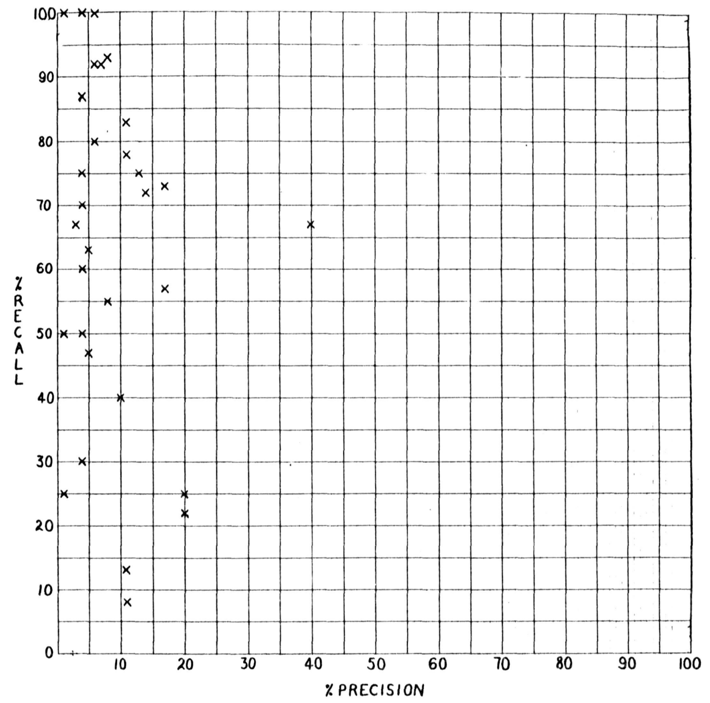


FIGURE 3.16P PLOT OF INDIVIDUAL RECALL AND PRECISION RATIOS OF 31 QUESTIONS SEARCHED AT A COORDINATION LEVEL OF 3 TERMS

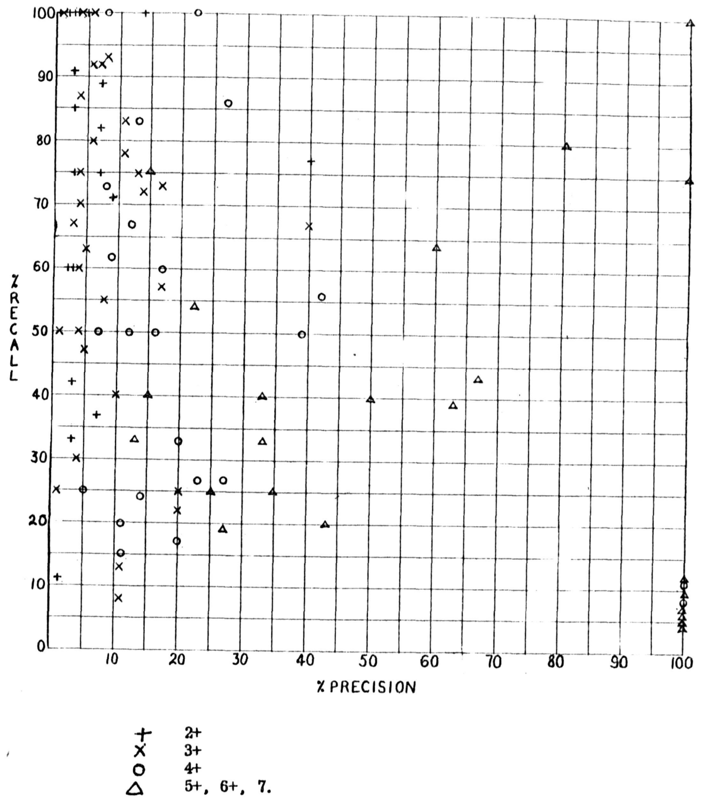


FIGURE 3.17P PLOT OF INDIVIDUAL RECALL AND PRECISION RATIOS OF 35 QUESTIONS SEARCHED AT COORDINATION LEVELS BETWEEN 2 AND 7 TERMS

The first method, as used in Cranfield I, involves obtaining total figures of the numbers of documents involved for the whole set of questions being used in the test, and then converting the one grand total into, say, recall and precision ratios. In the case of the 35 question set, a total of 287 relevant documents is sought; at a coordination level of 3+, 157 of the relevant documents are retrieved, together with 2,865 non-relevant documents. These totals are then used to calculate the ratios of:-

$$\text{Recall} = \frac{100a}{a + c} = \frac{157}{287} \times 100 = 54.7\%$$

$$\text{Precision} = \frac{100a}{a + b} = \frac{157}{157 + 2865} \times 100 = 5.2\%$$

$$\text{Fallout} = \frac{100b}{b + d} = \frac{2865}{(35 \times 1400) - 287} \times 100 = 5.9\%$$

These ratios are obtained for all of the seven possible coordination levels, and can then be plotted as points on a graph. While this procedure of averaging the numbers was used for presenting the results of the first Aslib-Cranfield Project and the Western Reserve University test, at the time of the latter test it was realised that this method results in certain searches affecting the final figures more than others. Non-typical questions, such as those which retrieve an exceptionally large number of non-relevant documents, will exert a disproportionate influence on the final figures, and, in the W.R.U. test, separate figures were given showing the change in performance when those questions that retrieved unusually large numbers of (mainly) non-relevant documents were deleted (Ref. 2, page 13).

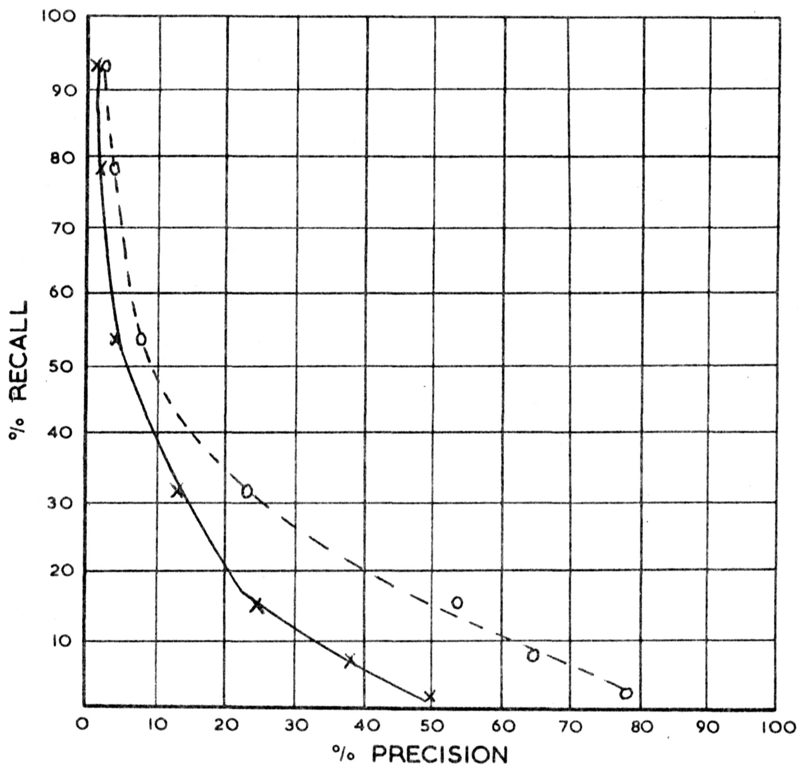
The second method of merging a set of results first converts the results of individual questions into recall, precision or fallout ratios and then obtains the final figures by using the average of the ratios of each question. In Fig. 3.18T are given the results of 35 questions which have been calculated in both ways, thus enabling a comparison of the 'average of numbers' and 'average of ratios' methods for these particular results. Recall, fallout and precision ratios for the two methods are compared in tabular form. It can be seen that there is no significant difference in the recall ratios between the two methods; at some coordination levels the average of ratios gives a slightly higher recall ratio, and at other levels the opposite is the case. The fallout values also show no significant difference. However, in the case of the precision ratios, it is clearly seen that the average of ratios gives a substantially higher figure for all coordination levels. Fig. 3.19P is a recall/precision plot of the two methods, where the 'better' curve results from averaging the ratios. As can be seen from the tables, a recall/fallout plot would have virtually overlapping

Coordination Level	Recall Ratio		Fallout Ratio		Precision Ratio	
	A	B	A	B	A	B
1+	93. 4%	93. 4%	48. 329%	48. 330%	1. 1%	1. 7%
2+	77. 3%	77. 3%	16. 500%	16. 502%	2. 7%	4. 2%
3+	54. 7%	55. 6%	6. 954%	6. 055%	5. 2%	8. 7%
4+	32. 8%	31. 2%	1. 540%	1. 538%	13. 5%	24. 4%
5+	16. 4%	14. 9%	0. 547%	0. 586%	25. 5%	54. 3%
6+	8. 0%	6. 9%	0. 381%	0. 377%	38. 3%	64. 2%
7	2. 8%	3. 3%	0. 192%	0. 190%	50. 0%	77. 8%

A Average of Numbers

B Average of Ratios

FIGURE 3. 18T COMPARISON OF RECALL, FALLOUT AND PRECISION RATIO WHEN TOTALLED BY (A) AVERAGING OF NUMBERS AND (B) AVERAGING THE RATIOS FOR 35 QUESTIONS.



— — — — AVERAGE OF THE RATIOS
———————— AVERAGE OF THE NUMBERS

FIGURE 3. 19P PLOT OF RECALL AND PRECISION RATIOS FROM FIGURE 3. 18T COMPARING TOTALLING BY AVERAGING THE NUMBERS AND AVERAGING THE RATIOS curves.

In the tests at Cranfield and in other tests where sufficient data has been available, the samples which have been processed by both methods have always shown this increase in precision with recall remaining much the same. However, we do not wish to be misquoted on this point and would emphasize that while it is probably true that the average of ratios will usually give a better performance figure, it would be wrong to assume that the proportional improvement would always be so pronounced as in the example shown.

An evaluation of the two methods which shows one method to be superior is not possible, since proponents of both methods can give good reasons for adopting one method in preference to the other. The theoretical cause of the discrepancy is the variation in the base from question to question: in the case of the recall ratio it is the number of relevant documents sought; in the precision ratio it is the total retrieved; and in the fallout ratio it is the total non-relevant. The average of numbers method weights the results of individual questions according to the base, and a larger base exerts a greater influence on the final result. The average of ratios completely ignores the base variation. In situations outside retrieval tests, where similar data has to be averaged, it is frequently advocated that the variation in base should be allowed for, and the average of numbers used (see, for instance Ref. 12, page 161). The difference in the results of the two methods is small except when the range and distribution of the variation in base becomes large, as is often the case with the precision ratio. However, both methods appear to be equally reasonable for use in retrieval situations, and the different results are really complementary viewpoints requiring careful interpretation.

A description of the different viewpoints represented by the two methods has been given by Salton (Ref. 13). He suggests that the average of ratios is 'a query-oriented viewpoint', and the average of numbers is a 'document-oriented viewpoint'; performance figures using the average of ratios indicate the performance of a single typical search question, typical that is of the set of questions used in the test. The use of average of numbers indicates the result of the whole set of questions, or indicates the success in performance of looking for a given set of relevant documents (287 in the example being used). This really ignores the actual individual questions involved, since one question with 287 relevant documents could in theory have the same result as 35 questions having in total 287 relevant documents. Thus the average of numbers gives an arithmetical mean value for a set of questions, and the average of ratios gives what approximates to a 'median' value which reflects the performance of a typical question.

Neither method appears to have any marked superiority over the other as a means of presenting results. However, the decision to use in this volume the average of numbers method was based on a most important practical advantage, namely the comparative ease of calculation. To have used the method of the average of ratios would have increased the calculations forty-fold; work that has taken hundreds of hours would have taken hundreds of weeks. The really important matter in any test is to know which method is being used and to use it consistently in all situations.

Method of totalling results

Apart from deciding on whether to use the average of ratios or the average of numbers, we were faced with the additional problem which is involved in totalling results of sets of searches where the co-ordination level cut-off is employed; an idea of what is involved in this problem can be seen in Appendix 4A. There are given the performance results, in actual figures, for the 221 questions (subset 3), being tested on Language I.1a (single terms, natural language and coordination), Exhaustivity 3, Search rule type A, Document Relevance 1 - 4 and searched on the 1400 document collection. The questions are arranged in numerical order, and for each question is given the total number of relevant documents in the whole collection, followed by the relevant and non-relevant documents actually retrieved at each coordination level. In the final column is given the sum of the total number of postings for the search terms; the total must, of course, equal the sum of the relevant and non-relevant documents at all coordination levels. The variations between the 221 questions that affect the problem of arriving at a single result of a single performance curve for the 221 questions can be seen in tabular form in Fig. 3.20T. Here the two characteristics of the 221 questions are listed, namely the numbers of terms initially selected from the search question and used as search terms (starting terms), and the number of retrieving terms, that is the maximum number of starting terms which, used in logical product coordination, may be put to the index and will still retrieve documents (whether relevant or non-relevant).

The table shows how, for this particular test, the starting terms ranged from 2 to 15, and the retrieving terms varied from 2 to 10. Within this 14 x 9 matrix the actual number of questions involved is recorded, so it can be seen, for example, that of the 35 questions having seven starting terms (column headed 7) only three of these questions could coordinate all seven terms and still retrieve some documents. The figures in the table refer only to the particular index language in use, and a different index language such as index language I.5 which includes synonyms, word endings and quasi-synonyms, would alter the distribution of the questions in relation to the retrieving terms, while any test involving a different basic index language (such as simple concepts as compared to single terms) would alter the starting term groups also.

Number of starting terms

Number of retrieving terms	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Totals
2	1	3	1	-	-	1	-	-	-	-	-	-	-	-	6
3		5	5	7	5	6	-	-	-	-	-	-	-	-	28
4			9	18	8	10	7	4	-	-	-	-	-	-	56
5				8	8	11	8	4	3	2	-	1	-	-	45
6					3	4	7	7	10	1	1	2	-	1	36
7						3	5	8	6	4	3	1	-	2	32
8							-	2	-	5	2	-	1	-	10
9								1	1	4	1	-	-	-	7
10									-	1	-	-	-	-	1
Totals	1	8	15	33	24	35	27	26	20	17	7	4	1	3	221

FIGURE 3.20T DISTRIBUTION OF THE 221 QUESTIONS BY STARTING TERMS AND RETRIEVING TERMS, IN ONE PARTICULAR TEST.

The table in Fig. 3.20T may be considered as showing how, in two respects, the 221 questions are a heterogeneous set of questions. Various subsets of the 221 can be picked to overcome the variations, and truly homogeneous subsets occupy each cell in the table, e.g. the five starting term group with four retrieving terms is the largest such subset, having a total of eighteen questions. A partially homogeneous subset, on the basis of one common characteristic only (either starting terms or retrieving terms), was the first to be examined in an attempt to find a method of totalling the whole set.

The subset of seven-starting-term questions was chosen and totalled by simply adding up each question at the seven possible coordination levels, resulting in seven totals. These totals are shown in Fig. 3.21T, and the recall precision percentages are recorded, these being calculated by using the average of numbers. The seven average recall and precision ratios are plotted in Fig. 3.21P, thus producing a performance curve for 35 questions, when the exhaustivity of search is altered by coordination levels. Since the characteristic of retrieving terms was ignored, not all the 35 questions provide results at all coordination levels, and, as was seen in Fig. 3.20T, one question is unable to retrieve any documents when more than two of the terms are demanded in coordination, and only three questions provide results at a coordination level of seven. The number of

Coordination Level	Documents Retrieved		Recall Ratio	Precision Ratio
	Rel.	Non-Rel.	a/a + c	b/b + d
1	268	23,681	93.4%	1.1%
2	221	8,086	77.0%	2.7%
3	157	2,869	54.7%	5.2%
4	94	600	32.8%	13.5%
5	47	137	16.4%	25.5%
6	23	37	8.0%	38.3%
7	8	8	2.8%	50.0%

Total relevant documents (a + c) = 287

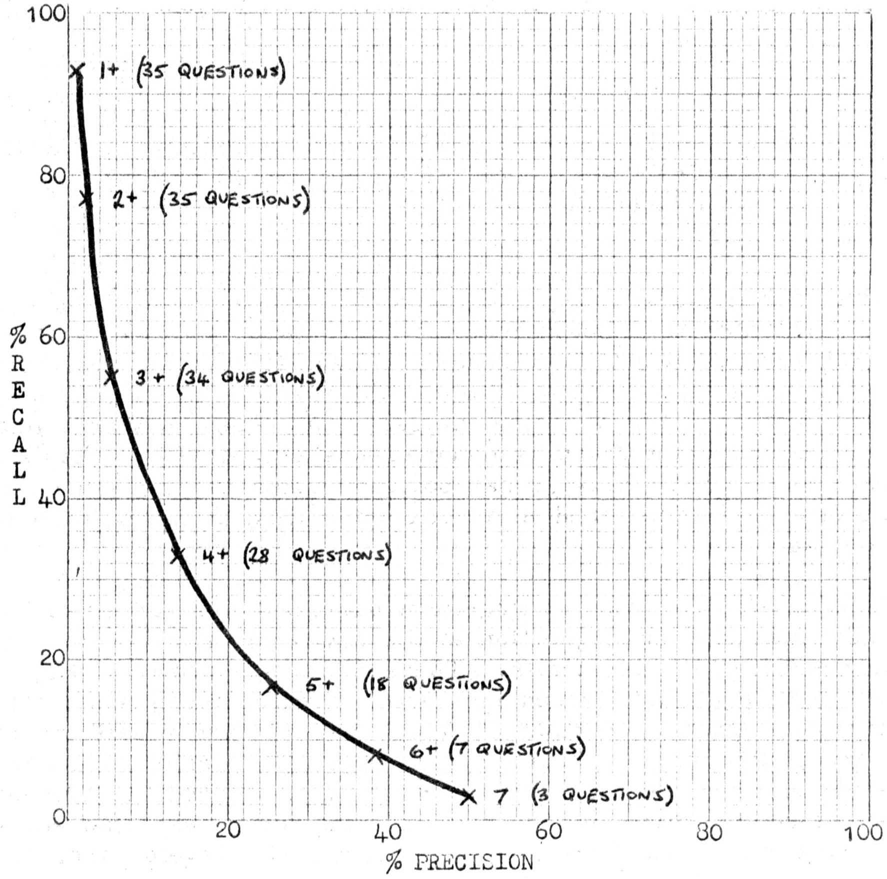


FIGURE 3.21TP TABLE AND PLOT FOR RESULTS OF 35 QUESTIONS WITH 7 STARTING TERMS TOTALLED BY COORDINATION LEVELS

questions that contribute results at each coordination level is recorded in Fig. 3.21P.

Although it has the bad characteristic in the reduced sample size at high coordination levels, it is suggested that totalling by starting term groups is a quite valid and satisfactory method.

On the other hand the totalling method using the retrieving term subset does not have this reduced sample size problem, and this was the next method to be investigated. The subset having five retrieving terms is obviously all composed of questions having five or more starting terms; as can be seen from Fig. 3.20T, there are 45 such questions and the results of this subset are given in Fig. 3.22TP. Here the low recall end of the curve does not sweep to high precision values, but stops at 26% precision at 15% recall. The main disadvantage of the retrieving terms subset totalling is that the composition of each subset alters whenever any language variable is introduced. This means that the generality number will be continually changing, and it therefore becomes more difficult to make comparisons.

While the matter of partly homogeneous sets presented little difficulty, the major problem lay in totalling the questions in the whole heterogeneous set of 221 questions; the results of our investigations on this point showed that no single method was conspicuously superior or satisfactory for all the different test situations. Many different methods were tried, but, with minor variations, they fell into six main groups. Summarised in Fig. 3.23T these are described in the following pages.

Method	Description
IA	Strict Coordination Levels.
IB	Strict Coordination Levels with adjustment for questions having no capability of retrieving.
2	Proportional Coordination Levels.
3	Maximum Starting Term Coordination Levels.
4	Maximum Retrieving Term Coordination Levels.
5	Recall Levels of Retrieving Term Groups.
6	Document Output Cutoff with ranked output derived from the coordination levels.

FIGURE 3.23T SUMMARY OF TOTALLING METHODS

Coordination Level	Documents Retrieved Rel.	Non-Rel.	Recall Ratio	Precision Ratio
1	302	29,898	94.7%	1.0%
2	257	10,917	80.6%	2.3%
3	191	3,292	59.9%	5.5%
4	119	899	37.3%	11.7%
5	47	132	14.7%	26.3%

Relevant documents (a + c) = 319

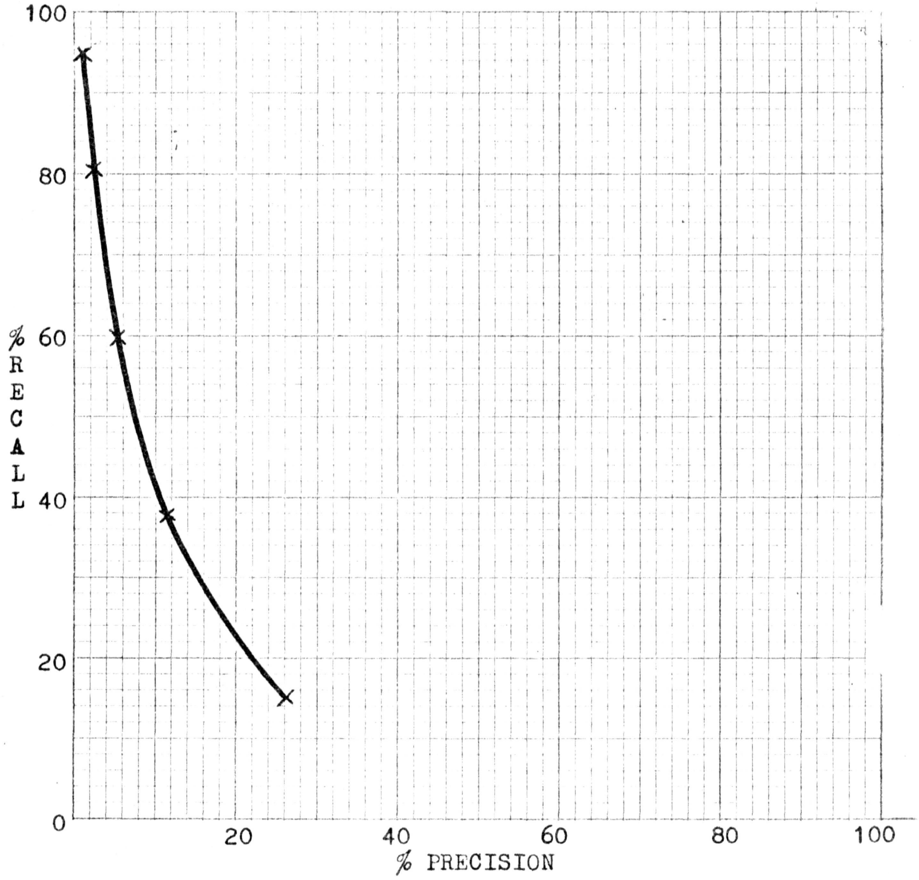


FIGURE 3.22TP TABLE AND PLOT FOR RESULTS OF 45 QUESTIONS WITH FIVE RETRIEVING TERMS TOTALLED BY COORDINATION LEVELS

Coordination Level	Recall Ratio	Precision Ratio
1	95. 0%	0. 9%
2	80. 7%	2. 2%
3	59. 5%	4. 1%
4	38. 1%	7. 6%
5	19. 7%	11. 6%
6	9. 7%	19. 0%
7	4. 7%	25. 5%
8	1. 4%	33. 8%
9	0. 5%	61. 5%
10	0. 1%	100. 0%

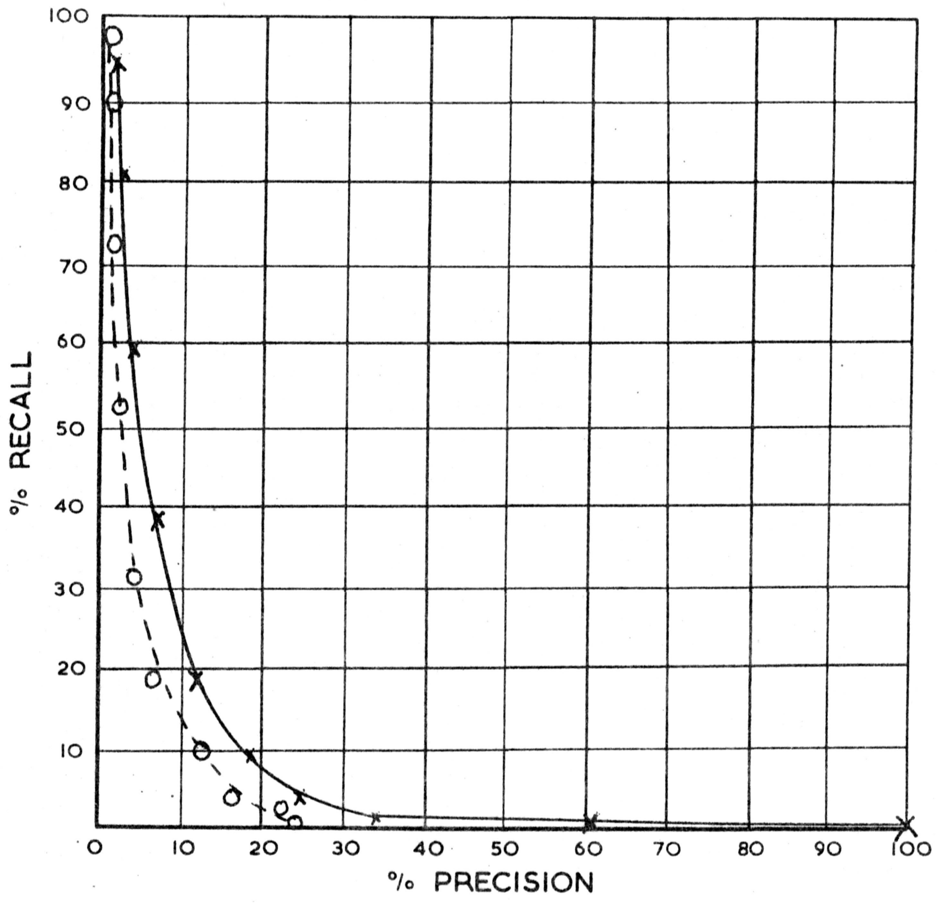


FIGURE 3.24 TP . TABLE AND PLOT FOR RESULTS OF 221 QUESTIONS TOTALLED BY METHOD 1A, COORDINATION LEVELS, FOR INDEX LANGUAGE I. 1. a. (INDEX LANGUAGE I. 6. a DOTTED LINE)

For Method 1, the questions were totalled in a similar manner to the starting term groups described above. This meant that for any given coordination level, (say, for example, four terms), the total results were obtained by adding the individual results for all the 221 questions, irrespective of the number of starting terms which each question had. Two variants of this strict coordination level totalling were considered. Method 1A involved totalling as described, and the resulting performance ratios are given in Fig. 3.24T, for Single Term Index Language I.1. The performance plot is given in Fig. 3.24P, with an additional curve of Language I.6 for comparison. In Method 1B, account is taken of the fact that at the higher coordination levels, many of the questions are not capable of contributing results, since the number of starting terms in the question is fewer than the coordination level. It is, for instance, quite impossible, at a coordination level of seven terms, to retrieve documents related to any of the questions which only have six, five, four, three or two starting terms. This effect increases, of course, with the coordination level. In this case, therefore, the recall ratio is calculated only for the questions that are capable of giving results. Fig. 3.25TP shows this, where it is seen that at a coordination level of 8+, only 704 relevant documents, i.e. less than half of the real total for this set of questions, are taken as the total of relevant documents being sought. This results in an increased recall ratio compared with Method 1A, but the precision ratio is not affected. A disadvantage of this method is that at each coordination level a change in generality occurs.

In Method 2, an attempt is made to allow for the fact that questions differ according to the number of starting terms. The strict coordination level of Method 1 can be faulted for equating, for example, the results of a five starting-term question searched at a coordination level of four terms, with the results of a ten starting-term question, also searched at four terms. The basic Method 2 can be described as 'totalling by proportional coordination levels', since it takes into account the potential range of coordination levels, which differs between questions. For example, a three starting-term question searched at a coordination level of two terms is demanding a match of two-thirds of the theoretical maximum, and in this method all questions having such a match would be included in the group. For a six starting-term question, for a nine starting-term question and for a twelve starting-term question, a two-thirds match would be four terms, six terms and eight terms respectively, although, for most other questions, no exact two-thirds match is possible. There are obviously many variations which are possible, but the example presented illustrates the use of this method when seven levels of match are chosen to obtain a total result.

There are obviously many ways in which this method could be applied; the example presented is where seven terms of match have been selected. Whatever the actual number of coordination levels in any particular question, the results are forced into the seven-term pattern. As can be seen from Figure 3.26T, this means that certain results are repeated, while for questions with more than seven starting-terms, certain results have to be omitted.

Coordination Level	Total Relevant	Relevant Retrieved	Recall Ratio	Precision Ratio
1	1590	1510	95. 0%	0. 9%
2	1590	1283	80. 7%	2. 2%
3	1583	946	59. 8%	4. 1%
4	1507	606	40. 2%	7. 6%
5	1401	314	22. 4%	11. 6%
6	1143	154	13. 5%	19. 0%
7	991	74	7. 5%	25. 5%
8	704	22	3. 1%	33. 8%
9	546	8	1. 5%	61. 5%
10	349	1	0. 3%	100. 0%

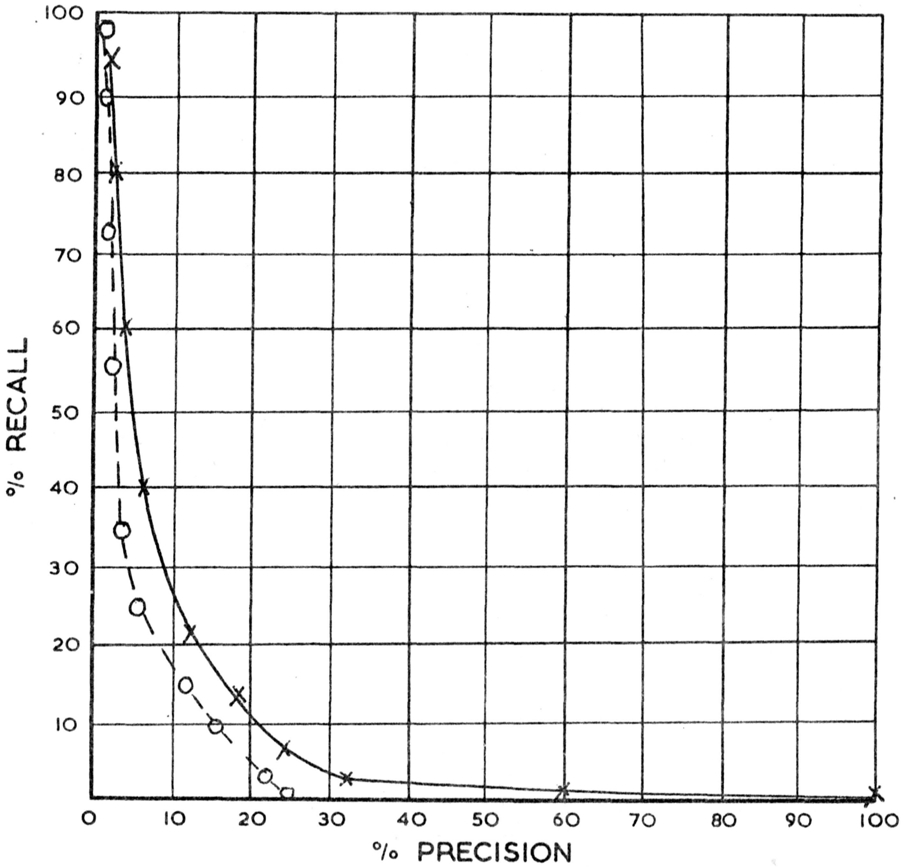


FIGURE 3. 25TP TABLE AND PLOT OF RESULTS FOR 221 QUESTIONS TOTALLED BY METHOD 1B, ADJUSTED COORDINATION LEVELS, FOR INDEX LANGUAGE I. 1. a. (INDEX LANGUAGE I. 6. a DOTTED LINE)

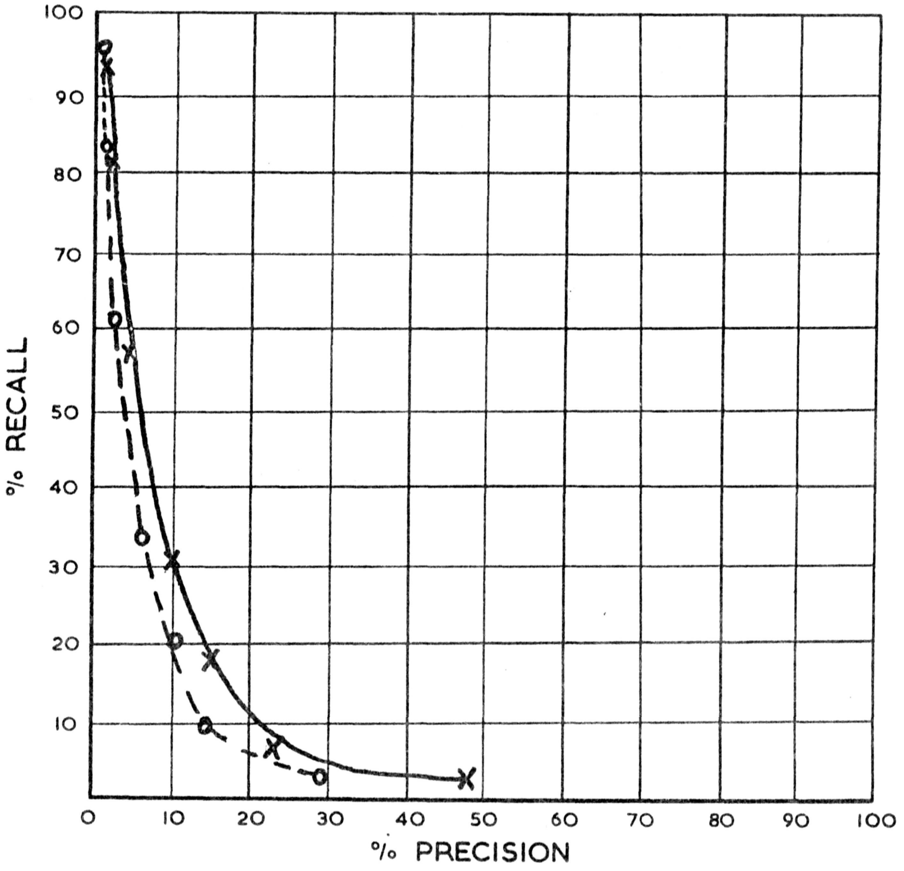
Starting Term Groups	Seven Proportional Coordination Levels						
	1/7	2/7	3/7	4/7	5/7	6/7	7/7
2/3	1	1	2	2	2	3	3
4	1	1	2	3	3	4	4
5	1	2	2	3	4	4	5
6	1	2	3	4	4	5	6
7	1	2	3	4	5	6	7
8	1	2	3	5	6	7	8
9	1	2	4	5	6	8	9
10	1	3	4	6	7	8	10
11	2	3	5	6	7	9	10
12/15	2	3	5	7	8	10	11

FIGURE 3.26T PROPORTIONAL PLACEMENT OF COORDINATION LEVELS IN STARTING TERM GROUPS FOR METHOD 2

Such a method is very arbitrary; some of the results for questions having less than seven starting terms had to be used more than once, whilst some of the results for questions having more than seven starting terms could not be used. The performance figures resulting from this method are given in Fig. 3.27TP.

For Method 3, described as 'maximum starting term coordination levels', the questions were totalled by grouping according to the maximum number of starting terms. Thus the three-starting-term questions searched at a level of 3 would be totalled with the four-starting-term questions searched at 4, with the five-starting-term questions searched at 5 and so on. A single coordination level is dropped off at a time, working from right to left in the diagram given in Table 3.28T. It can be seen that questions having only a small number of starting terms are soon reduced to a single term search; therefore the results at this level are maintained together with those questions that still have terms that can be dropped off, until all questions are being searched on a single term. Results by this method are given in Fig. 3.29TP.

Coordination Level	Recall Ratio	Precision Ratio
1/7	94. 4%	0. 9%
2/7	81. 0%	2. 4%
3/7	58. 7%	4. 2%
4/7	31. 3%	9. 9%
5/7	18. 7%	15. 5%
6/7	8. 4%	23. 5%
7/7	3. 2%	48. 1%



3. 27TP TABLE AND PLOT FOR RESULTS BY METHOD 2, PROPORTIONAL COORDINATION LEVELS, FOR INDEX LANGUAGE I. 1. a. (INDEX LANGUAGE I. 6. a DOTTED LINE)

	Twelve Coordination Levels											
Starting Term Groups	Minus 11	Minus 10	Minus 9	Minus 8	Minus 7	Minus 6	Minus 5	Minus 4	Minus 3	Minus 2	Minus 1	Maximum
2/3	1	1	1	1	1	1	1	1	1	1	2	3
4	1	1	1	1	1	1	1	1	1	2	3	4
5	1	1	1	1	1	1	1	1	2	3	4	5
6	1	1	1	1	1	1	1	2	3	4	5	6
7	1	1	1	1	1	1	2	3	4	5	6	7
8	1	1	1	1	1	2	3	4	5	6	7	8
9	1	1	1	1	2	3	4	5	6	7	8	9
10	1	1	1	2	3	4	5	6	7	8	9	10
11	1	1	2	3	4	5	6	7	8	9	10	11
12/15	1	2	3	4	5	6	7	8	9	10	11	12

FIGURE 3.28T GROUPINGS FOR MAXIMUM STARTING TERM COORDINATION LEVELS FOR METHOD 3

All methods discussed so far have results which, at higher coordination levels, are based on increasingly smaller sets of questions. Method 4 overcomes this particular drawback, by totalling all questions at the highest coordination levels that retrieve documents in every question. Known as the method of 'maximum retrieving term coordination levels', all questions are first aligned at the highest coordination level at which, in every question, at least one document is retrieved, irrespective of whether or not it is relevant. This level will vary from question to question, and by referring back to Fig. 3.20T, it can be seen that in the particular conditions of that test, some questions only began retrieving documents when several of their starting terms had been dropped off in coordination. For example, none of the twelve starting-term questions retrieved documents at a coordination level higher than nine. When the search results have been aligned by the coordination level at which each question gives a result, the figures are totalled, the coordination level in each question being relaxed one term at a time, until every question is reduced to a single term search. The results by this method are given in Fig3.30T, and the curve plotted in Fig. 3.30P. It will be noted that the lower end of the curve terminates at 17% recall and 22% precision; this point has been derived from the individual results

Coordination Level	Recall Ratio	Precision Ratio
Minus 11	94. 5%	0. 9%
Minus 10	94. 2%	1. 0%
Minus 9	93. 2%	1. 2%
Minus 8	91. 2%	1. 4%
Minus 7	86. 3%	1. 8%
Minus 6	78. 8%	2. 4%
Minus 5	67. 7%	3. 9%
Minus 4	54. 2%	6. 8%
Minus 3	38. 7%	8. 7%
Minus 2	25. 3%	9. 3%
Minus 1	11. 0%	14. 4%
Maximum	3. 1%	47. 6%

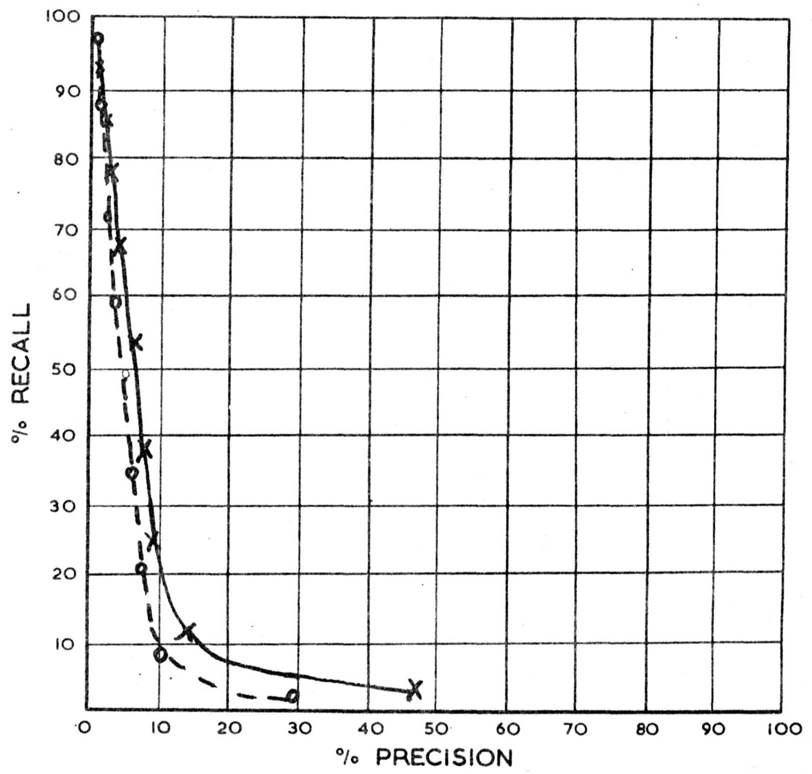


FIGURE 3. 29TP TABLE AND PLOT FOR RESULTS BY METHOD 3, MAXIMUM STARTING TERM COORDINATION LEVELS, FOR INDEX LANGUAGE I. 1. a.
(INDEX LANGUAGE I. 6. a. DOTTED LINE)

Coordination Level	Recall Ratio	Precision Ratio
Minus 9	94. 9%	0. 9%
Minus 8	94. 9%	1. 0%
Minus 7	94. 9%	1. 1%
Minus 6	94. 7%	1. 3%
Minus 5	93. 7%	1. 6%
Minus 4	88. 5%	2. 7%
Minus 3	78. 3%	4. 6%
Minus 2	62. 3%	6. 0%
Minus 1	40. 7%	8. 9%
Maximum	16. 9%	22. 2%

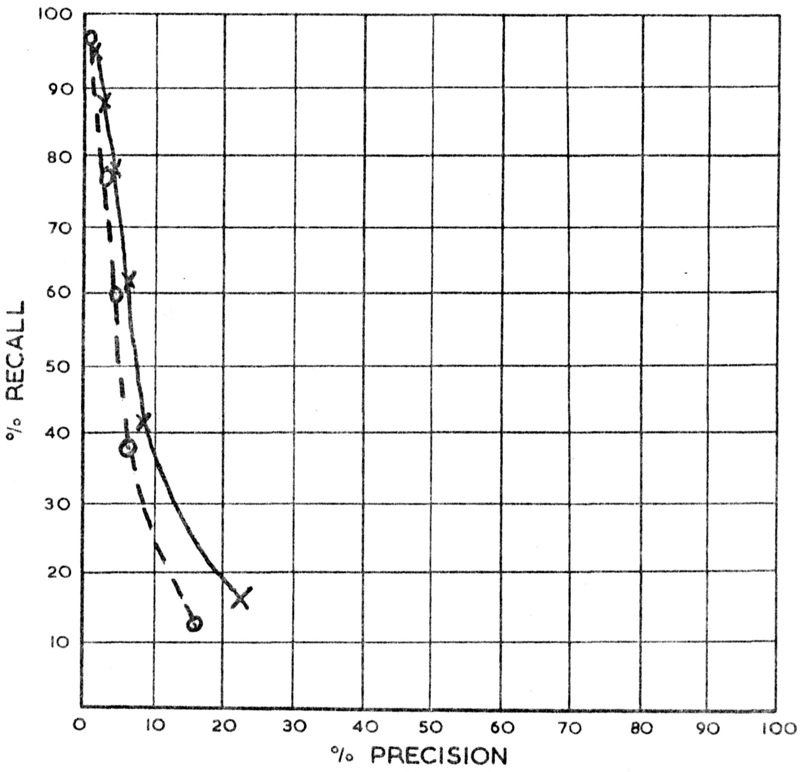


FIGURE 3. 30TP TABLE AND PLOT FOR RESULTS BY METHOD 4, MAXIMUM RETRIEVING TERM COORDINATION LEVELS, FOR INDEX LANGUAGE I. 1. a. (INDEX LANGUAGE I. 6. a DOTTED LINE)

of every one of the 221 questions.

Method 5 differs from all other methods described so far in being based on actual retrieval results obtained in testing. The method was generally known as 'recall levels', because a series of recall ratios is chosen in advance, and the performance results closest to the chosen recall levels are used to obtain the totals, irrespective of the coordination level of the search terms. Ideally this method should be applied to each individual question in a set, with the recall and precision ratios attained by each question being recorded when closest to 5% recall, then 10% recall, and so on. The calculations by Method 5 approximated to this by using the recall levels of the nine retrieving term groups. The recall ratios of these retrieving term groups were arranged by a set of twenty-one recall levels, being 0%, 5%, 10% etc. to 100%, and then the results in figures thus arranged were used to obtain twenty-one sets of recall and precision ratios. Fig. 3.31TP gives the table and plot of results, and the large number of performance points on the plot show a slight scatter through which the performance curve is drawn.

Method 6 was known as 'Document output cutoff method', and was based on quite different principles to those already discussed. To explain this method, it is first necessary to consider the effect of the 'conventional' search cutoff method used in the test. This, as has been explained, was based on the coordination level, which is to say that with, for instance, a six-term question, the search result would be recorded for a coordination of all six terms, then it would be recorded for a coordination of five terms, then for a coordination of four terms and so on. It was this method of search cutoff, with questions having a range of different potential coordination levels, that caused the problem in totalling the results of the whole set of questions, and Method 6, involving a document output cutoff, seemed to overcome this problem.

To apply this method, it was first necessary to obtain a ranked order of documents for every question, and, in our case, this had to be based on the coordination level cutoff results. A method of doing this was developed, but it entailed a considerable amount of effort.

The decision as to which method to use for presentation of the results was not easy to make and has probably involved more discussion, both amongst ourselves and with other people, than any other single aspect of the test. The necessity for the particular series of attempts to total the results was due to the problem created by the coordination level cutoff. It seems reasonable to assume that the final method discussed, the document output cutoff method would be most satisfactory since it eliminated the basic problem of totalling different sets of results but it appeared to involve more effort than could be afforded.

Recall Levels	Recall Ratio	Precision Ratio
100%	94. 4%	0. 9%
95%	93. 7%	1. 1%
90%	93. 5%	1. 2%
85%	82. 3%	2. 5%
80%	81. 4%	2. 7%
75%	77. 9%	3. 1%
70%	67. 8%	4. 2%
65%	62. 2%	5. 2%
60%	61. 7%	5. 4%
55%	54. 0%	7. 2%
50%	52. 3%	7. 4%
45%	47. 4%	8. 2%
40%	42. 6%	10. 2%
35%	27. 6%	15. 4%
30%	27. 5%	15. 3%
25%	22. 1%	17. 7%
20%	22. 1%	17. 7%
15%	16. 9%	22. 2%
10%	16. 9%	22. 2%
5%	16. 9%	22. 2%
0%	16. 9%	22. 2%

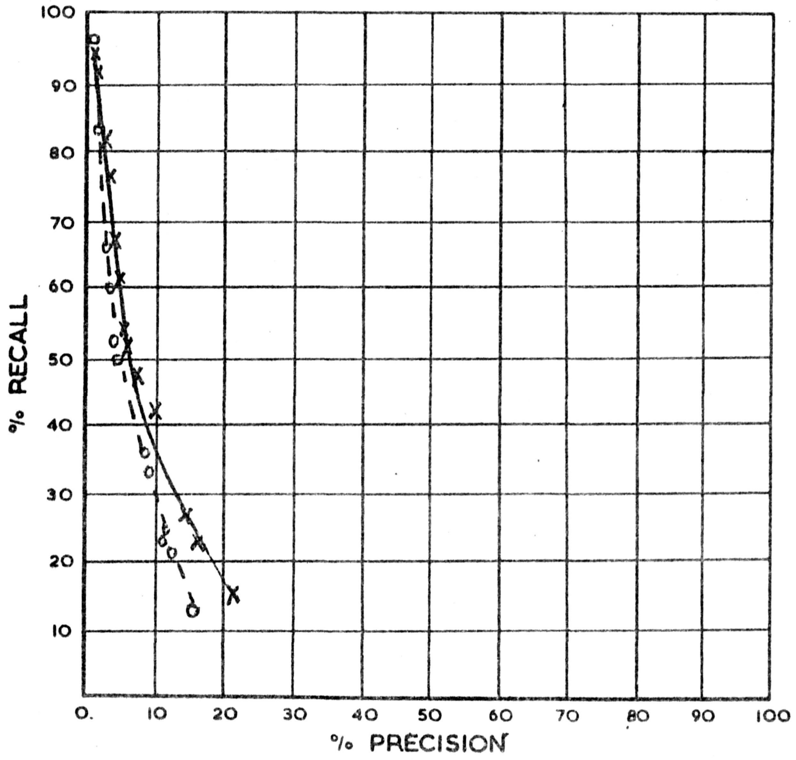


FIGURE 3. 31TP TABLE AND PLOT FOR RESULTS BY METHOD 5, RECALL LEVEL OF RETRIEVING TERM GROUPS, FOR INDEX LANGUAGE I. 1. a (INDEX LANGUAGE I. 6. a DOTTED LINE)

With all the methods that have been discussed and illustrated, each consistently showed Index Language I.1a to have a seemingly superior performance to Index Language I.6a. Whatever weakness there might have been in any of the methods, in no case were the results sufficiently distorted to mask this change in performance. In this situation, it again seemed most sensible to adopt a method which was relatively simple to apply, and Method 1A, using starting term coordination levels was therefore selected.

After this decision had been taken and the main sets of results had been prepared, a simpler method of obtaining a ranked output was found, and the majority of the results have been recalculated by the document output cutoff method. However, the decision to present the main results by Method 1 was not reversed, so the results obtained by this alternative Method 6 are presented separately in Chapter 5.

In view of the decision to use the starting term coordination level method, it is necessary to mention one further point. Using this method means that average results obtained at high coordination levels are based on an increasingly smaller number of questions in the set due to two reasons - firstly the variation in the number of starting terms, resulting in questions with a small number of starting terms never being capable of contributing results when the coordination level exceeds the number of starting terms. This has already been discussed in Chapter 2, where it was stated that this information was given in each table of test results in column z (see Fig. 2.15). The second reason was the variation in the number of terms that actually retrieve any documents, since the higher coordination levels in some questions demand a match that is too strong for any documents to be retrieved. Data on this point is presented in Column x which gives the total number of questions which actually retrieved any documents. As can be seen in Fig. 2.15, although z decreases at the higher coordination levels, x is smaller than z at all coordination levels of 4 or more. This was the normal experience, since there were usually some questions where the demand for a coordination of four terms would not retrieve a single document.

<u>The generality number</u> $\frac{1,000(a+c)}{N}$

To return to the matter of the generality number, it is now possible to consider this in more detail. It is known that, in situations where the generality numbers are different, varying performance figures will be obtained, even though the actual operational performance may be similar. In experimental tests such situations exist when the average numbers of documents relevant to the questions differ in two cases of identical file size or, vice versa, where the file sizes are the same but the numbers of relevant documents are different. A third situation is where both the numbers of the relevant documents and the file sizes are different.

As an example, two collections are hypothesised (see Fig. 3.32T), Collection A having 1000 documents and Collection B having 10,000 documents. In both collections there are assumed to be ten relevant documents for a given question, giving a generality number of 10 for Collection A and 1 for Collection B. It is hypothesised that the recall ratio is 50% and that the proportion of non-relevant retrieved to collection size remains the same. The fact that the proportion of non-relevant retrieved remains the same means that the fallout ratio will be 1.0%*, although the precision ratio changes from 33.3% in Collection A to 4.8% in Collection B, reflecting the decrease in the generality number. A recall/fallout plot would indicate an identical performance, concealing the information that in Collection A a fallout ratio of 1.0% means the retrieval of ten non-relevant documents and in Collection B it means the retrieval of one hundred non-relevant documents. On the other hand a plot of recall/precision would correctly indicate this change.

COLLECTION A 1000 DOCUMENTS

	Relevant	Non-Relevant		Generality 10
Retrieved	5	10	15	Recall 50%
Not Retrieved	5	980	985	Fallout 1.0%
	10	990	1,000	Precision 33.3%

COLLECTION B 10,000 DOCUMENTS

	Relevant	Non-Relevant		Generality 1
Retrieved	5	100	105	Recall 50%
Not Retrieved	5	9890	9895	Fallout 1.0%
	10	9990	10,000	Precision 4.8%

FIGURE 3.32T TWO SETS OF PERFORMANCE RESULTS WITH DIFFERENT GENERALITY NUMBERS AND CONSTANT RECALL AND FALLOUT RATIOS.

For a comparison of retrieval performance, it can be argued that the result revealed by the fallout ratios is more useful, since the change in precision ratio is solely due to the change in the environmental factor of the generality number. However, we have earlier stated our intention to present the main body of results with recall/precision plots, on the ground that these, in general, make a more useful and comprehensible

*This is correct to one decimal place; the actual figures are, respectively, 1.0101%, recurring and 1.001001% recurring.

presentation of performance. It is therefore necessary to make adjustments to the precision ratios in certain situations (which have been considered in Chapter 2) where sets of varying generality have to be compared. This is reasonably straightforward and is obtained by the following equation:-

$$P_A \text{ (Adjusted Precision Ratio)} = \frac{R_1 \times G_2}{(R_1 \times G_2) + F_1(1000 - G_2)}$$

where R_1 = Recall ratio obtained for a given system, in a situation of a known generality number
F_1 = Fallout ratio obtained for the given system, in a situation of a known generality number
G_2 = Generality number to which it is desired to alter the results, to obtain the adjusted precision

Thus two sets of performance figures obtained with systems of differing generality can be compared by adjusting the precision ratio of one case, so that it is based on the generality number of the other. If the example in Fig. 3.32T were to be corrected, and if it were decided to alter the result of Collection A to fit the generality of Collection B, then, from the equation given above,

$$P_A = \frac{.50 \times 1}{(.50 \times 1) + .01(1000 - 1)} = \frac{.50}{.50 + 9.99} = .048$$

The answer, expressed as a percentage is 4.8% and this result is clearly correct, with both cases now having an identical recall ratio, fallout ratio and precision ratio,

This however, is a simplified example, and in practice the matter is complicated by what at present seems to be the most difficult problem in performance comparison, namely the determination of the correct N. (the size of the collection). To consider this, an actual result is taken from a particular set of 42 questions that were searched on collections A and B where N equals 200 and 1400 documents respectively, the documents in collection A being a subset of the documents in collection B. The details are given in Fig. 3.33T, with the two sets of performance figures obtained in exactly the same conditions. While the precision ratio for collection A has increased with the increased generality number, yet there is also a significant difference in the fallout ratio.

SYSTEMS DATA

	Collection A	Collection B
No. of documents	200	1400
No. of questions	42	42
Total No. of relevant documents	198	198
Generality Number	23.6	3.4

PERFORMANCE AT COORDINATION LEVEL OF 3

	Collection A	Collection B
Relevant retrieved	132	132
Non-relevant retrieved	761	3,984
Recall Ratio	66.7%	66.7%
Precision Ratio	14.8%	3.2%
Fallout Ratio	9.278%	6.798%

FIGURE 3.33T SYSTEMS AND PERFORMANCE DATA FOR COMPARISON OF GENERALITY NUMBERS.

If the fallout in both collections were exactly the same, this would mean that the ratio of the change of the number of non-relevant retrieved (b) would be the same as the ratio of the change of the total non-relevant (b + d) i.e.

$$\frac{\text{b(Collection B)}}{\text{b(Collection A)}} = \frac{\text{(b + d)(Collection B)}}{\text{(b + d)(Collection A)}}$$

$$\therefore \quad \frac{\text{b(Collection B)/b(Collection A)}}{\text{(b + d)(Collection B)/(b + d)Collection A}} = 1$$

Bearing in mind that these figures represent the sum of a series of searches for 42 questions having 198 relevant documents, the result from Fig. 3.33T is, in fact,

$$\frac{\frac{3984}{761}}{\frac{(42 \text{ x } 1400) - 198}{(42 \text{ x } 200) - 198}} = \frac{5.2352}{7.1448} = 0.7327$$

It is therefore shown that b(non-relevant retrieved) has increased by a factor of 5.2352 while the total number of non-relevant documents (b + d) has increased by a factor of 7.1448. Proof of the accuracy of this can be shown by assuming that collection B had retrieved 7.1448 times as many non-relevant documents as collection A in which case it would have retrieved 761 x 7.1448 ≈ 5437 documents, as against the actual total of 3,984 documents.

The fallout ratio would now be $\frac{5437}{58602} = 9.278\%$.

This fallout is now identical with that of collection A in Fig. 3.33T; it should be noted however, that these figures would result in the precision ratio falling from 3.2% to 2.4%.

One has various options as to how to correct the precision ratio according to generality; it is possible to convert A to B (i.e. 23.6 to 3.4), B to A (i.e. 3.4 to 23.6) or to take a figure intermediate between A and B, such as 11. The effect of these three possible changes would result in the following figures:-

	Uncorrected Precision Ratio	Adjusted Precision Ratio			Fallout Ratio
		G = 3.4	G = 23.6	G = 11	
Collection A	14.8%	2.4%	14.8%	7.3%	9.278%
Collection B	3.2%	3.2%	19.0%	9.7%	6.798%

Whereas uncorrected precision ratio shows A to be superior, all adjusted precision ratios show B to be superior. To discover what is the factor which, in terms of the two collections, causes the difference in performance, Collection A will be taken as giving the expected result, and we will investigate the reasons why B should show the improved performance after precision ratio has been adjusted.

The problem is why, with collection B, fewer non-relevant documents are retrieved than expected. This can be explained by saying that there is more diversification in the indexing terms (and, therefore, presumably of the subject) of some of the documents in the larger file in relation to the search terms of the questions. The 42 questions in the test were all specifically on aerodynamics, as were all the 200 documents in collection A. Howevér, it is known that 257 of the documents in collection B were included in relation to questions dealing with the theory of aircraft structures; if it is assumed that these were never retrieved by any of the 42 questions on aerodynamics, then this would reduce N for collection B from 1400 to 1143, which is shown as B_1 in Fig. 3.34T, where the new generality number and fallout ratio are given. The fallout, at 8.333%, is now closer to, but still does not reach, the level for collection A.

It is therefore clear that if the performances are to be equated, it is necessary to hypothesise that in collection B there is a further subset of documents which are not retrieved by the questions. This number can be found by calculating the size of a hypothetical collection, B_2, which would result in an identical performance as collection A; the size of this collection, B_2, is calculated to be 1027 documents, which means that a further 116 documents must be deleted from collection B_1. As will be seen in Fig. 3.34T, the collections are now equated with the fallout ratio being, in both cases, 9.278%.

Collection	No. of Documents	Generality Number	Fallout Ratio	Collection	No. of Documents	Generality Number	Fallout Ratio
A	200	23.6	9.278%	B	1400	3.4	6.798%
				B_1	1143	4.1	8.333%
				B_2	1027	4.6	9.278%
A_1	271	17.4	6.798%				

FIGURE 3.34T CORRECTED COLLECTION SIZES TO FIT GENERALITY NUMBERS.

If instead of correcting collection B to collection A, the reverse step had been taken, then it can be seen that it would have meant adding 71 documents to collection A making A_1, which would then have a fallout of 6.798%, the same as the original collection B.

As a result of doing this, the precision ratio of collection A, can now be converted by the equation given earlier, and, since recall, fallout and generality are equal, the adjusted precision ratio must be 3.2% as for collection B.

While the above may seem to be somewhat involved, it is, in fact, a simplification of the real situation in that 42 questions have been taken as a block. A more detailed analysis would require that each question should be treated separately. Then, again, the analysis has been done in a single fixed situation, namely a certain index language at a certain level of coordination, and clearly it could be repeated over many hundreds of situations of a similar type. However the implications of such analysis are far-reaching, going beyond the scope of this chapter, so they will be considered later in this report.

In addition to explaining the performance measures adopted in this report, this chapter has also attempted to cover, albeit in a non-exhaustive manner, the main considerations regarding their use and effect. For ourselves, we feel that it is foolish, at the present stage of development, to be dogmatic on this subject. Wherever it has been necessary to make a choice between different methods, in most cases the decision has been taken for reasons which could be considered peculiar to this project. Other experimenters may well find that different measures better suit their purpose; hopefully, in this survey the relationship between different measures has now been established, and so long as complete sets of figures are given in reporting test results, there should be no serious difficulty in converting from one set of measures to another. Ultimately, one assumes that something approaching general agreement will be reached on the measures to be used. All that we would claim is that the measures used in this report appear to be as good as any others so far proposed.