SECTION 1 MEASURES OF RETRIEVAL PERFORMANCE

INTRODUCTION

Problems of performance measurement are considered from the viewpoint of experimental tests, that is pure research investigations carried out in fixed environments in which the components of systems are varied in a strictly controlled manner. In operational tests, that is in tests of real systems in their own environments, performance measurement does not normally constitute a problem, since in such tests the main objective is an analysis of the operating characteristics of every part of the system, and any measurement of retrieval performance is usually straightforward. In experimental tests however, it is the basic components of systems, such as index languages, search rules, etc., that are being investigated, and since the main objective is the comparative evaluation of such variables, the use of measures accurately to reflect changes in retrieval performance is essential. The treatment of the subject in this note does not explore all possible theoretical considerations, and does not make involved mathematical excursions, but gives the main advantages and disadvantages of each measure with practical examples taken from some of the results being obtained in the present Aslib-Cranfield Project.

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

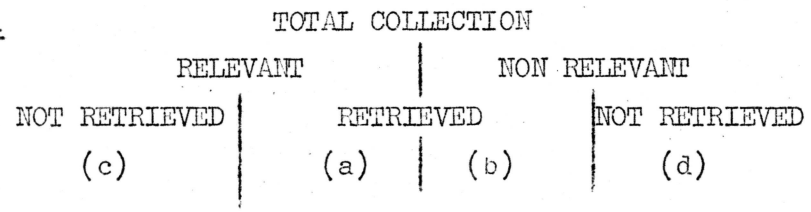
1. A document collection of known size to be used in the test;

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, which usually gives 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. 1, page 174) by the following table:-

FIG. 1



The most frequent case where the search of a question retrieves documents which are both relevant and non relevant is illustrated, and the resulting four categories of documents labelled (c), (a), (b), and (d) are called by Vickery: Missed, Hit, Wasted and Dodged respectively. It is the numbers of

documents that fall into these four categories as the result of a search that indicate the retrieval performance of the system, and various different equations or measures have been proposed to quantify the performance achieved. Before considering some of these proposed measures two topics must be introduced here, although they are treated in greater detail later in the note.

The first concerns the theoretical way in which the parameters or values in Fig. 1 are to be treated. It is more usual to present the categories in Fig. 1 in the form of a 2 x 2 contingency table, as follows:-

FIG. 2

	RELEVANT	NON RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d (TOTAL COLLECTION)

This table displays the different categories more clearly, showing the four single values together with the five associated totals, and this notation will be used in this note. 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 at the moment. The use of the table is purely for convenience at this stage, since it may be that retrieval tests represent an entirely different situation to which conventional statistics do not apply.

The second topic concerns a distinction between two different test situations, where retrieval results alter in different ways. The purpose of using measures of retrieval performance is primarily to enable some comparison to be made, either a comparison of several sets of results obtained in different conditions, or the comparison of a single result with some theoretically possible perfect result. The latter is quite feasible, since a perfect result is the retrieval of all the relevant documents with none of the non relevant ones. But in the former case a reliable comparison can only be made when it is known exactly what variables are altered in the different situations, and two ways in which different variables alter the retrieval results are considered.

1. Some variables alter the values a, b, c, d, a+b, and c+d in Fig. 2. The total retrieved (a+b) and total not retrieved (c+d) can remain unaltered, while the proportions of Hit, Wasted, Missed, and Dodged (a,b,c,d) change; or the totals of retrieved and not retrieved may change proportionally, so affecting all six values. In a test situation of this type, the collection size, total relevant (a+c), and total non relevant (b+d) do not alter at all. When the proportion of retrieved and not retrieved changes this is caused by a change in the 'Cutoff Point' applied. This is the place in a search where the rules do not allow any further documents to be examined,

and so the search is stopped and a record made of the documents retrieved, both relevant and non relevant, which determines all the values in Fig. 2. Some of the rules for establishing a cutoff cause problems in calculating test results, and the whole problem is considered in section 2 . It may be noted here that experimental computer systems, such as the SMART system of G. Salton at Harvard University, which always retrieve the total collection in response to a search, cannot use the performance measures described here unless some cutoff is applied. In the case of the SMART system, the collection is ranked in an order of decreasing similarity with the search question, and special measures of performance have been developed at Harvard to meet this case, as described on page 14.

2. Other variables alter the values of a+c, b+d and a+b+c+d. If the decision as to what is relevant is altered then the first two values change, and if the collection size is changed other values in the table may change. Although such changes occur very rarely in retrieval system tests it is necessary to consider these rarer cases, as experimental tests can involve this. Both types of change alter the number of relevant documents in relation to the collection size and such a variation can occur in a fixed test situation if different sets or subsets of questions are used, or if certain totalling procedures are adopted. It is convenient to express this variation as a parameter, and C.W. Cleverdon has suggested 'Generality', with Generality Ratio = $\frac{1000(a+c)}{a+b+c+d}$, that is, 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 a generality ratio is 5, this means that five relevant documents are found in every thousand documents in the collection, whatever the actual size is. Therefore the only significance of a change in either the relevance decisions or the collection size as far as retrieval performance is concerned is that in both cases it is the generality ratio that alters.

TYPES OF PERFORMANCE MEASURES

The major measures of performance which have been used in tests, or have been proposed, will be described and examined. The single measures derived from one row or one column of Fig. 2 will be examined first, before looking at the combined measures.

Single Measures

These measures fall into two groups, with the purpose of measuring two different things, the first being the ability to present relevant documents.

The success of a search in retrieving the relevant documents is usually measured by the 'Recall Ratio', defined as $\frac{a}{a+c}$. Perry (in ref. 2)

called it the Recall Factor, and Cleverdon (ref. 3) uses $\frac{100a}{a+c}$ to express the fraction as a percentage. It is an unfortunate proliferation of terminology that the ratio 'Specificity' (Western Reserve, ref. 4) is another name for the same thing. The ratio which is complementary to recall in Fig. 2 is $\frac{c}{a+c}$, and it clearly gives no different information but just gives the proportion of relevant not retrieved, rather than the proportion that is retrieved. In some notes by R.A. Fairthorne (ref. 5) this is defined as the Snobbery Ratio. No other measures which reflect the retrieval of relevant documents have been suggested, and the recall ratio has been used in practically all tests carried out so far.

The second group of single measures reflects the ability to withhold non relevant documents.

Two different measures have been suggested to reflect the success of a search in not retrieving the non relevant documents. The first proposed was the Pertinency Factor, defined as $\frac{a}{a+b}$ (Perry, ref. 2), which is the number of relevant documents retrieved as a proportion of the total number of documents retrieved (both relevant and non relevant). It is more usually known as the Relevance Ratio or Precision Ratio, the latter term being the one now agreed on by several groups. The complementary ratio, $\frac{b}{a+b}$, is called the Noise factor by Perry (ref. 2). The other suggested ratio is $\frac{b}{b+d}$, which is the number of non relevant documents retrieved as a proportion of the total number of non relevant documents in the collection. The actual values of 'b' were used in a combined measure by D.R. Swanson (ref. 6), and the ratio has been suggested by several people, probably first by J.A. Swets (ref. 7). No name has been given to this ratio, so a suggestion of 'Fallout Ratio' made by C.W. Cleverdon will be adopted here. The complementary ratio, $\frac{d}{b+d}$, is called Specificity (Western Reserve, ref. 4).

These two measures each use a different 'slice' of Fig. 2, with precision using the top row and fallout the middle column. The relative merits of precision and fallout will be examined in the next sub-section on combined measures.

The use of one of these single measures from one of the two groups only, either a measure reflecting the retrieval of the relevant items, or one reflecting the retrieval of the non relevant items, is clearly inadequate to fully reflect the retrieval performance of a system. Even if it is known that a search has a recall ratio of 95%, this does not indicate a good performance, since the search might be retrieving three-quarters of the collection in order to get this recall, with a precision ratio of less than 1%. The reverse is true: although a precision ratio of 80% indicates a strong suppression of the non relevant documents, if the recall is only 5% then performance is far from perfect. Although in different situations either high recall or high precision may be paramount, the real picture of retrieval performance can only be seen when one of each of the single measures are put together in some way, to reflect both the retrieval of the relevant and retrieval of the non relevant.

Combined Measures

Many different combinations of the two types of single measure have been proposed, but the combined ones fall into two groups: twin variable measures and composite measures.

Twin Variable Measures.

For these measures 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.

A comparison of the recall and precision ratios has been used in quite a number of tests. For example, in the Aslib-Cranfield test of the index of metallurgical literature of Western Reserve University, it was found that the W.R.U. index was operating at 75.8% Recall with 17.7% precision, in one of the conditions used in the test (ref. 8, page 15). The individual recall ratio is quite a good figure, and the individual precision ratio seems a bad result, but a high figure for recall together with a low figure for precision is an expected result for document retrieval systems. The accumulating experimental data strongly supports Cleverdon's postulate of an inevitable inverse relationship between recall and precision ratios (but see ref. 9, page 5) for the conditions in which this applies), and this fact can be demonstrated more clearly in an experimental test, rather than a test of an operational system such as at W.R.U. In an experimental test such as the present Aslib-Cranfield Project, the strict laboratory control of the variables in a system enables one variable to be altered at a time to give many sets of different performance ratios. For example, a set of figures from the present test is shown in table 1. The one variable being altered is the coordination level, or number of search terms demanded to match with terms in the document indexing. This has the effect of varying the exhaustivity and specificity of the search, and at each different level a cutoff is applied in the search to obtain the recall and precision ratios shown. It is seen that a demand of five search terms (in a logical product) gives a performance of 27.9% recall with 29.3% precision; four terms gives 40.4% recall with 11.7% precision; and so on, down to a single term search giving 95.2% recall with 0.8% precision. Such results may be displayed in tabular form as in table 1, or they may be put on a graph where recall is plotted against precision. Table 2 shows such a plot, with the five performance points connected to make a recall : precision curve. The joining of the points by a curve seems a reasonable procedure, since the alteration of the search rule demanding different coordination levels clearly results in an inverse change in the recall and precision ratios.

A further step in testing variables is to repeat the same search procedure at the five different levels, but altering one other variable in

the system. Table 3 adds another set of figures to table 1; the first result was obtained when the search rules allowed any combination of terms to be accepted, but in the second case more intellect is applied to the search rules, and only certain sensible combinations of terms are now accepted. The difference between the searches S and T can be seen in the table, but is seen better by table 4, where the two performance curves are plotted together. The intelligent search gives a generally 'better' performance, since the curve is positioned nearer the point of perfect retrieval (the top right corner), but it should be noted that the top end of the curve for search T does not reach the same maximum recall but stops at 69.4% compared to 95.2% in search S. Also, each coordination level in search T shows a worse recall ratio but better precision ratio than search S. Many more similar results obtained by varying other types of variables are being published.

No evaluation of the advantages and disadvantages of using this comparison of the twin variables of recall and precision will be made until the other combined measure has been described.

A comparison of the recall ratio with the fallout ratio can be made in exactly the same manner. The corresponding performance in the W.R.U. test is 75.8% recall with 1.3% fallout. The precision figures of table 3 are changed to fallout ratios and shown in table 5, and it is apparent that the range of values in the fallout ratio are concentrated at the low end, and some figures will be less than 1%. For this reason, when the plot of recall against fallout is presented in table 6, a log scale is used to spread out the fallout values. It will be seen that the range of fallout values is 'reversed' compared with precision, since now a 100% value indicates the worst performance, and a low value a good performance. The ratio that complements fallout, namely specificity, does compare with precision in this way, but the fallout ratio is preferred in this note to avoid visual confusion, since a plot of recall : specificity looks similar to a plot of recall : precision, (see table 7).

Looking at the plot of the two performance curves again, the recall : fallout plot is very similar to the recall : precision one, with a similar improvement seen in the intellectual search, and the same drop in recall at all points, with a corresponding decrease in fallout at each point. The plot of recall : specificity enables closer visual comparison, but the log scale obscures any obvious difference. In fact the current discussion on the relative merits of the two plots is in most cases concerned with trivial differences, as will be shown.

In making a comparison of the two plots, some practical points concerning comprehension and presentation will be made, before theoretical accuracy is examined.

The comprehension of the plots - the ease of understanding quickly just what is being plotted - is important. In the case of recall : precision a result of 50% recall with 10% precision may be quickly grasped

as indicating that half the relevance documents were found in the search, and one tenth of those retrieved were relevant. This is a correct interpretation, whatever the collection size or number of relevant sought happened to be in actual figures. A result of 50% recall with 2.0% fallout can be interpreted as retrieving half the relevant documents together with 2.0% of the collection (or strictly the non relevant in the collection). Without the actual collection size the fallout ratio does not mean much, and does not show how much non relevant material had to be tolerated in the retrieved set. This may be aleviated in a given situation if the actual numbers of non relevant documents are recorded on the x axis together with the fallout ratios. However, because the precision ratio gives an easily grasped indication of the contents of the retrieved set, its use may be preferred to the fallout ratio.

The presentation of the ratios on a plot in graphical form, with recall on the y axis, is the only way to fully see the retrieval performance when variables are being altered. The precision ratio may be plotted on a linear scale, but the low precision values around 0% to 5% represent large changes in actual figures and might be better plotted on a log scale. However this is not normally necessary since such low values indicate such a bad performance that accuracy is rarely needed here. High values of precision are quite satisfactory, except in cases where factors quite unconnected with performance measures cause problems, such as certain methods of totalling sets of questions (see section 2). The fallout ratio cannot usually be plotted on a linear scale because typical values do not give an even variation in percentage value, with the minimum value possible depending on the collection size being tested. A range of 0.01% to 100% will cover most situations, and with a log scale in use is satisfactory over most parts of the range, except that changes at low fallout may be distorted on the plot a little. The two different plots are really complementary as far as presentation goes, as recall : precision shows the low recall at high precision area a little better, and recall : fallout shows the high recall at high fallout a little better, although the satisfactory use of a linear scale with recall : precision is an advantage.

In comparing the two plots for theoretical accuracy in indicating retrieval performance, different test situations must be taken into account. If a test of an operational system is being made, which will result in one performance result such as the W.R.U. figure, either set of measures may be used. It is really in the controlled environment of experimental testing that accuracy becomes important since the object is to draw certain conclusions about the effect of different variables. A comparison of the two plots will be made for the two types of variables given on page 2.

In cases where the values of a, b, c, and d alter, and probably the cutoff point as well, it is possible to show both the precision and fallout values on one plot. Taking the results of searches S and T with the five coordination levels in tables 4 and 6, they are combined into one plot in table 8. Here the plot is basically a recall : fallout one, but the

precision values are shown by 'Precision curves' which sweep down to the bottom left corner. The position of these precision curves can be determined when recall and fallout is known, since if values for recall and fallout are obtained in a test the corresponding precision ratios can be simply calculated. In fact the position of the precision lines can be calculated in advance for any test situation, provided that the generality ratio is known and is always unaltered. This means that, given the fallout values, the correct precision values can be just read off the plot. Some visual idea of the difference in the plots may be gained if the precision lines are 'straightened out' so bending the performance curves back to the position they held on the precision plot. Table 9 shows the reverse of this, with fallout plotted ascurves on a recall : precision plot. However, assuming that a choice between plotting recall : precision or recall : fallout has to be made, the chief objections against the use of precision will be examined first.

It is stated that a plot of recall : precision is not a valid comparison from which reliable deductions can be made because both ratios contain 'a' (relevant retrieved) in them, and that in plotting $\frac{a}{a+c} : \frac{a}{a+b}$ all the a's cancel out, in a sense, leaving the real factors being plotted as c : b.

It has never been suggested that a plot of recall : precision is a comparison of completely independent variables, but this does not necessarily mean that the plot is useless. The claim that the real factors being plotted are c and b is partly true, since c and b will vary inversely when the recall and precision ratios vary inversely. But to suggest that because of this a plot of recall : fallout is better is quite false, since a plot of c against b also represents the main factors in this plot also; $\frac{a}{a+c} : \frac{b}{b+d}$, where the a's and b's cancel out, leaving a plot of c : d, but since d varies inversely as b in Fig. 2, the complementary plot is c : b. The crux of the matter is that in the precision ratio, $\frac{a}{a+b}$, the actual value is determined mainly by the value of b in most cases. It is true that 'a' must have some effect, but in most cases 'a' is a small value compared with 'b'. When high precision ratios are achieved, towards the right of the plot, the effect of 'a' will be greater, and the precision ratio will be affected by the recall value. But even here the effect is usually negligible, as can be shown by using a 'Corrected Precision' ratio suggested by Fairthorne (ref. 5). This measure, known as the Distillation Ratio, is $\frac{a}{a+b} - \frac{c}{d}$, that is, the precision ratio less a correction factor of the relevant not retrieved as a proportion of the non relevant not retrieved. When this correction factor is 'negligible compared with the precision ratio, the latter is a valid measure' (ref. 5). In the results presented in table 4, the factor at a coordination level of five terms in Search T is 0.38%, which is definitely negligible. If this correction factor $\frac{c}{d}$ is correct, then it will usually be a small value since it will approximate to the generality ratio in many cases, and the range of

generality ratios encountered in many tests so far is approximately 1 to 6 (i.e. no more than 6 in 1000), except in a few cases where tests have been conducted on very small collections.

Another argument advanced for using a fallout type measure is because it 'takes into account one of the vital parameters in a retrieval system - size of file' (ref. 10, page 7). It has been shown that it is really the generality ratio that is important, and in test situations where it is constant it is unnecessary to use a ratio that does not vary with generality. However, even in cases where generality does alter, it is a simple matter to adjust the precision ratio to allow for this, as will be demonstrated.

Table 10 shows two sets of performance results, which it is desired to accurately compare. For this hypothetical example, case A is a collection of 1000 documents, with 10 relevant; case B still only has 10 relevant, but the collection size is 10,000, resulting in a very large generality change from 10 (case A) to 1 (case B). In both cases the recall is 50%, and the proportion of non relevant retrieved to collection size remains the same (10 documents in case A, 100 in case B) resulting in a fallout ratio of 1.0% in both cases. But the precision ratio alters considerably, from 33.3% in case A to 4.8% in case B, as generality is decreased. A recall : fallout plot would indicate an identical performance for the two cases, although the information that in case A 1.0% fallout means retrieving 10 documents, and in case B it means retrieving 10 times more again, would be highly desirable. A plot of recall : precision would show a large change - quite correctly since the increase in collection size was matched by a corresponding increase in non relevant retrieved, but with no increase in the relevant documents. But if a strict comparison of the two cases using the precision ratio is desired, the generality ratio must be held constant in some way.

Four different ways of choosing a constant generality ratio are suggested, in order to adjust the precision ratios to enable accurate comparison:

1. Case A altered to the generality of case B; the situation with the higher generality altered to the lower one.
2. Case B altered to the generality of case A; the situation with the lower generality altered to the higher one.
3. Cases A and B altered to the average generality of the two cases.
4. Cases A and B altered to a standard generality ratio chosen for presenting test results.

To adjust the precision ratio in this way, the following formula may be used:-

$$\text{adjusted precision ratio} = \frac{(R \times 1000G)}{(R \times 1000G) + (F(1000 - 1000G))} \times 100$$

where R = recall ratio

G = generality ratio

F = fallout ratio

In the example being considered, if method 1 is adopted, the adjusted precision ratio for case A can be calculated using the formula, and involving the generality ratio of case B ($\frac{1}{1000}$), and the result will be 4.8% precision. This is clearly correct, since with both cases now having a constant generality ratio the precision ratio must be the same in both cases. To illustrate the use of three of the four suggested methods for adjusting the precision ratio a more complex example is given in table II. The two cases for comparison, C and D, are given in the form of retrieval tables of figures, with both cases having a similar recall ratio, but case D having the better fallout ratio, 1.0% compared to 1.2% in case C. However, the precision ratios give a conflicting result, with case C giving 20%, and case D 16.7%, due to the change in generality ratio from 5.0 to 3.4. The first three methods have been used to adjust the precision ratios for three standardised generality ratios, the highest (5.0), the lowest (3.4), and the mean (4.2). The adjusted precision ratios are given in table 12, where it is seen that the correct superiority of case D is now seen in all results. This table shows how a given precision ratio increases with an increase in generality, and how the increase of case D over case C will be shown up more clearly at the higher generality values, since at a generality of 5 the precision ratio is increased by 2.8%, compared with 2.1% at a generality of 3.4. It is suspected that the increases will just be proportional to the size of the numbers involved, with bigger ratios showing up a proportionally bigger difference. The fourth suggested method of adjusting the precision ratios is the adoption of a chosen standard generality value for the reporting of retrieval tests, but choice of such a value would be difficult to meet all needs.

It has been shown that a plot of recall and precision can be used in retrieval tests to make every kind of comparison that is possible. No final statement as to which set of twin-variable measures is best can be made, since both methods give a comparable and in some ways complementary indication of retrieval performance. In the rare cases where the generality ratio does alter, extra adjustment is needed for the precision ratio which is not required for the fallout ratio. The recall and precision plot can be be criticised for including two variables which in certain cases have a degree of dependant variation in their values, but it is clear that in the majority of retrieval tests the factor will be negligible.

The use of such twin variable measures as these is described as 'an unnecessarily weak procedure' by J. Swets (Ref. 8, page 248). However, he qualifies this by assuming that a 'real retrieval system has a constant

effectiveness, independant of the various form of queries it will handle' and then continues by stating that such an assumption is open to question. The assumption clearly does not hold for an experimental test situation, or one where major variables in the system are being varied, really resulting in quite different systems. In such tests the twin variable measures are necessary to see all the changes over the whole range of performance. In tests of operational systems, where each part of the system is operated as close to a typical real life situation as possible, it may be that twin variable measures will not be required and some type of composite measure, considered next, may be of use. But in no cases can the twin variable measures be weaker than the composite ones, since all composite ones present some compressed and simplified combination of the whole range of values shown by twin variable measures.

Composite Measures

The discovery of a single measure to reflect retrieval performance has an alluring appeal, and quite a number of suggestions have been put forward. Since any such measures can only use various combinations of figures from the retrieval table and since it has been shown that the twin variable plots accurately reflect retrieval performance, 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 very near 0% recall at 0% precision at 100% fallout. So all composite measures have some scale of values between those two extremes, which can be plotted for visual examination on both a recall : fallout and recall : precision plot.

Some of the proposed measures may be described as linear composite measures, when their values vary in some linear way when 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, or recall and fallout ratios. Table 13 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 as good a performance as 10% recall at 70% precision, and many other different levels along the diagonal line. Some 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, table 14. The performance curves from table 4, plotted on both tables are seen to have composite values which generally indicate the superior performance of search T, but of course the detailed differences at the cutoff points and maximum recall loss cannot be indicated by any composite measures.

Two measures of this type have been proposed. The first is used by

J.D. Sinnett in his thesis describing a test of role indicators, (Ref. 11), where he uses an effectiveness measure 'R' being $R = 100(\frac{a}{a+c} - \frac{b}{a+b})$ 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, as table 13, and have a range of values from - 100 to + 100, with the centre diagonal being 0. The second proposal, by Western Reserve University is the measure 'Effectiveness', being the sum of recall and specificity (ref. 4), and appears as straight lines on a plot which just reverses recall : fallout, as is seen in table 15.

Other composite measures proposed use more complex combinations of values from the retrieval table, and may be described as non-linear measures because the scale of values varies when recall or precision (or fallout) are varied. When a measure of this type includes the value of 'd' in its equation, the associated lines on a Recall : Precision plot will vary in position according to the generality ratio, so a ratio of 5:1000 is used in tables 16 and 18.

The simplest measure proposed by Verhoeff (ref. 14) is a 'measure of merit' involving the formula: a - b - c + d. This can also be written as:- (a + d) - (b + c), and is the sum of the successes minus the sum of failures. Tables 16 and 17 plot the values on the two plots, together with the performance curves, and show how the high values of the measure occur at the high precision and low fallout areas of the plots.

A more complex version of this is the 'Q' factor, proposed for use in retrieval tests by J.E.L. Farradane. This is a well known statistical coefficient of association proposed by Yule (ref. 13). The formula is $Q = \frac{ad - bc}{ad + bc}$, 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. Tables 18 and 19 show the two plots with 'Q' curves plotted, with the performance curves. The position of these Q curves is closer to the performance curves plotted, and other results from the present project, than the other composite measures.

A measure suggested by B.C. 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 table 20, and since the equation does not include 'd' the position of the curves does not vary with generality. 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. These curves are again somewhat similar to some observed performance curves, although do not fit the plotted results very closely.

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 all these measures cannot show the very large differences that occur inbetween these two points, in the different positions at which systems actually operate. If it is accepted that the curves in tables 4 and 6 are accurate indicators of retrieval performance when a component of a system is varied to give results over the largest possible operating range, then all composite measures can only reflect just one point of such curves. It is unfortunate that the point on the curves which determines the value assigned to that test by a given composite measure is usually either the point of maximum recall, or of maximum precision (top or bottom ends), both of which frequently may not be the best points to use. It is a reasonable conclusion that for experimental tests where changes in the variables in systems are being examined, the composite measures are inadequate. For tests where a single cutoff point is chosen, or a single cutoff is applied to two systems in a comparable manner, some of the composite measures may be used.

Having examined the main suggested performance measures, it may be asked whether any theoretical objective methods are known which might 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. 14, page 264) gives data on a population of 77 people, showing the numbers that were both inoculated and uninoculated, and the numbers that were infected and not infected (table 21). The usual purpose of such a table is to ask a question, e.g. 'Is there really some degree of association between the events?', or in this case, 'Is the proportion of people that were uninoculated and became infected significantly different from the proportion of people that were inoculated and were not 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, as was suggested by B.C. Vickery, at the N.A.T.O. Advanced Study Institute, July 1965. He pointed out that no question such as 'Is the proportion of relevant documents in the retrieved set significantly different from the proportion in the set not retrieved' makes 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 thedefficiencies of the composite measures have been noted, and no assistance or confirmation of the twin variable measures being used seems to be given. The tentative conclusion is that statistics does not help at all at this point.

Measures for systems using no cutoff

The measures of performance developed for the SMART system are designed for use with systems that retrieve the whole collection in a 'ranked' order of decreasing correlation with the search request. Since no cutoff is involved, every search retrieves all the relevant documents and all the non-relevant ones as well, and the conventional performance measures described already can not be used. New measures to meet this new situation have been developed, and involve measuring the positions in the ranked list of the relevant and non relevant documents. Perfect performance is obviously achieved if all the relevant documents to a question are at the 'front' of the ranked list (i.e. having the highest correlation with the search), and the worst possible case occurs if the relevant documents are all at the 'back' of the ranked list. Various slightly different measures have been proposed to provide a scale of values in between the best and worst cases, and the two main pairs of measures being known as 'Rank Recall' with 'Log Precision' and 'Normalised Recall' with 'Normalized Precision'. Various additional combinations of measures have been proposed, some including weights to normalize the scale of values, but the interpretation of the measures remains unchanged. Descriptions of the measures are given in ref. 15, papers III and IV. A simplified example is given in ref. 15, page IV 18, figure 6, where a hypothetical question to which there are five relevant documents is put to a collection of 25 documents. The normalized recall measure is calculated for the different cases of retrieval, as follows:-

1. The ideal situation, where the relevant documents have ranks 1, 2, 3, 4 and 5, in the ranked list of 25 documents. Normalized Recall is 1.0.

2. The worst case, where the relevant documents have ranks 21, 22, 23, 24 and 25 in the ranked list. Normalized Recall is 0.

3. A typical case, where the relevant documents have ranks 3, 5, 6, 11 and 16 in the ranked list. Normalized Recall is 0.74.

The equation for normalized recall together with results for the three cases is given in table 22. The normalized measures are preferred to the others proposed because the equation includes the collection size and number of relevant documents, thus allowing comparison between different test

situations. The normalized precision measure is derived in a similar manner to the normalized recall, and employs log values in the equation.

There is little difference between the normalized recall and normalized precision measures. The first reflects the proximity of the relevant documents to the 'front' of the list, the second reflects the proximity of the non-relevant documents to the 'back' of the list. The two measures are thus completely dependent on each other: it is impossible for the value of one measure to change without altering the value of the other, and a perfect value (i.e. 1) of one measure will result in a perfect value for the other also. An account of these measures is given by J. Rocchio, (ref. 15, paper III, page 114), in the following paragraph:-

'Since both these indices reflect over-all performance, a value of 1 for either implies a value of 1 for the other, in opposition to the conventional recall and relevance ratios. The difference between these two over-all measures lies in the weighting given to the relative position of the relevant documents in the retrieved ranked list. The recall index weights rank order uniformly, since it is sensitive to each relevant document. The precision index, however, weights initial ranks more strongly, since it is sensitive to having a high percentage of relevant documents in the initial part of the retrieved list.'

Therefore these normalized measures are in no way comparable with the conventional measures of recall and precision, and they do not and cannot show any 'inverse relationship' of the type observed in the conventional measures. In order to use the normalized measures the output of a search in the SMART system must always be the total collection, and use of conventional recall in such a situation would always give a recall of 100%. If the precision ratio also was calculated for such a situation, it would always be equivalent to the total relevant divided by the total collection.

Some confusion could arise from a statement made by Salton in the following quotation:

'The normalized measures used in the SMART system are equivalent to the values obtained by using the average standard recall and average standard precision for all possible retrieval levels' (Ref. 16, page 100). This would be carried out by making a cutoff after every document as it is 'retrieved', calculating the conventional recall and precision ratios (called by Salton Standard recall and precision), and then averaging all the resulting ratios to give one final pair of ratios called average standard recall and average standard precision. These averaged measures are stated to be equivalent to the normalized measures in both refs. 16 and 17, with ref. 17, pages 8-9 containing an algebraic proof. This equivalence has since been acknowledged to be an error (private communcation, 15th October, 1965) and the new equations are stated to be 'only an approximation'. The average standard recall does give a value very close to normalized recall, but averaged standard precision gives a very different value from normalized precision.

Both the averaged and the normalized measures only give a single pair of values for a given search, and it is not possible to use them to draw any performance curve on a plot of recall and precision. With both averaged and normalized measures, recall and precision will tend to vary in direct relationship, as the recall value increases so also will the precision value. No real similarity between these averaged and normalized measures on the one hand, and the use of conventional measures with a cutoff on the other can be demonstrated. Any retrieval test that does not use a cutoff and does not obtain figures for a, b, c and d in the retrieval table cannot yet be directly compared with the results of a conventional test employing a cutoff.

SECTION 2 METHODS OF AVERAGING SETS OF RESULTS

INTRODUCTION

When the work of the second Aslib Cranfield Project was proceeding to investigate index language devices and other variables, it was realised that work of a similar nature was being done by G. Salton at Harvard University, U.S.A. Although the work at Harvard centred on a computer for the searching and used natural language abstracts as the 'indexing', the different 'options' being tested closely corresponded to the devices and variables being tested at Cranfield using conventional searching and indexing. The area of overlap was not very great, since the Harvard team were using a set of document abstracts and questions in a different subject area, and were not able to investigate some of the things being covered at Cranfield. But from the Cranfield viewpoint the ability of the Harvard system (known as SMART) to use computer searching to obtain speedy results, since the necessary programming of the SMART system had been completed, and the fact that abstract searching was possible, made the idea of co-operation attractive. It was decided to make available to Harvard the set of documents and questions, together with the carefully controlled relevance assessments, that had been obtained at Cranfield. For the Harvard team, this permitted a new test in a different subject area, and provided the twin essentials of a set of document/question relevance assessments, and several available 'dictionaries' or groupings of terms in the subject area. For the Cranfield group, such a test would be a validation of the results and conclusions of the tests at Cranfield, and would enable some extra interesting comparisons to be made. Since the indexing performed on the document collection at Cranfield, in addition to the author abstracts of the documents, was supplied to Harvard, some comparison of searches on the indexing and the abstracts would be possible. For initial tests, a subset of the document collection was supplied, being a set of 200 of the documents used for many of the tests at Cranfield, together with the set of 42 questions having their relevant documents among the 200. In later testing larger sets of documents and questions will be used.

When the first results of the testing at Harvard became available, the interpretation of the results and comparison with results obtained at Cranfield was seen to be a bigger problem than had been realised. The major differences between the SMART system and the conventional methods used at Cranfield is in the form of output from a given search. At Cranfield, various 'cutoffs' are applied in the search, resulting in sets of retrieved documents. In the SMART system no cutoff is used for test purposes, and the output of a search is the whole document file, but ranked in an order of decreasing correlation with the search terms. For this situation, the team at Harvard developed a series of new performance measures, the major ones being Normalised Recall and Normalised Precision. These are described on page 14 , and it can be seen that they are not comparable with the measures used at Cranfield, nor with any of the other measures described in Section 1.

Accurate comparison of the SMART results with the Cranfield ones is only possible if some form of cutoff is applied to the SMART system, so that the measures previously described, and particularly the preferred twin-variable ones, can be used. However, this problem was foreseen before co-operation with Cranfield was suggested, and one method of applying a cutoff in the SMART searches was used in the first test of 17 requests on computers (see pages IV 30-34, Ref. 15). A closer examination of this technique showed that in two respects it was different from the methods used at Cranfield:

The method of averaging the results of individual questions to obtain one final performance figure was different to that in use at Cranfield.

The cutoff method used at Harvard was different and quite incompatible with the methods used at Cranfield.

These two topics, the averaging methods and cutoff methods will be considered in the remainder of this report.

AVERAGING METHODS

To present reliable results of performance, the figures from a set of questions must be averaged in some way. The size of this 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 tables 23 and 24. In these plots of recall : precision the individual results from a set of 35 questions are plotted, where single term natural language indexing and searching is being tested. Table 23 shows the points that result when any 3 out of a possible total of 7 of the search terms in each question are demanded in 'logical product' co-ordination (31 points only are plotted since 4 of the questions retrieve no documents at this level of search) and table 24 shows points from the same questions when the level of search terms demanded in co-ordination is varied from 2 to 7. The scatter is quite wide, in table 24 ranging from 11% recall at 1% precision in the bottom left corner, to 100% recall at 100% precision at the top right corner. But a trend is clearly present down the left side of the plot and at the bottom right corner, with a clear tendency for high co-ordination level results to give high precision and low recall, and lower co-ordination resulting in an inverse change. Two different methods of averaging these results, at each of the 'co-ordination levels', may be used.

The first method, usually used by the Cranfield group, involves obtaining grand total figures of the numbers of documents involved for the whole set of questions, 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 are sought: at a co-ordination level of 3+, 157 of the relevant are retrieved, together with 2,865 non-relevant documents. These totals are then used to calculate the ratios of:-

Recall $\frac{157}{287} \times 100 = 54.7\%$

Precision $\frac{157}{157 \times 2865} \times 100 = 5.2\%$

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

These ratios are obtained for all of the seven possible co-ordination levels, and can then be plotted as points on a graph. Table 25 shows a plot of recall and precision using these particular results, with the seven points plotted and joined by a curve. This procedure of averaging the numbers was used for presenting the results of the first Aslib-Cranfield Project, and the Western Reserve University test. However, even at the time of the latter test it was realised that this method of averaging the numbers results in certain questions affecting the final figures more than others. Non-typical questions, such as those that retrieve an exceptionally large number of non-relevant documents, will exert a lot of 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 amounts of documents were deleted (Ref. 8, page 13).

The second method, which has been used by the Harvard team, converts the results of individual questions into ratios and obtains a total average ratio by using the average of the ratios of each question. The results from the 35 question set have been calculated this way, and Tables 26, 27 and 28 enable a comparison of the 'average of numbers' and 'average of ratios' methods for these particular results. In Table 26 the recall, fallout and precision ratios for the two methods are compared in tabular form, giving five of the seven possible co-ordination levels. It can be seen that there is little difference in the recall ratios between the two methods, at some co-ordination 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 little significant difference. However, in the case of the precision ratios it is clearly seen that the average of ratios gives a substantially higher result for all co-ordination levels, the average increase over the average of ratios being 19.4%. Table 27 is a recall precision plot of the two methods, where the 'better' curve results from averaging the ratios, and table 28 is a similar Recall : Fallout plot.

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 (Ref. 18, page 161). Of course the difference in the results of the two methods is not very great except when the range and distribution of the variation in base becomes large, as is the case with the precision ratio, but not significantly so in the recall and fallout ratios in these particular figures. But 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 G. Salton (Ref. 17). He says that the average of ratios is 'a query-oriented viewpoint', and the average of numbers is a 'document-oriented viewpoint' (page 17). 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 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 (provided that other relevant documents were the same). Thus the average of numbers gives an arithmetical mean value for a set of questions, and the average ratios gives what is really a 'median' value which reflects the performance of a typical question.

In the results processed at Cranfield, the small samples that have been calculated by the average of ratios all show a large increase in precision and an improved performance curve over the average of numbers. The variation in recall can be significant also, as is seen in table 29. Here the results are based on 42 questions supplied to Harvard, and contrary to the Cranfield results, the precision stays exactly the same. This is due to the cutoff method used to establish the performance points, which in this case involves a constant base (total retrieved) in every question, but the recall ratio is improved by averaging the ratios, resulting in a slightly better curve.

The use of the precision ratio in tests causes a problem when the results include a total average of ratios value. The difficulty occurs when the figures in the top row of the retrieval table are $a = 0$ and $b \geq 1$; that is when the retrieved set contains no relevant documents but only non-relevant ones. This could occur when the question being tested actually has no answer in the collection at all, a case which Fairthorne rightly says

is one which an operational retrieval system is faced with, and should be able to meet (i.e. no documents retrieved would be a 'perfect' performance, correctly indicating the non-existence of any relevant). However, the inclusion of such questions in experimental tests seems an unnecessary complication, and no good purpose seems to be served by including them. But this retrieval case, where the retrieved set is all non-relevant, does occur in testing, particularly when the cutoff used retrieves only a few documents, and where the match of the search prescription to document description of the relevant documents is less close than some of the non-relevant ones.

An example is given in Table 30, where in Case 1 question 5 retrieves 50 non-relevant documents and no relevant ones at all. As there are 5 relevant documents sought for question 5, and the collection size is 1000 in this hypothetical example, the fallout value is easily calculated as 5.5%. But the precision ratio is clearly 0%, and would still be 0% even if no non-relevant documents had been retrieved. Thus the precision ratio can only be directly used when at least one relevant document is retrieved. Case 2 gives another set of hypothetical results, where the total retrieved and relevant retrieved is unaltered for questions 1 to 4, but for question 5 no documents are retrieved at all. It must be emphasised that this is a hypothetical result, used for purposes of comparison only: it is improbable that two searches of 5 questions would produce the figures of cases 1 and 2 and some totalling methods discussed later would ignore the question that retrieves no documents in case 2. However, the purpose is to show the two cases of question 5; firstly when 50 non-relevant are retrieved; secondly when no documents at all are retrieved. Clearly the second case is a better performance for question 5, but the precision ratio for that question is still 0%.

No problem arises if, in totalling the five questions, the average of numbers method is used. In this case, both the fallout and precision values are easily calculated, and are seen to be 6.2% and 6.4% respectively. It can also be seen that in Case 2 the values rise to 5.3% fallout and 7.5% precision, because question 5 has a 'better' performance. But if the average of ratios is used, the precision value remains at 7.4% for both cases, although the fallout value shows the correct increase in performance (6.4% to 5.3%). If absolute accuracy is required with use of the average of ratios, in these retrieval cases the fallout value must be used, or the correct fallout value must be obtained to get an adjusted precision. This adjusted precision is easily obtained in the same way that precision is adjusted for generality (page 9). Since the recall, fallout and generality ratios are all known, the correct precision value can be calculated, and is done here with the result of 6.2% precision, which shows the correct superiority of case 2 over case 1.

METHODS FOR ESTABLISHING CUTOFF POINTS

The point at which a search is terminated in a retrieval test is the cutoff point, and this point is reached when the rules being followed do

not allow any fresh documents to be examined. The rules used to establish a cutoff point, or a series of cutoff points, may be based on the following principles:-

1. Some degree of match between the search prescription and document description.

2. The number of documents retrieved.

3. The number of relevant documents retrieved.

The three methods may be used individually, or can be combined to make compound rules. The application of the rules chosen for the tests of operational systems will be a subjective decision which will vary from question to question, and from the requirements of one question to another. For example, if method 1 was in use, one question containing an infrequently used 'potent' retrieval term might be extended in its search to that one term alone; but in another question involving a set of frequently used terms the search might never be extended to using less than three of those terms in co-ordination. The use of methods 2 and 3 in operational tests also requires a subjective decision on the part of the questioner, as to how many documents he is willing to examine (method 2), or how many relevant documents would satisfy his needs (method 3). All these subjective decisions can be made either before the search is carried out, or can be done as the search proceeds, with the questioner or searcher using the feedback being obtained.

For experimental tests, it is usually desirable to eliminate all such subjective decisions, to allow no feedback or variation in rules from one question to another, but to use a fixed rule that can be used for all questions in the test. An exception to this may obtain when a test of different rules about cutoff points is being carried out, but even in such a case strict control to remove as much of the subjective element as possible is essential. Experimental tests also frequently require rules that give a whole series of cutoff points, where each one relaxes the previous search requirements by a controlled amount to retrieve a further set of documents, rather than the requirement, frequently needed for operational tests, of only one final cutoff to terminate the whole search. The purpose of this section is to describe various different rules that are being used to establish series of cutoff points for experimental tests.

An examination of the three cutoff methods above will reveal that method 3 requires that the search output be examined for relevance as the search is being carried out, so that for example, whenever a relevant document is found in the retrieved set, a cutoff is established at that point. As will be shown later, although this method can be consistently applied to a set of questions in a test, the resulting overall performance in absolute terms is very different from that obtained when methods 1 or 2 are used. Rules using all three cutoff methods have been applied to the SMART system, but at Cranfield rules based on method 1 only have been used in the main tests. Since the difference between the Cranfield and Harvard methods is quite large,

the cutoff rules and associated problems will be discussed in two stages, to cover the two distinct types of system that each represents.

Systems which retrieve sets of unranked documents

In such conventional systems, a set of documents less in number than the total collection is distinguished as the retrieved set, in response to a search prescription. Although methods of establishing a cutoff to define the retrieved set can include types 2 and 3 listed above (number retrieved and number of relevant retrieved), these are not normally used in experimental tests since accurate control of the number of documents to be retrieved is impossible. In the Aslib-Cranfield Project a cutoff of the first type is used, and is described as the co-ordination level cutoff.

To illustrate this, two sets of performance figures for question 145 are given in table 31, each set corresponding to a particular index language being tested. The seven major columns are the seven possible co-ordination levels for this particular question, since it has seven search terms in the prescription, and the seven different levels range from the least exhaustive and specific demand of any single term (headed 1+), up to the most exhaustive and specific demand of all seven terms in logical product co-ordination. The actual combinations of terms accepted at each of these co-ordination levels are recorded in a separate search prescription, but the performance results just give the actual numbers of documents retrieved (divided into relevant and non-relevant) at each co-ordination level, that is at each cutoff point. The figures are cumulated, that is at 4+ all the documents retrieved with a co-ordination of four or more terms (in language 4a co-ordinations 4, 5 and 6) are included. It can be seen that for language 4a a co-ordination level of seven terms is too strong, since no documents at all are retrieved, but relaxing the requirements to six terms retrieves ten documents, five relevant and five non-relevant. The results for language 6a show that the change in the language to accepting quasi-synonymous terms in addition to synonyms and word forms only (language 4a) results in one relevant document being retrieved at a co-ordination level of seven, and more documents being retrieved at co-ordination levels 2+ to 6+ as well.

The use of a single co-ordination level as the degree of match between the search prescription and document description at which a cutoff is made is not the only possible technique to use. If the search prescription involves sets of terms in a logical product and sum relationship for example, a cutoff could be made every time a single individual term in the prescription is altered in any way (e.g. dropped off or replaced by another term). For experimental testing however, the choice of cutoff points is usually made in a way that can be applied with equal sense and consistency to different questions, to eliminate the subjective elements in making search prescriptions as much as possible. The cutoff point is really the choice of the point in a search at which the progress of the search is recorded, and so is only of

importance in obtaining search figures for test purposes. However the basic problem with regard to cutoff points occurs when it is desired to amalgamate the results of whole sets of search questions, in order to get some average results for the whole set.

In totalling the results of a set of questions resulting from a test of an operational system, and where only one cutoff point is established in the search (just to terminate it), no problem occurs in amalgamating results of the set. It is in experimental test situations, where search questions are conducted with several cutoffs recording the progress of the search at different points, that problems of totalling occur. This arises because different questions provide differing numbers of cutoff points, and because questions behave differently when variables, such as a change in the index language, are being tested. Although question 145 in table 31 has seven search terms, or seven 'starting' terms in the original request, other questions being used differ from this, and have anything between two and fifteen search terms in the original question, resulting in cutoff points that vary between two and fifteen. The totalling of results of questions that have such 'varying cutoff points' is difficult, and is further complicated by the fact that all co-ordination levels or cutoff points do not actually retrieve any documents in some cases. A case of this was noted in table 31, where a co-ordination level of seven terms in language 4a retrieves no documents, and in many cases the number of co-ordination levels that retrieve documents varies considerably as different languages and other variables are tested within the same question.

Some idea of the variations between questions that are encountered in a test of one variable is seen in table 32. This gives data on 221 questions used in some of the tests, showing how the starting terms vary from 2 to 15, and the maximum terms that retrieve vary from two to ten. For example, there are 35 questions having seven starting terms (column headed 7, total 35 at bottom) and only three of these questions could co-ordinate all 7 terms and still retrieve some items, while one of the questions could co-ordinate no more than two terms as a maximum. The figures in the table alter with any change in language or any other variable. Some methods of totalling sub-sets of these questions, and the whole set of 221 questions will be described next. The aim is to find some method of totalling sets of questions using varying cutoffs, a problem that not only occurs with the co-ordination level cutoff used here but also with the relevant documents cutoff and correlation coefficient cutoff described later.

In table 32, showing the characteristics of a set of heterogeneous questions, sub-sets of the total set are seen to be formed by two different principles. The first principle concerns the number of starting terms or initial search terms contained in the question, and fourteen homogeneous sub-sets are formed by this criteria in the particular test involved. The totalling of the questions in one of these sub-sets (homogeneous starting-term sub-sets) can be simply accomplished by strict co-ordination levels; that is, if the sub-set of 35 questions having seven starting terms is used,

the performance results for each question are totalled for each of the seven co-ordination levels possible, resulting in seven average results. This method ignores the fact that some questions retrieve no documents at all at the higher co-ordination levels, so that at higher levels the final average performance will be based on less documents, or less questions that contribute any results. This is seen in table 33, where the results for this sub-set are plotted as a recall : precision curve, and where it is recorded that the number of questions that contribute results drops from 35 at co-ordination level 2+, to 3 at level 7 (also derivable from table 32). This indicates that the position of the curve at the low recall end will be based on a very small sample size with this totalling method, and will usually sweep across the bottom of the plot to reach high precision values.

A second principle used in table 32 involves the number of terms in co-ordination that actually retrieve documents (always either equal to or less than the number of starting terms in a question), and the total set is divided into nine homogeneous sub-sets by this criterion in the test involved. These homogeneous retrieving-term sub-sets may also be totalled by strict co-ordination levels, and the 45 questions in the five retrieving term sub-set are plotted as a recall : precision curve, connecting the five totals, in table 34. In this method all the 45 questions contribute results (documents retrieved) at all the five co-ordination levels. The effect of this is seen in the recall : precision curve, which at the low recall end terminates at 15% recall at 26% precision.

It is recognized that the homogeneous sub-sets described are only homogeneous in a single criterion, either starting terms or retrieving terms, and each sub-set viewed from the other criterion becomes a heterogeneous one. The sets that are truly homogeneous from both criteria are small, the largest in table 32 being the eighteen five starting term questions that have four retrieving terms. The following ideas on totalling heterogeneous sets therefore can also be applied to these homogeneous ones described, and will precede any comparison of the methods described already.

The first method of totalling heterogeneous sets is by strict co-ordination levels, already mentioned. The characteristics of the performance curves at high co-ordination levels have been noted: when sets of questions having differing number of starting terms are totalled this way the figures and curves will have the reduced sample size and corresponding curve shape of the homogeneous starting term sub-sets and table 33, respectively. Such heterogeneous starting term sets cause additional problems at the higher co-ordination levels: for example, at a co-ordination level of 6+, all questions having less than six starting terms will never even have the possibility of contributing results. This can be allowed for, but only by reducing the question sample size ~~sample size~~ and correcting for changes in generality. The principle behind strict co-ordination level totalling may be faulted for some reasons, since it involves, for example, totalling the results of a three starting term question searched at a co-ordination level of two together with a ten starting term question also searched at a level of two terms. There are other good reasons why this method should be used for displaying certain

types of results.

In order to meet this last problem, the second totalling method is by proportional co-ordination levels. Here questions are totalled by aligning co-ordination levels proportionally; for example, the three starting term questions at a co-ordination of two terms would be totalled with the six starting term questions at co-ordination of four terms - and all other questions using approximately $^2/_3$ of their available terms in co-ordination. Correlation of different questions is difficult to do this way, but several techniques are satisfactory. Some techniques are forced to 'ignore' some of the performance results produced by questions that retrieve with a large number of terms in co-ordination, and another technique uses the performance results which are nearest certain selected recall position on the plot (e.g. at 5%, 10%, 15% etc. to 100%) to obtain a performance curve drawn through the slight scatter of the 20 points produced by this method. The curves drawn by all these techniques are of the type shown in table 33 (with one exception), and usually have somewhat diminished sample sizes at low recall.

The third totalling method involves aligning the results at some maximum co-ordination level, either the maximum co-ordination level possible (starting terms), or the maximum co-ordination level that retrieves any documents. The latter criterion is usually adopted, and gives performance curves similar to the one in table 34. This method is best understood by assuming that a set of heterogeneous search questions are put to the system, with their total starting terms demanded in co-ordination. The levels of co-ordination are relaxed until each question retrieves some documents, and at that point (the maximum co-ordination level that retrieves in each question) the total performance results are calculated. From this point onwards, a single co-ordination level at a time is dropped off each question, until all questions are reduced to a single term, and maximum recall is attained. Thus the bottom end of the performance curve is always based on results from the whole set of questions, and the results in the present test takes up the shape shown in table 34.

These three somewhat complex methods are not being examined in detail, but given as the main solutions that have been tried so far to solve the problem caused by varying cutoffs. Several techniques based on each method have been tried, and detailed comparison would require long explanations. The differences between the three main methods really involve a question of test technique, being the choice of the points during the search at which the performance achieved is recorded, in order to total with other questions to obtain a single average result. The choice of a method depends entirely on the particular purpose of the test being made, since some methods are best used to display particular variables. For example, if a series of different index languages are to be compared, the strict co-ordination level methods may be preferable, since at each co-ordination level the points plotted on the recall : precision graph will show accurately the change in performance between different index languages. But if a performance curve

that shows a typical range of performance for a particular index language is required, it may be desirable to use a method that does not involve a reduced sample size at the bottom end of the curve, such as the maximum co-ordination level method. Where performance results obtained from less than a total set of questions can be used, the choice of some homogeneous groups can ease the problems; also the comparison in performance of different homogeneous groups (which will involve variations in generality) may be an important part of a test.

To return to the present subject of the use of co-ordination levels as a cutoff in conventional systems: the performance curves finally obtained are satisfactory for all the types of test likely to be made in any experimental test. However, the problem of the varying cutoffs does cause a large problem in practice, and must slightly affect the accuracy of the results. Cutoffs of types 2 and 3 (page 22) are not readily applicable to conventional systems that distinguish sets of retrieved from sets of non-retrieved documents, so these types will be covered in the next section, together with one new method devised to enable a cutoff of type 2 to be used in a test of a conventional system.

Systems producing a ranked output

It has already been noted that the SMART system is normally operated without a cutoff for the purposes of experimental tests, and that the use of a cutoff is necessary both to allow use of the conventional measures of performance described in Section 1, and to compare the SMART test results with the Cranfield ones. Systems producing ranked output can more easily use cutoff methods 2 and 3 than conventional systems, i.e. cutoffs applied after a given number of documents have been examined or after a given number of relevant documents have been examined. Conventional systems cannot control the numbers of documents retrieved very accurately, but in a ranked output where the correlation coefficient gives a value that is different for nearly all documents in the collection, the control of the output and use of such cutoffs can be done in a more refined way. Cutoff rules based on all three different methods have been used in the SMART tests, and each will be considered, commencing with method 3, then method 2, and finally method 1.

The method first used at Harvard is described in Ref. 15, IV pages 30-31, and involves applying a cutoff immediately a relevant document is found. This produces performance curves known as 'Quasi-Cleverdon Graphs', but here described as 'Relevant Documents Cutoff'. The SMART output is always a ranked list of documents, arranged in an order of decreasing correlation with the question, and Table 35 gives an example showing the ranks of the relevant documents for five questions, searched on a collection of 200 documents. To apply the relevant documents cutoff, the correct recall and precision values are calculated after each relevant document is reached in the ranked list, and the average of these ratios taken over a set of questions. Table 36 shows the last cutoff point used for the 5

questions in Table 35; i.e. at 100% recall the precision ratio is calculated by dividing the total relevant found by the rank of the last relevant document found (being the total retrieved) and an average of the ratios taken over the questions. In practice the procedure is slightly more complex than this, since ten recall values are chosen (10%, 20% etc. to 100%) and then actual results 'smoothed out' to get ten sets of ratios, to solve varying cutoff problems.

Although this method results in a reliable plot of recall and precision, the position of the curve on the recall : precision plot is greatly affected by the cutoff being applied immediately a relevant document is found. It can be seen intuitively that the resulting precision ratio is unusually high, for example, 39% precision at 100% recall in Table 36 is extremely high. This cutoff rule chooses the optimum point at which to make a cutoff, and could only be achieved in a real-life situation if the questioner assessed the output document by document. Even if such a situation is being simulated here, the establishing of a final cutoff just when the last relevant document in the collection is reached is virtually the use of hindsight, since knowledge as to exactly which relevant document was the last one would not be known. This does not mean that for testing purposes the relevant documents cutoff is necessarily unreliable in any way, or that test results showing differences in performance between different index languages are not accurate. Further investigation is needed for final proof, but it is fairly certain that relative differences between two performance curves derived from different index languages are quite reliable, but take up a quite different and greatly improved position on a performance plot when compared with other cutoff rules. In order to compare the SMART and Cranfield results it is desirable to use a similar or identical cutoff rule, and the co-ordination level used at Cranfield does not use the intellect and virtual hindsight involved in the relevant documents cutoff. Another example of the difference caused by these different rules is seen in the results for question 264 in Table 35, where the SMART system gives the two relevant documents the ranks 1 and 2. With the relevant documents cutoff the precision ratio cannot be less than 100%, but in the Cranfield searches the rules progressively relax the search requirements and so reduce the precision ratio, even if the first cutoff happened to produce 100% recall and 100% precision.

A second method of cutoff, proposed at Cranfield, is the use of a Document Output Cutoff. In this case a cutoff is applied to the ranked output of a search as soon as a certain number of documents has been examined, whether this includes many relevant ones or not, and the recall and precision values are calculated at that point. For testing purposes, a fixed set of cutoff points is chosen, say after the first 5 documents in the output, then after 10, and so on, probably up to the last document in the collection in order to reach 100% recall over a set of questions. Table 37 gives an example of this method, as it is used for calculating the results for question 230, with the ranks of the relevant documents given in Table 35. Eleven cutoff points are chosen, and the recall and precision values calculated at each

point, ranging from 28½% recall at 40% precision after 5 documents, to 100% recall at 3½% precision at the 200th document. For question 264 already referred to, it can be seen that only a cutoff of two documents gives 100% recall at 100% precision: at 5 documents precision would be 40%, and at the 200th document 1%. The five questions in Table 36, at 100% recall would now have 2.6% precision. These figures are similar to the results at Cranfield - an apparently shocking performance, but for good reasons not being explored here.

The document output cutoff has two favourable characteristics, in relation to the problems previously mentioned. In totalling sets of questions, this cutoff does not give varying results for different questions, since a controlled number of documents is retrieved at each successive cutoff, and the totalling of results for a set of questions is straightforward. For the same reason, there can never be any difference in the final average precision ratio whichever averaging method is used, because at a given cutoff the base value in the ratio remains constant for all questions.

The magnitude of the difference in the curves produced by the relevant documents cutoff compared with the document output cutoff is seen in Table 38. The plot gives results of the 42 questions supplied by Cranfield, being the Null thesaurus option with numeric vectors tested on the abstracts. The curve for the output cutoff results does not extend below 29% recall, since that figure is obtained at an output cutoff of only two documents. Very similar curves result from the Cranfield results, when either of two cutoff methods are employed. In terms of real life operation of a system such as SMART, the use of such an output cutoff seems a reasonable and useful method to use in order to examine only a portion of the documents in the system. But it can also be shown that this curve is representative of the performance of SMART if a third possible method were used.

Since the ranked output of the SMART system is obtained by calculation of a correlation coefficient, in this case a decimal value ranging from 0 to 1, the choice of a minimum acceptable value for the correlation coefficient seems a reasonable method of making a cutoff. For test purposes, the correlation cutoff is applied by choosing a standard set of values, say 1.0, 0.95, 0.90, 0.85 etc. down to a minimum of 0.05. At each point the documents with a correlation measure equal to, or greater than, that value are taken as retrieved, and the recall and precision ratios calculated. This method has been tried on the 42 questions, tested by the Null thesaurus with logical vectors testing the indexing, and the resulting curve is shown in Table 39. The performance at high recall has not been calculated, only because the total output was not available at Cranfield, the output received only listing the fifteen documents with the highest correlation with each question. The position of the curve at low recall can be calculated with fair accuracy, and is seen in Table 39, together with the curve resulting from the output cutoff. The use of the correlation cutoff for testing purposes is undesirable, because exactly similar problems to those encountered by the Cranfield results using 'co-ordination cutoff' occur again: i.e. the maximum cutoff that 'retrieves'

any documents varies from question to question, and also varies within each question when different options are tested, resulting in problems of totalling sets of questions. A totalling method of the 'strict co-ordination level' type has been used here; if a 'maximum co-ordination level' method is used, the position of the lower end of the curve is 17% recall at 42% precision.

The comparison of SMART and Cranfield cutoff methods

Although it is suggested that comparisons between the SMART and Cranfield results can be made using the document output cutoff and co-ordination level cutoff respectively, it must be admitted that the correspondence is not perfect. A slightly closer correspondence is possible if the SMART system uses a correlation coefficient cutoff, but this rule has not been applied to any of the results on any scale, and the problem of varying cutoffs would cause similar difficulties to those encountered at Cranfield. An ideal comparison would be possible if the document output cutoff could be used for both tests, and so a suggested method of applying this rule to the Cranfield output is described.

In order to use the output cutoff at all, a ranked set of documents in order of decreasing match with the search question must be provided for each search. The product of the Cranfield searches is always several groups of documents at differing co-ordination levels, each level being a lesser match with the question, so a partial ordering can be achieved in this way. Table 40 shows these groups in order of decreasing co-ordination level from left to right and at each level is recorded the numbers of relevant and non-relevant documents retrieved. In order to simulate a ranked output, the figures are first re-processed to remove the cumulative nature of the results (since the documents retrieved at level 5+ are included in 4+, 3+,, 2+, etc.), to indicate at each co-ordination level how many new relevant and new total documents are retrieved. This is seen in Table 40 rows 2 and 3, where it is seen that at co-ordination level 5+ three documents are retrieved, with one relevant; at co-ordination level 4+ an additional ten are retrieved with two being relevant, and so on. The total retrieved (row 2) can now be given a set of ranks (row 4), and the relevant documents are put in the 'middle' position in rank within the group concerned. For example, in the first group the ranks are 1, 2 and 3, therefore the single relevant document is given position 2. Where an exact middle position doesn't exist, the rank nearer the beginning is given; e.g. at a co-ordination level 2+ the middle position between documents 35 and 82 falls between documents 58 and 59, so number 58 is given to the relevant one. The final result for the six relevant documents gives them ranks of 2, 8, 9, 23, 24 and 98 (row 5), and these can be used with an output cutoff to obtain results. The samples done this way show close correlation of the resulting performance curve with the curve produced by a co-ordination level cutoff, although difficulties in certain details remain to be solved.

This method has two advantages:-

1. It would provide one solution to the problem of varying cutoffs encountered in the cutoff rules used at Cranfield.

2. It would also enable a close and accurate comparison to be made between the SMART and Cranfield test results.

References

1. VICKERY, B.C. — On Retrieval System Theory. London, Butterworths, 2nd edition 1965.

2. PERRY, J.W. — Operational Criteria for designing Information Retrieval Systems. American Documentation 6, 1955, 93-101 (No. 2), also in Machine Literature Searching.

3. CLEVERDON, C.W. — Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, 1962.

4. GOFFMAN, W., and NEWILL, V.A. — Methodology for Test and Evaluation of Information Retrieval Systems. Comparative Systems Laboratory Technical Report No. 2. C.D.C.R., Western Reserve University, Cleveland, 1964.

5. FAIRTHORNE, R.A. — Unpublished Notes.

6. SWANSON, D.R. — Paper presented at the Congress of the International Federation of Information Processing Societies, Munich, 1962.

7. SWETS, J.A. — Information Retrieval Systems. Science, 141, 1963, pp. 245-250.

8. AITCHISON, J. and CLEVERDON, C.W. — Report of a test on the index of metallurgical literature of Western Reserve University, Cranfield, 1965.

9. CLEVERDON, C.W., MILLS, J. and KEEN, E.M. — Factors determining the performance of index languages. Vol. I, Cranfield, 1966.

10. REES, A.M. — The evaluation of Retrieval Systems. Comparative Systems Laboratory Technical Report No. 5. C.D.C.R., Western Reserve University, Cleveland, 1965.

11. SINNETT, J.D. — An evaluation of links and roles used in information retrieval. Dayton, Ohio, Air Force Materials Laboratory, 1964.

12. VERHOEFF, J., GOFFMAN, W., and BELZER, J. — Inefficiency of the use of Boolean functions for information retrieval systems. Communication of the Association for Computing Machinery, 4, 1961, 557-558, 594.

13. KENDALL, M.G., and STUART, A. — The Advanced Theory of Statistics. (Griffin) 1961.

14. MORONEY, M.J. — Facts from figures. Penguin, 3rd edition 1956.

15. THE COMPUTATION LABORATORY, HARVARD UNIVERSITY. — Information Storage and Retrieval, Scientific Report ISR-8, Cambridge, Massachusetts, 1964.

16. SALTON, G. — Progress in automatic information retrieval I.E.E.E. Spectrum, August 1965, p. 90-103.

17. SALTON, G. — The evaluation of computer-based information retrieval systems. Paper given at F.I.D. Congress, Washington, October, 1965.

18. CROXTON, F.E. and COWDEN, D.J. — Applied general statistics. New York, 1939.

Co-ordination Level	Recall Ratio	Precision Ratio
1+	95.2%	0.8% *
2+	76.9%	1.6% *
3+	57.1%	4.2%
4+	40.4%	11.7%
5	27.9%	29.3%

* Estimated Results

TABLE 1 TABULAR COMPARISON OF RECALL AND PRECISION RATIOS AT FIVE CO-ORDINATION LEVELS, BASED ON 20 QUESTIONS.

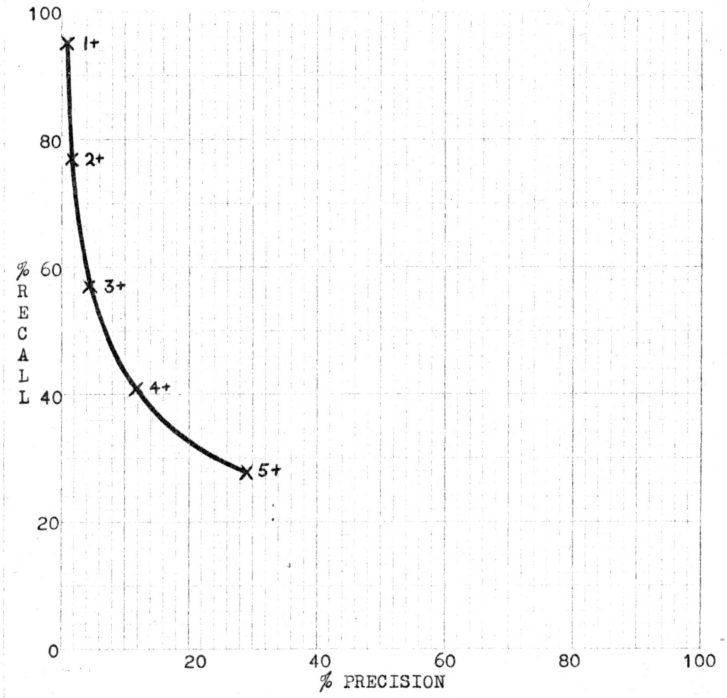


TABLE 2 PLOT OF THE RECALL AND PRECISION RATIOS IN TABLE 1, WITH CO-ORDINATION LEVELS INDICATED.

Co-ordination Level	Recall Ratio		Precision Ratio	
	S	T	S	T
1+	95.2%	69.4%	0.8%*	3.5%
2+	76.9%	55.1%	1.6%*	8.9%
3+	57.1%	40.3%	4.2%	20.3%
4+	40.4%	26.6%	11.7%	55.5%
5	27.9%	23.3%	29.3%	65.7%

S Search Rule S

T Search Rule T

* Estimated Results

TABLE 3 TABULAR COMPARISON OF RECALL AND PRECISION RATIOS AT FIVE CO-ORDINATION LEVELS WITH TWO SEARCH RULES, BASED ON 20 QUESTIONS.

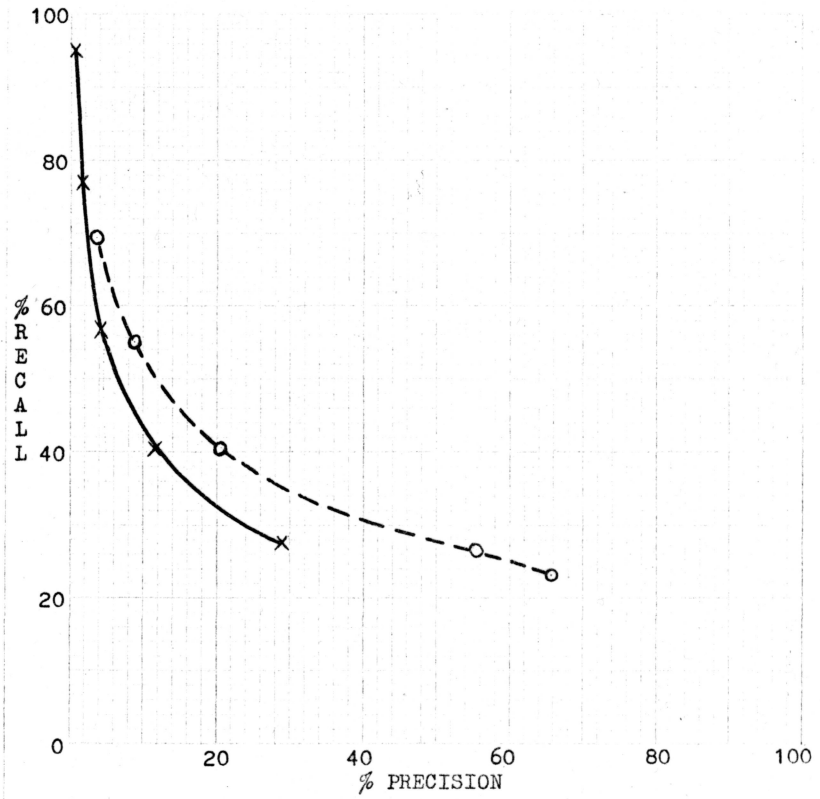


TABLE 4 PLOT OF THE RECALL AND PRECISION RATIOS IN TABLE 3.

Co-ordination Level	Recall Ratio		Fallout Ratio	
	S	T	S	T
1+	95.2%	69.4%	59.3%	9.6%
2+	76.9%	55.1%	24.1%	2.8%
3+	57.1%	40.3%	6.6%	0.8%
4+	40.4%	26.6%	1.5%	0.1%
5	27.9%	23.3%	0.3%	0.06%

S Search Rule S

T Search Rule T

TABLE 5 TABULAR COMPARISON OF RECALL AND FALLOUT RATIOS AT FIVE CO-ORDINATION LEVELS WITH TWO SEARCH RULES, BASED ON 20 QUESTIONS.

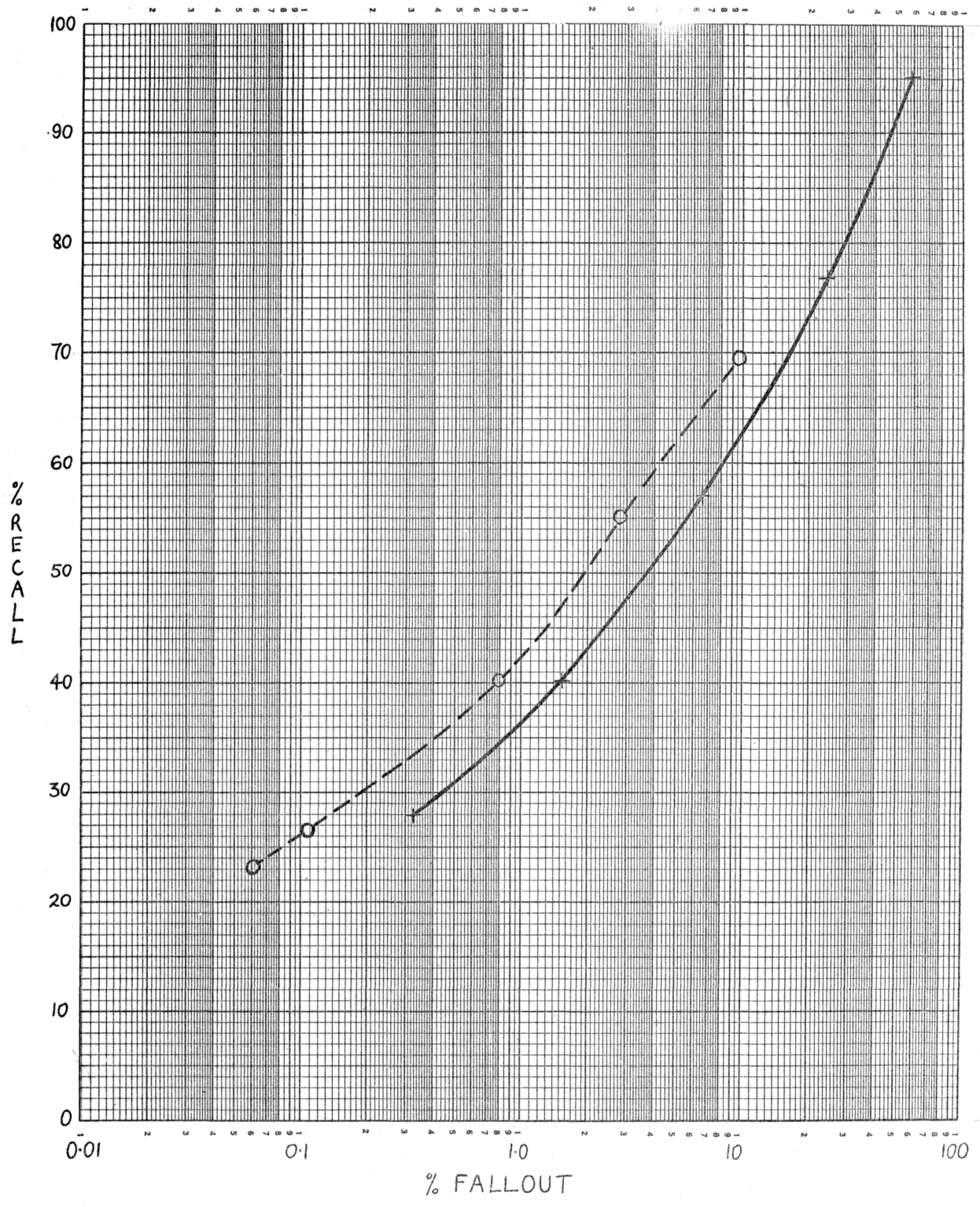


TABLE 6 PLOT OF THE RECALL AND FALLOUT RATIOS IN TABLE 5.

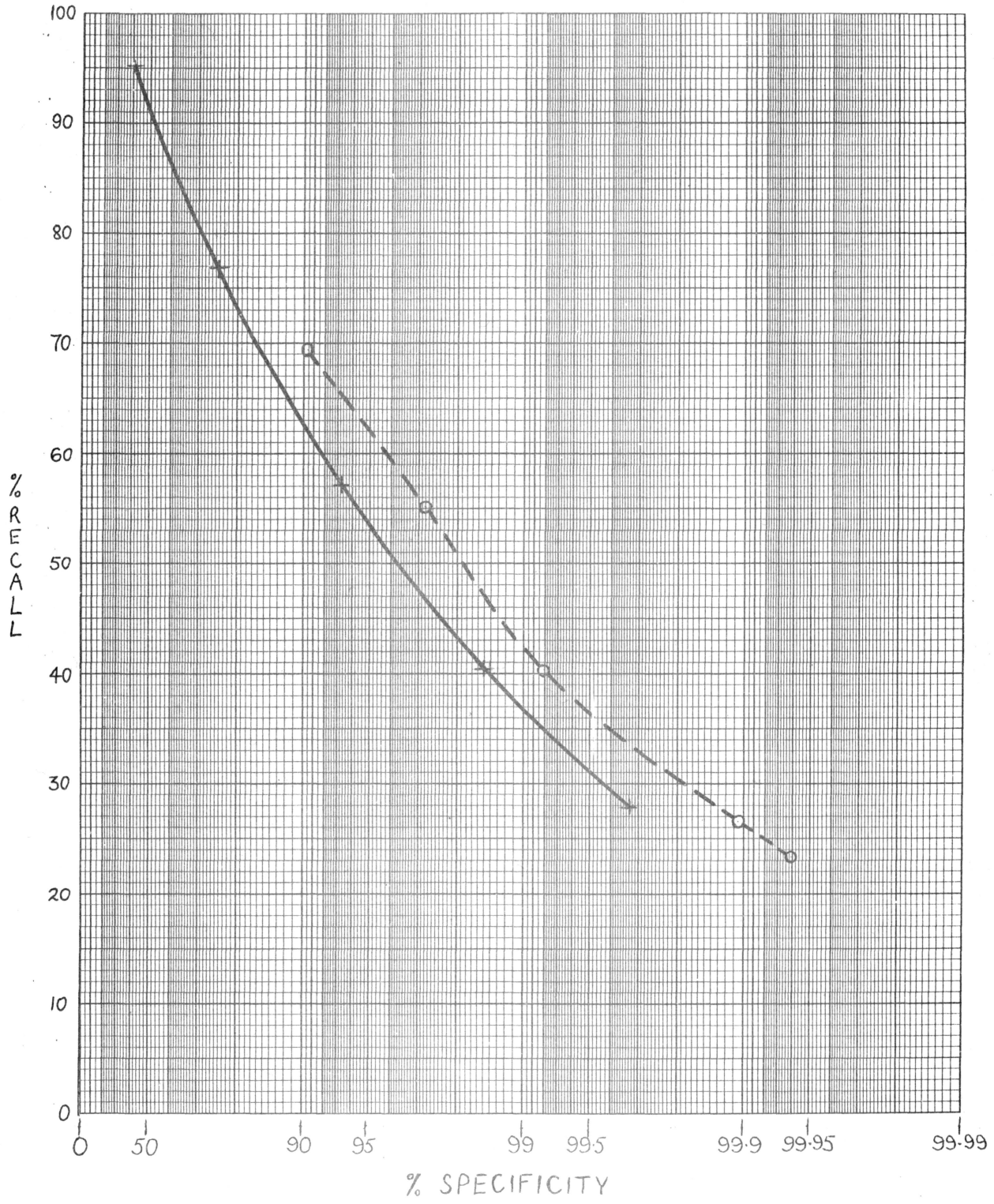


TABLE 7 PLOT OF RECALL AND SPECIFICITY RATIOS DERIVED FROM TABLE 5.

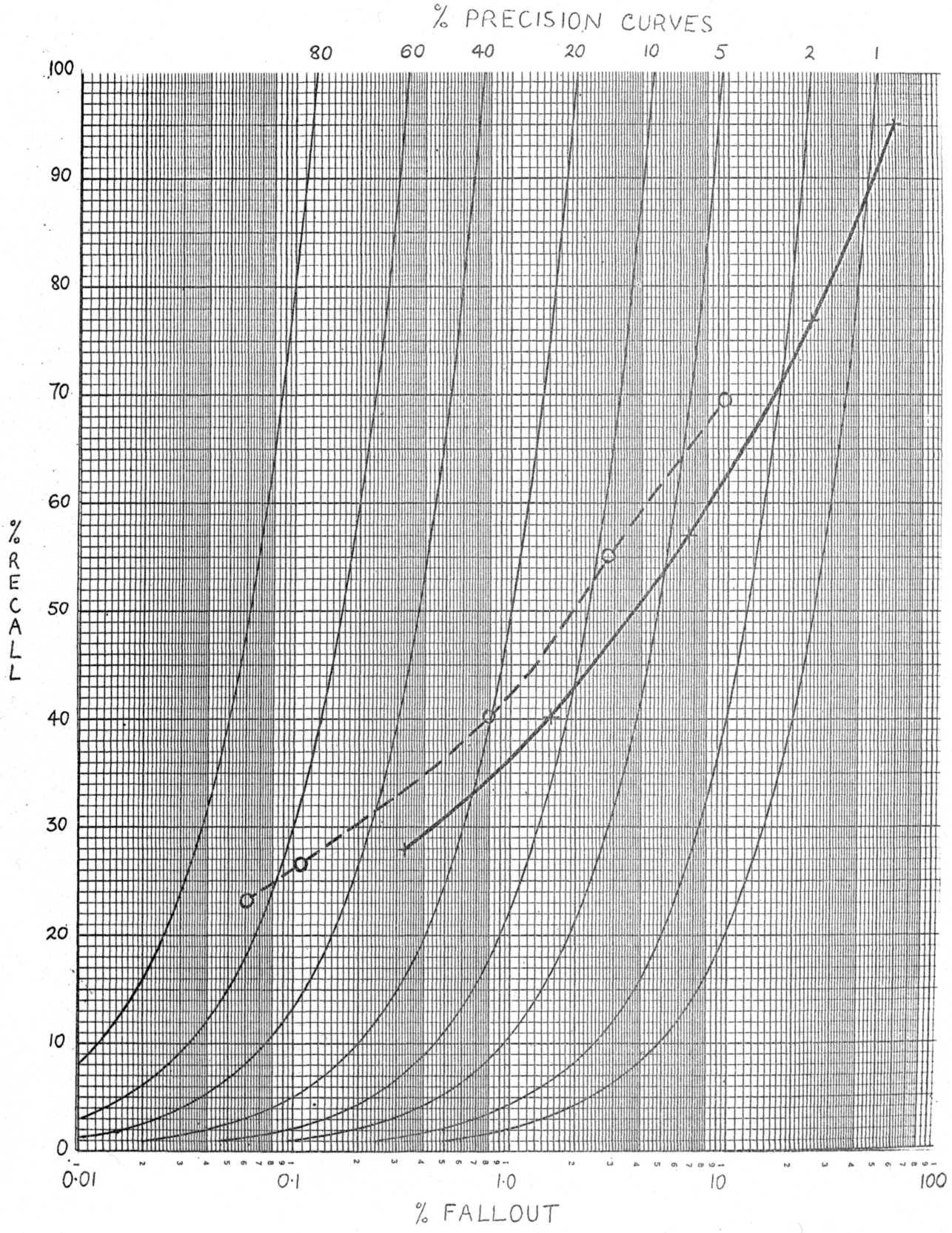


TABLE 8 PLOT OF RECALL AND FALLOUT AS TABLE 6 SHOWING THE PRECISION RATIO CURVES.

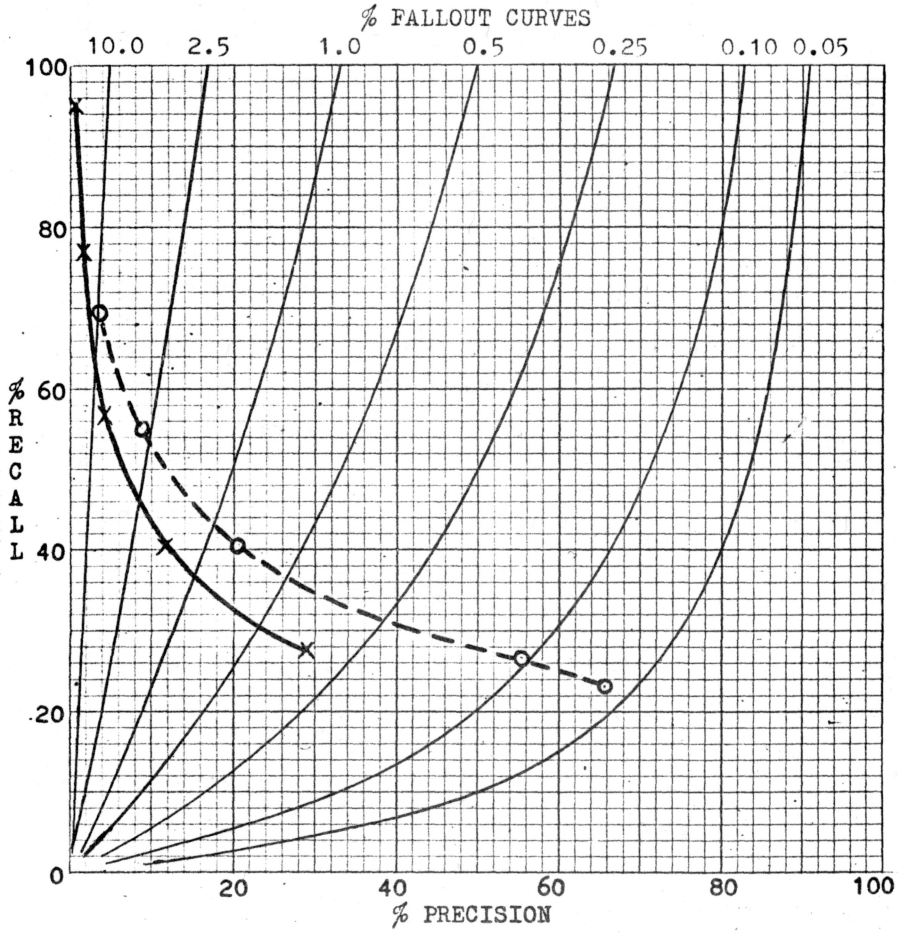


TABLE 9 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING THE FALLOUT RATIO CURVES.

CASE A

	Relevant	Non Relevant		Generality 10:1000
Retrieved	5	10	15	Recall 50%
Not Retrieved	5	980	985	Fallout 1.0%
	10	990	1,000	Precision 33.3%

CASE B

	Relevant	Non Relevant		Generality 1:1000
Retrieved	5	100	105	Recall 50%
Not Retrieved	5	9890	9895	Fallout 1.0%
	10	9990	10,000	Precision 4.8%

TABLE 10 TWO SETS OF PERFORMANCE RESULTS WITH DIFFERENT GENERALITY RATIOS AND CONSTANT RECALL AND FALLOUT RATIOS

CASE C

	Relevant	Non Relevant		Generality 5:1000
Retrieved	3	12	15	Recall 60%
Not Retrieved	2	983	985	Fallout 1.2%
	5	995	1,000	Precision 20%

CASE D

	Relevant	Non Relevant		Generality 3.4:1000
Retrieved	10	50	60	Recall 58.8%
Not Retrieved	7	4933	4940	Fallout 1.0%
	17	4983	5000	Precision 16.7%

TABLE 11 TWO CASES OF PERFORMANCE RESULTS WITH DIFFERENT GENERALITY AND FALLOUT RATIOS

Generality	Adjusted Precision Ratios Case C	Case D
3.4	14.6%	16.7%
4.2	17.4%	19.9%
5.0	20.0%	22.8%

TABLE 12 ADJUSTED PRECISION RATIOS FOR THE TWO CASES OF TABLE 11 AT THREE CONSTANT GENERALITY RATIOS

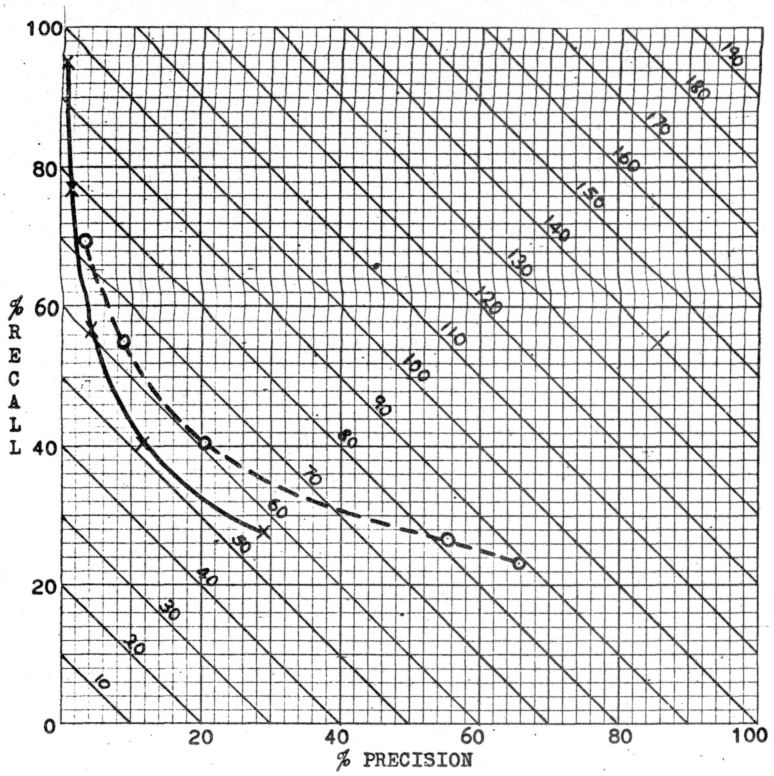


TABLE 13 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING THE 'RECALL + PRECISION' LINES.

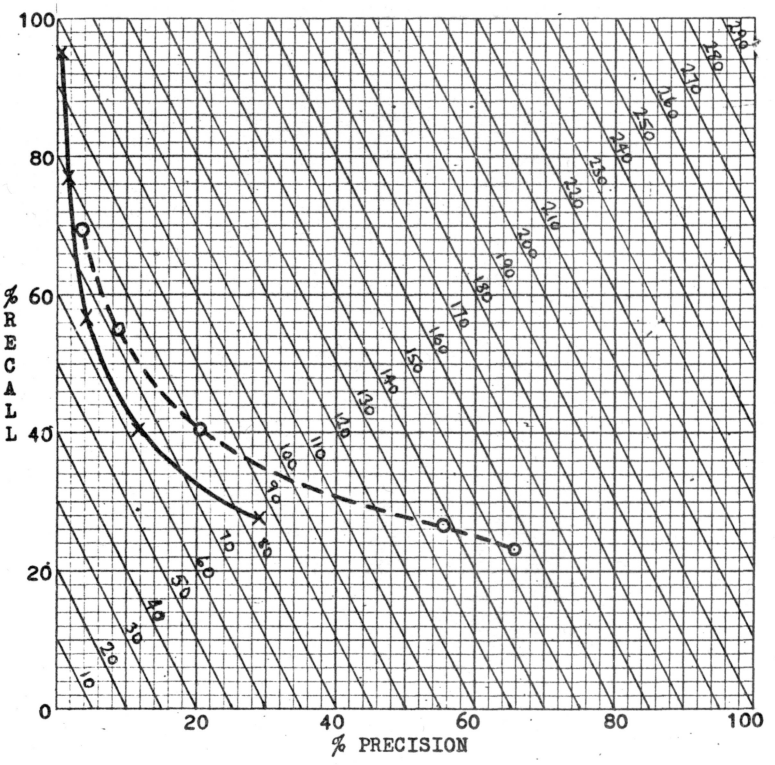


TABLE 14 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING THE "RECALL + PRECISION X2" LINES.

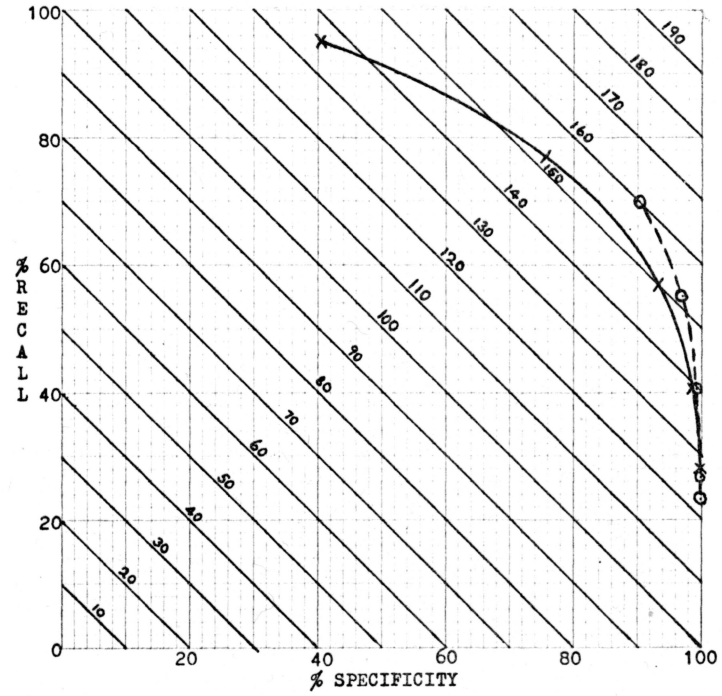


TABLE 15 PLOT OF RECALL AND SPECIFICITY AS TABLE 7 (BUT ON A LINEAR SCALE) SHOWING THE "EFFECTIVENESS" LINES.

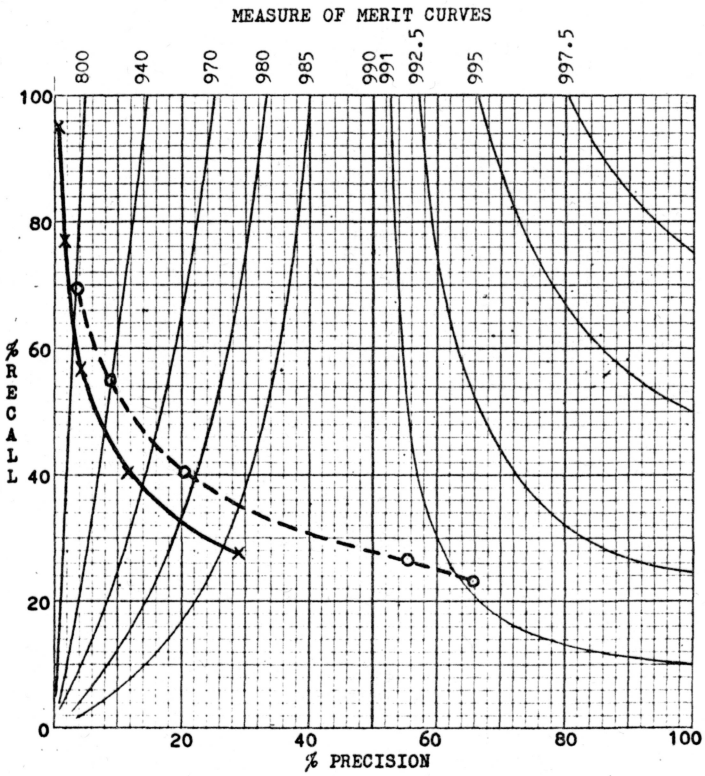


TABLE 16 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING "MEASURE OF MERIT" CURVES.

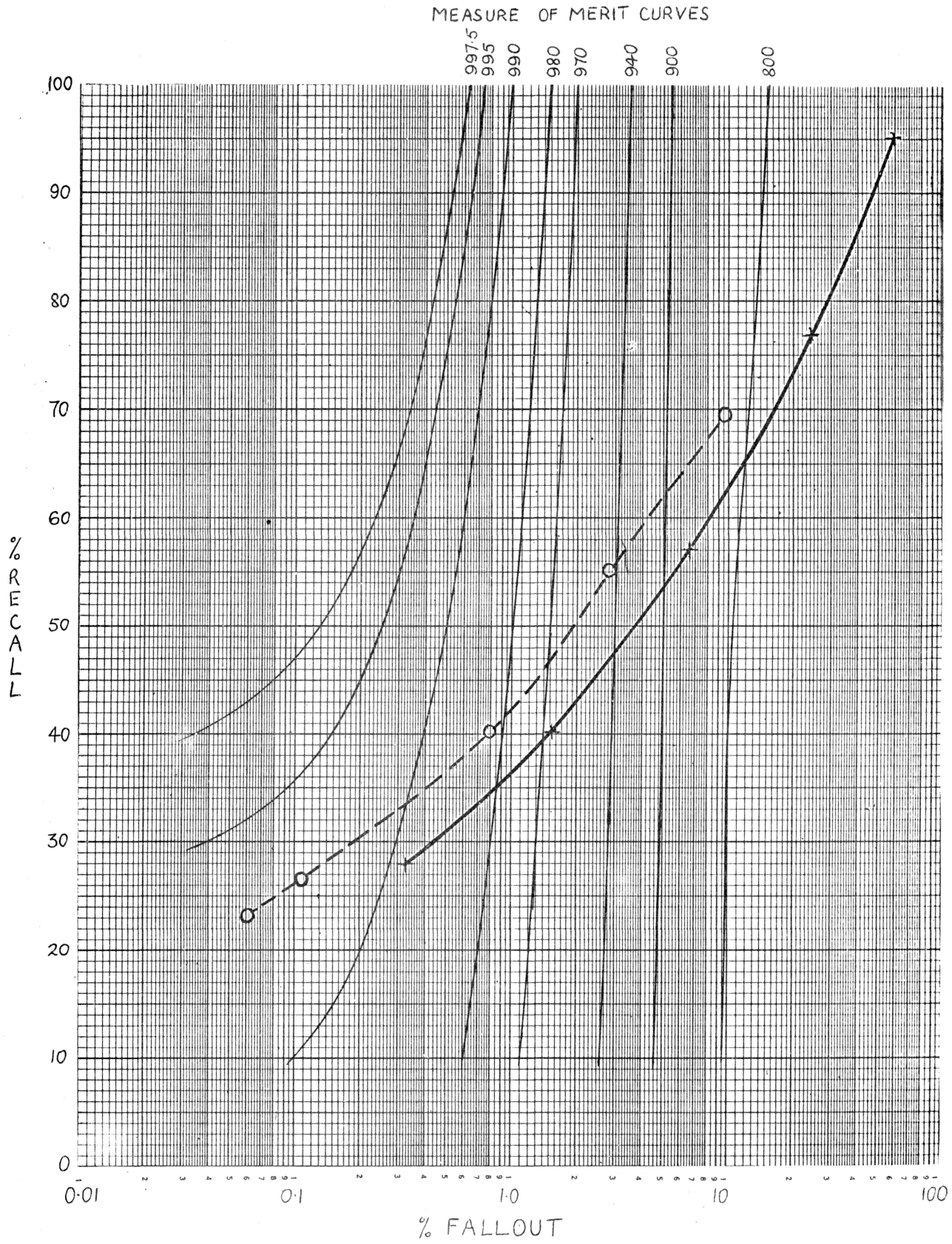


TABLE 17 PLOT OF RECALL AND FALLOUT AS TABLE 6 SHOWING "MEASURE OF MERIT" CURVES.

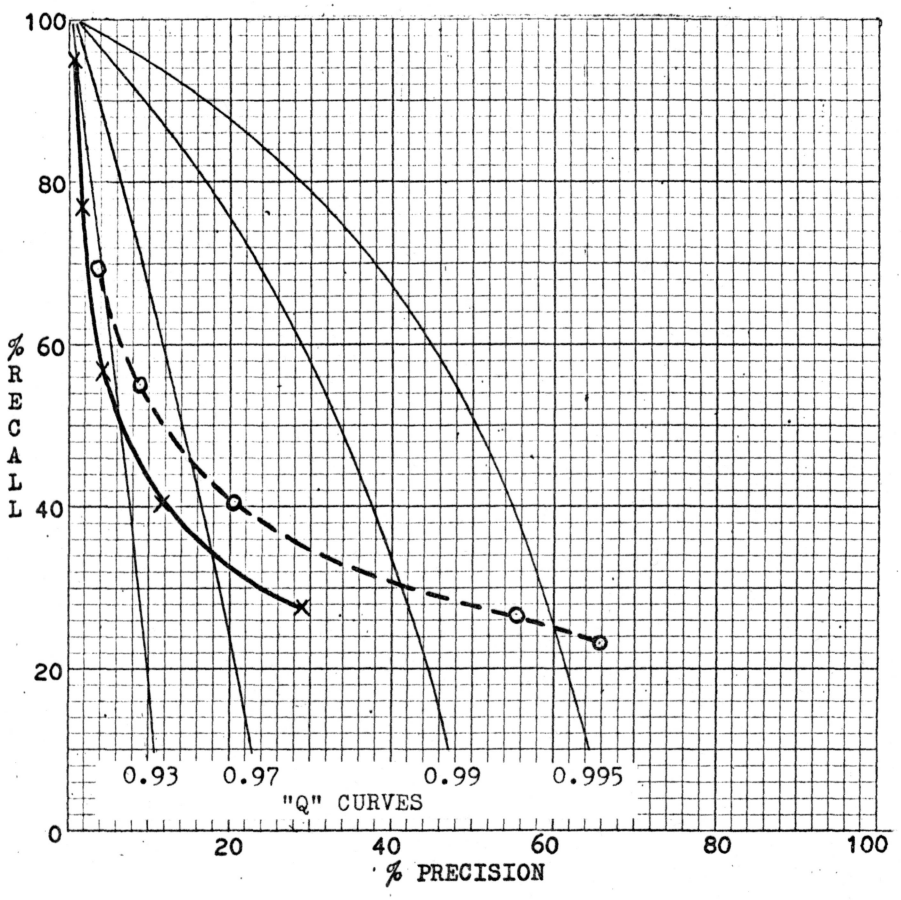


TABLE 18 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING 'Q' CURVES.

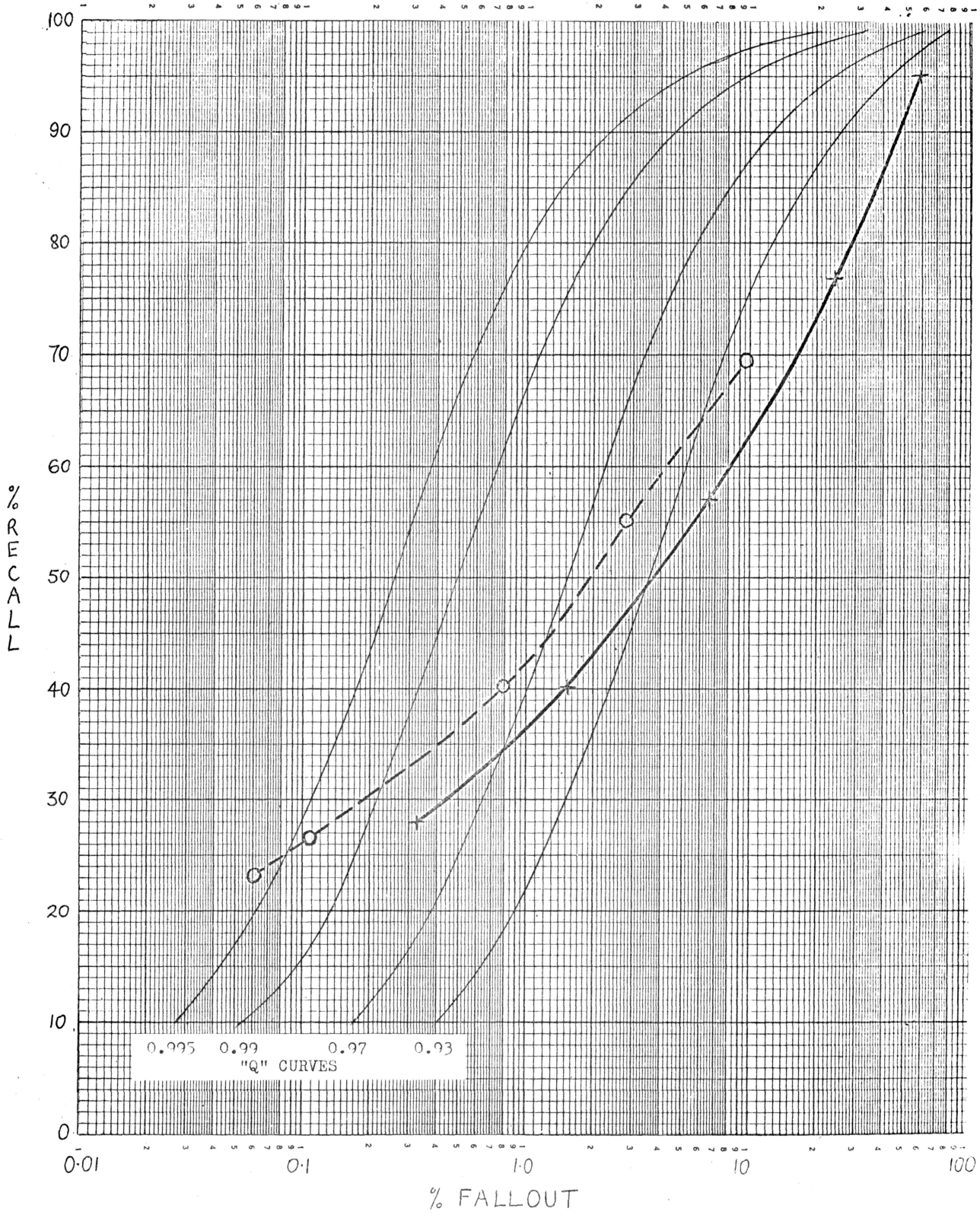


TABLE 19 PLOT OF RECALL AND FALLOUT AS TABLE 6 SHOWING "Q" CURVES

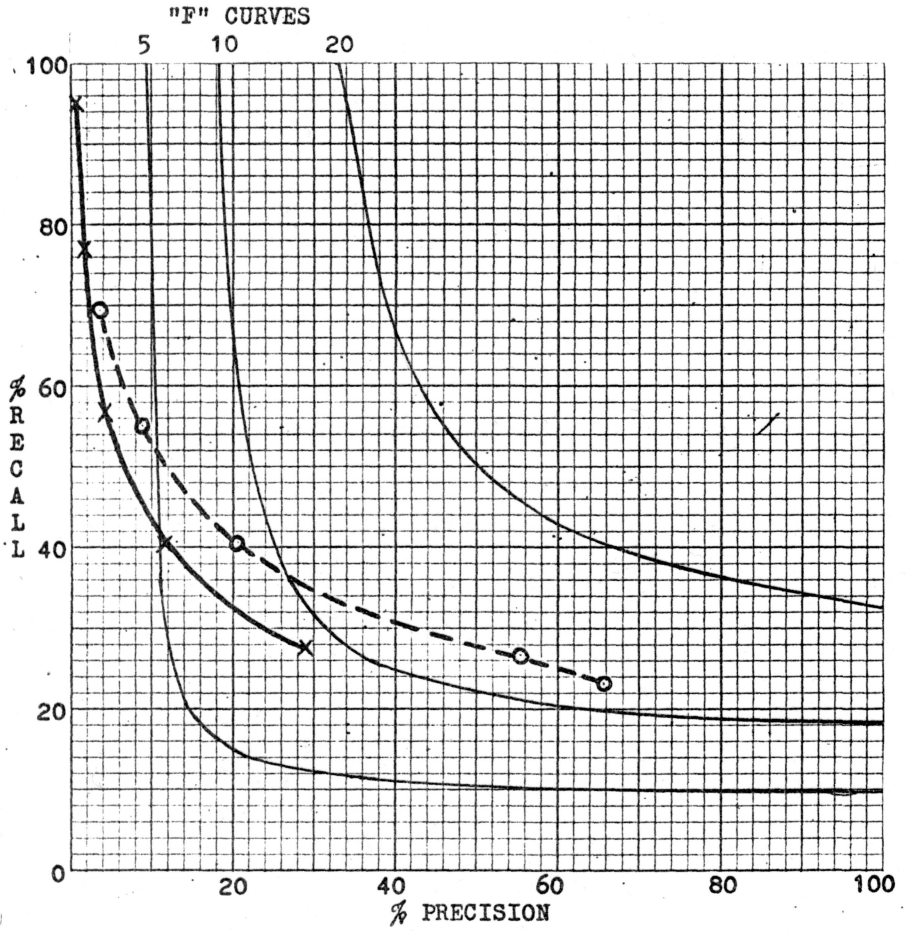


TABLE 20 PLOT OF RECALL AND PRECISION AS TABLE 4 SHOWING 'F' CURVES.

	Infected	Not Infected
Uninoculated	2	5
Inoculated	20	50

TABLE 21 A TYPICAL 2 x 2 CONTINGENCY TABLE

$$R_{norm} \text{ (normalized recall)} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N - n)}$$

where n = number of documents relevant to a question (5)

N = number of documents in the collection (25)

r_i = rank order of i^{th} relevant document in output

CASE 1

Rank order of relevant documents 1, 2, 3, 4, 5

$$R_{norm} = 1 - \frac{\sum 1 + 2 + 3 + 4 + 5 \quad - \quad \sum 1 + 2 + 3 + 4 + 5}{5(25 - 5)} = 1.0$$

CASE 2

Rank order of relevant documents 21, 22, 23, 24, 25

$$R_{norm} = 1 - \frac{\sum 21 + 22 + 23 + 24 + 25 \quad - \quad \sum 1 + 2 + 3 + 4 + 5}{5(25 - 5)} = 0$$

CASE 3

Rank order of relevant documents 3, 5, 6, 11, 16

$$R_{norm} = 1 - \frac{\sum 3 + 5 + 6 + 11 + 16 \quad - \quad \sum 1 + 2 + 3 + 4 + 5}{5(25 - 5)} = 0.74$$

TABLE 22 THE NORMALIZED RECALL MEASURE

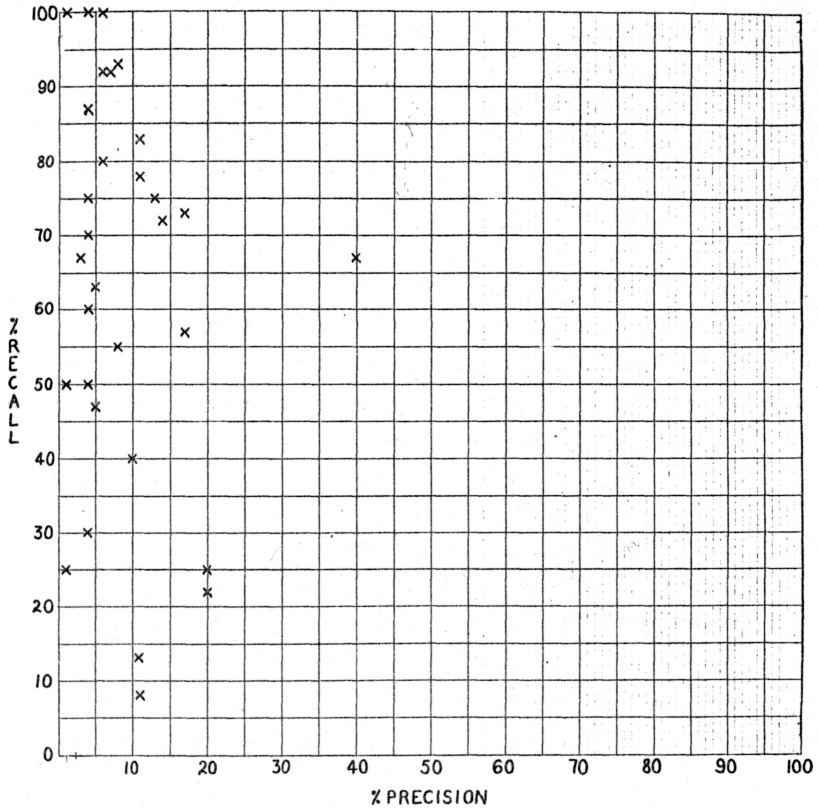


TABLE 23 PLOT OF INDIVIDUAL RECALL AND PRECISION RATIOS OF 31 QUESTIONS SEARCHED AT A CO-ORDINATION LEVEL OF 3 TERMS.

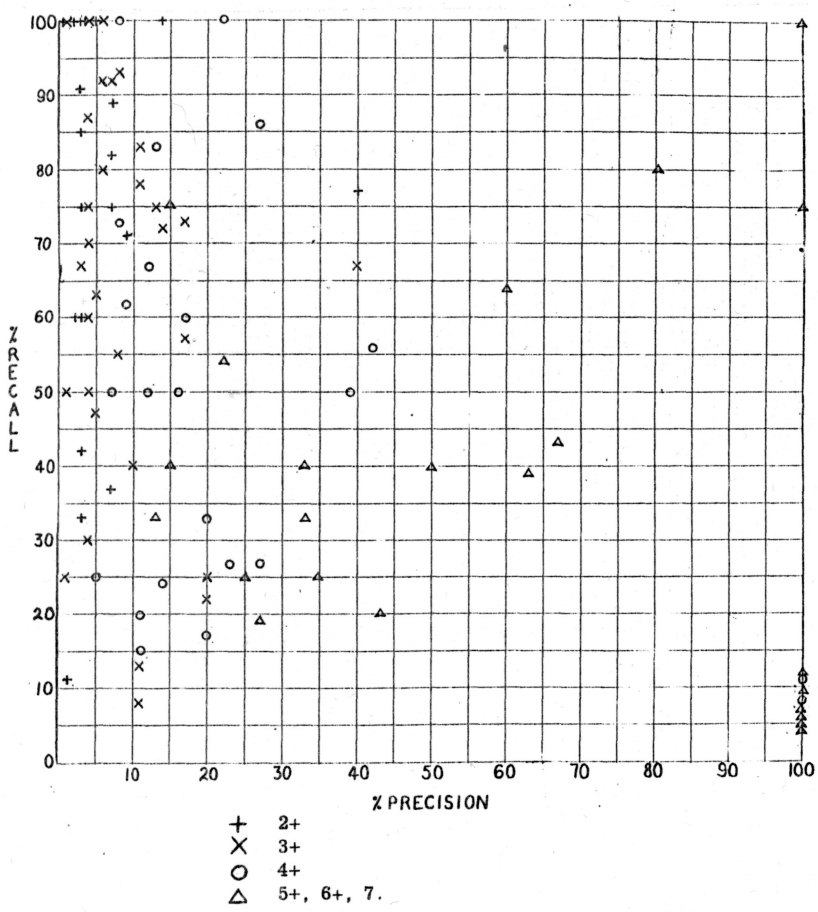


TABLE 24 PLOT OF INDIVIDUAL RECALL AND PRECISION RATIOS OF 35 QUESTIONS SEARCHED AT CO-ORDINATION LEVELS BETWEEN 2 AND 7 TERMS.

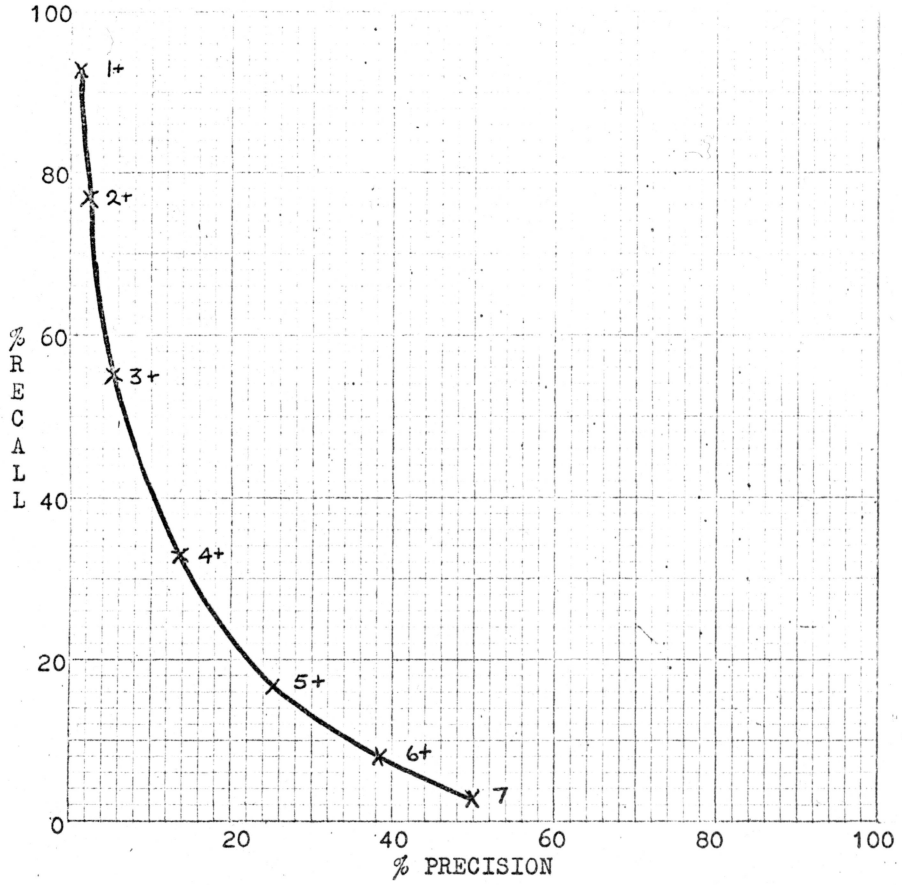


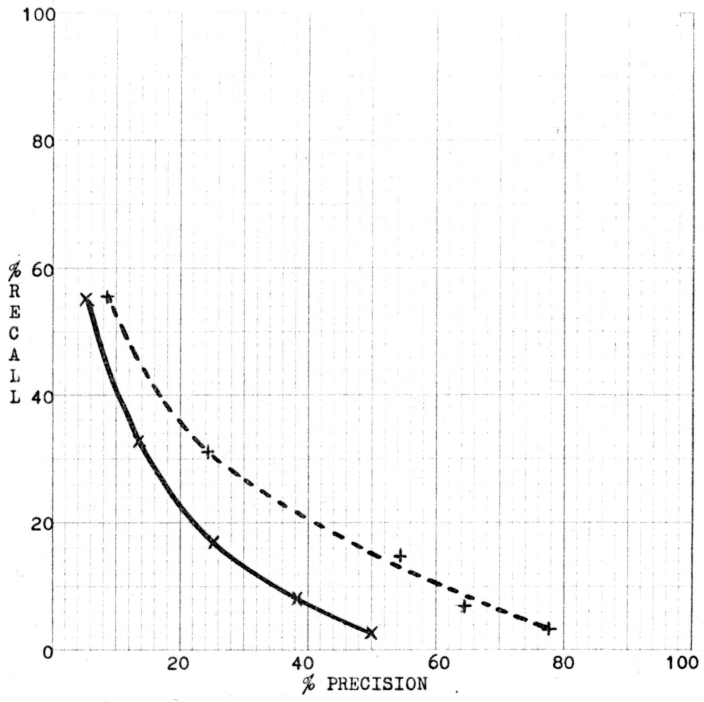
TABLE 25 PLOT OF RECALL AND PRECISION RATIOS, WITH CO-ORDINATION LEVELS INDICATED, THE 35 QUESTIONS TOTALLED BY AVERAGING THE NUMBERS.

Co-ordination Level	Recall Ratio		Fallout Ratio		Precision Ratio	
	A	B	A	B	A	B
3+	54.7%	55.6%	6.054%	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

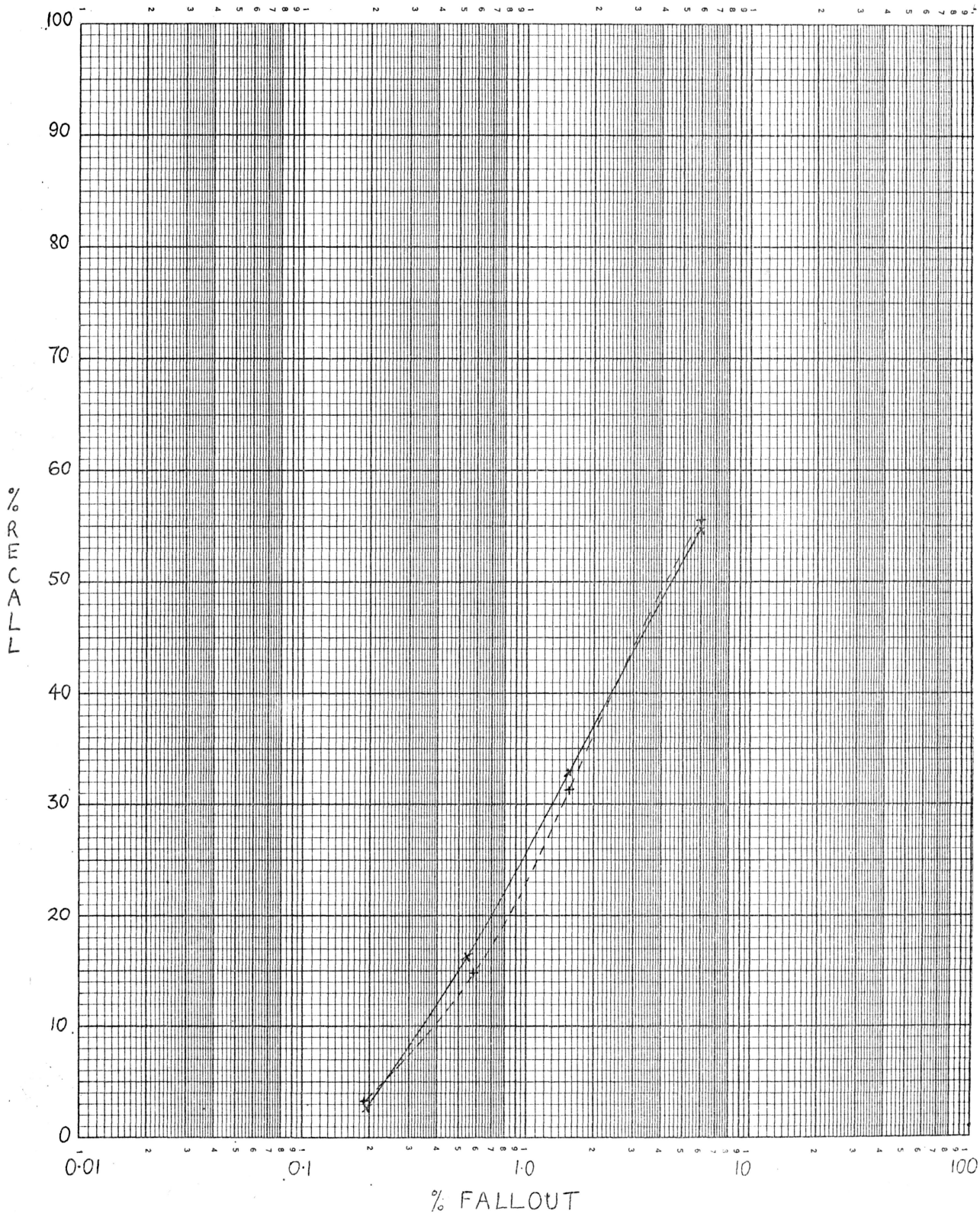
TABLE 26 COMPARISON OF RECALL, FALLOUT AND PRECISION RATIOS WHEN TOTALLED BY AVERAGING THE NUMBERS AND AVERAGING THE RATIOS.



---+--- AVERAGE OF THE RATIOS

—X— AVERAGE OF THE NUMBERS

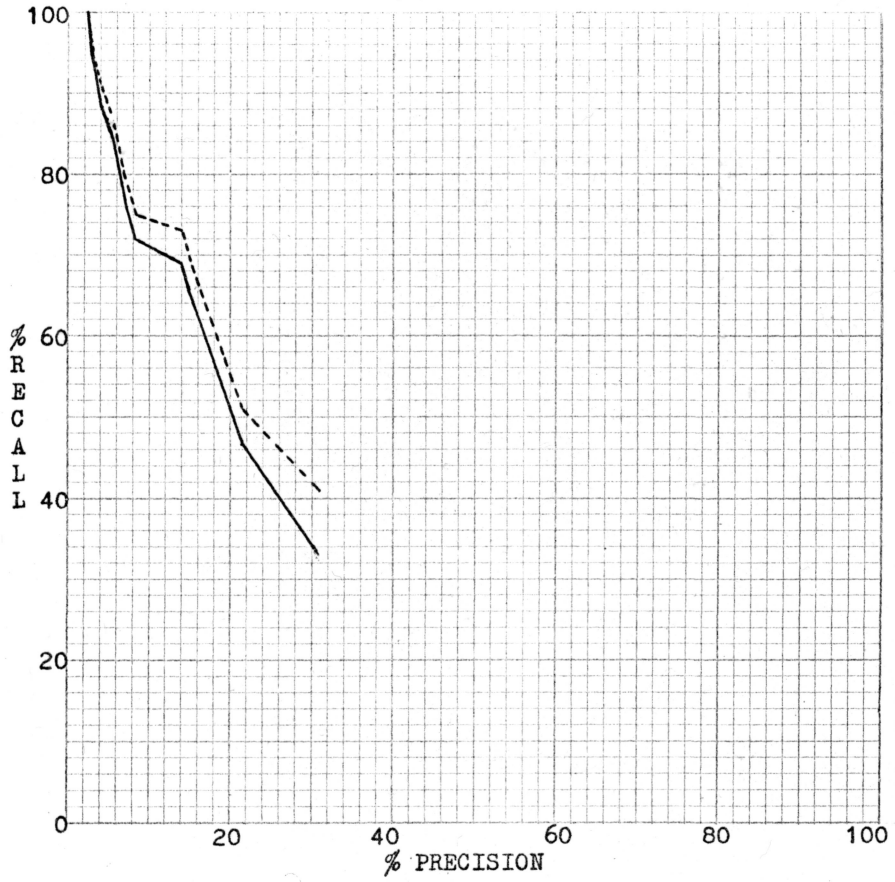
TABLE 27 PLOT OF RECALL AND PRECISION RATIOS FROM TABLE 26 COMPARING TOTALLING BY AVERAGING THE NUMBERS AND AVERAGING THE RATIOS.



---+--- AVERAGE OF THE RATIOS

——x—— AVERAGE OF THE NUMBERS

TABLE 28 PLOT OF RECALL AND FALLOUT RATIOS FROM TABLE 26 COMPARING TOTALLING BY AVERAGING THE NUMBERS AND AVERAGING THE RATIOS.



------ AVERAGE OF THE RATIOS

——— AVERAGE OF THE NUMBERS

TABLE 29 PLOT OF RECALL AND PRECISION RATIOS OF 42 QUESTIONS FROM CRANFIELD SEARCHED BY THE SMART SYSTEM COMPARING TOTALLING BY AVERAGING THE NUMBERS AND AVERAGING THE RATIOS.

FIGURES

		CASE 1				CASE 2			
Question	Relevant Sought	Total Retrieved	Relevant Retrieved	Fallout	Precision	Total Retrieved	Relevant Retrieved	Fallout	Precision
1	10	60	6	5.5%	10%	60	6	5.5%	10%
2	6	100	2	9.9%	2%	100	2	9.9%	2%
3	4	20	3	1.7%	15%	20	3	1.7%	15%
4	20	100	10	9.2%	10%	100	10	9.2%	10%
5	5	50	0	5.5%	0%	0	0	0%	0%

RESULTS

	Fallout		Precision	
Averaging Method	Case 1	Case 2	Case 1	Case 2
Average of Numbers	6.2%	5.3%	6.4%	7.5%
Average of Ratios	6.4%	5.3%	7.4%*	7.4%

* Adjusted Precision Ratio = 6.2%

TABLE 30 TWO CASES OF PERFORMANCE OF FALLOUT AND PRECISION RATIOS SHOWING AVERAGING METHODS AND THE ADJUSTED PRECISION RATIO.

Co-ordination Level	Index Language 4a		Index Language 6a	
	R	N-R	R	N-R
1+	13	*	13	*
2+	13	*	13	*
3+	12	180	13	400
4+	8	88	8	146
5+	7	31	7	46
6+	5	5	5	8
7	0	0	1	0

R Number of relevant retrieved
N-R Number of non relevant retrieved
* Figures not obtained

TABLE 31 PERFORMANCE RESULTS FOR QUESTION 145 FOR TWO INDEX LANGUAGES AND SEVEN CO-ORDINATION LEVELS

Number of retrieving terms	Number of starting 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

TABLE 32 DISTRIBUTION OF THE 221 QUESTIONS BY STARTING TERMS AND RETRIEVING TERMS, IN ONE PARTICULAR TEST.

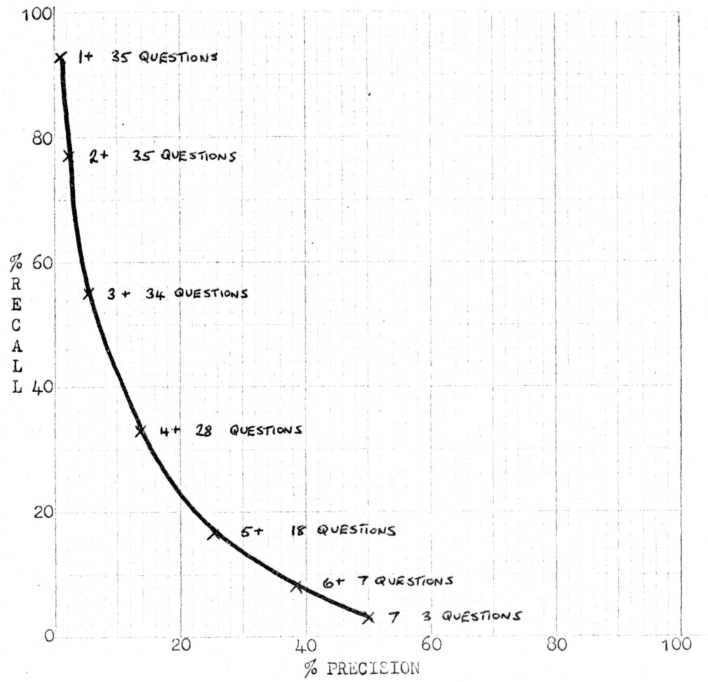


TABLE 33 PLOT OF RECALL AND PRECISION RATIOS SHOWING REDUCED SAMPLE SIZE AT HIGH CO-ORDINATION LEVELS. THE 35 QUESTIONS ARE A HOMOGENEOUS SEVEN STARTING TERM SUBSET, TOTALLED BY STRICT CO-ORDINATION LEVELS, AVERAGING THE NUMBERS.

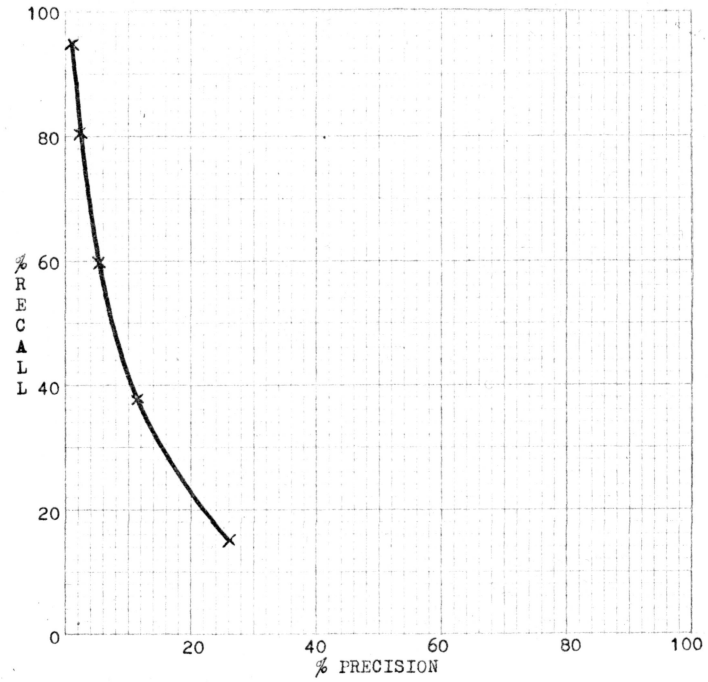


TABLE 34 PLOT OF RECALL AND PRECISION BY CO-ORDINATION LEVELS OF 45 QUESTIONS IN THE HOMOGENEOUS FIVE RETRIEVING TERM SUBSET, TOTALLED BY STRICT CO-ORDINATION LEVELS, AVERAGING THE NUMBERS.

Question Number	Number of Relevant	The Ranks of the Relevant Documents
230	7	1, 3, 7, 17, 66, 80, 190.
250	8	1, 2, 3, 6, 7, 14, 16, 171.
261	4	1, 2, 3, 5.
264	2	1, 2.
266	5	10, 12, 13, 27, 72.

TABLE 35 EXAMPLE OF SMART RESULTS FOR FIVE QUESTIONS.

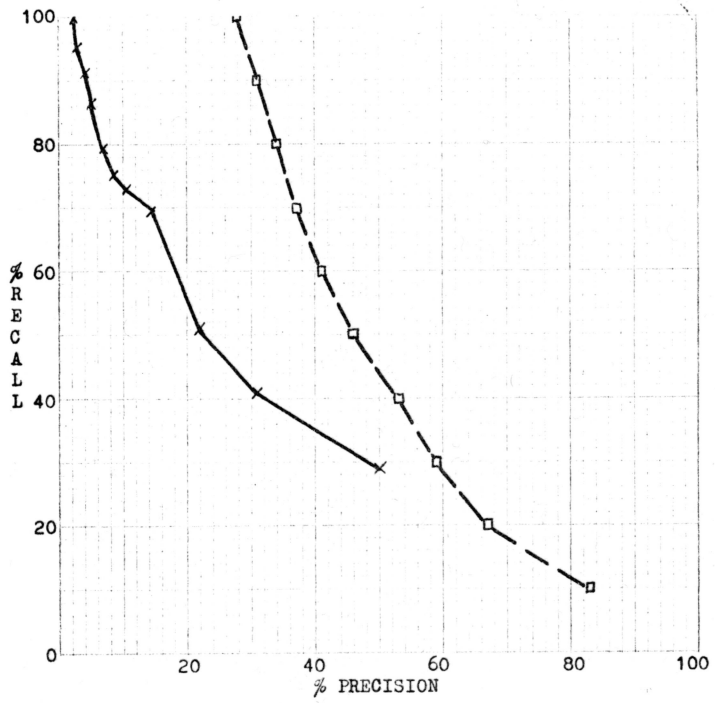
Question Number	Precision Ratio at Final Cutoff
230	$\frac{7}{190} \times 100 = 3.7\%$
250	$\frac{8}{171} \times 100 = 4.7\%$
261	$\frac{4}{5} \times 100 = 80\%$
264	$\frac{2}{2} \times 100 = 100\%$
266	$\frac{5}{72} \times 100 = 6.9\%$

Precision ratio (average of ratios) = 39.1%

TABLE 36 PRECISION RATIOS FOR THE FIVE QUESTIONS IN TABLE 35 CALCULATED BY THE RELEVANT DOCUMENTS CUTOFF AT 100% RECALL.

CUTOFF APPLIED AFTER n DOCUMENTS RETRIEVED	RELEVANT RETRIEVED INDICATED BY A CROSS	RECALL RATIO	PRECISION RATIO
n = 5	X X	$\frac{2}{7}$ x 100 = 28.6%	$\frac{2}{5}$ x 100 = 40.0%
n = 10	X	$\frac{3}{7}$ x 100 = 42.9%	$\frac{3}{10}$ x 100 = 30.0%
n = 20	X	$\frac{4}{7}$ x 100 = 57.1%	$\frac{4}{20}$ x 100 = 20.0%
n = 30		$\frac{4}{7}$ x 100 = 57.1%	$\frac{4}{30}$ x 100 = 13.3%
n = 40		$\frac{4}{7}$ x 100 = 57.1%	$\frac{4}{40}$ x 100 = 10.0%
n = 50		$\frac{4}{7}$ x 100 = 57.1%	$\frac{4}{50}$ x 100 = 8.0%
n = 60		$\frac{4}{7}$ x 100 = 57.1%	$\frac{4}{60}$ x 100 = 6.7%
n = 70	X	$\frac{5}{7}$ x 100 = 71.4%	$\frac{5}{70}$ x 100 = 7.1%
n = 100	X	$\frac{6}{7}$ x 100 = 85.7%	$\frac{6}{100}$ x 100 = 6.0%
n = 150		$\frac{6}{7}$ x 100 = 85.7%	$\frac{6}{150}$ x 100 = 4.0%
n = 200	X	$\frac{7}{7}$ x 100 = 100%	$\frac{7}{200}$ x 100 = 3.5%

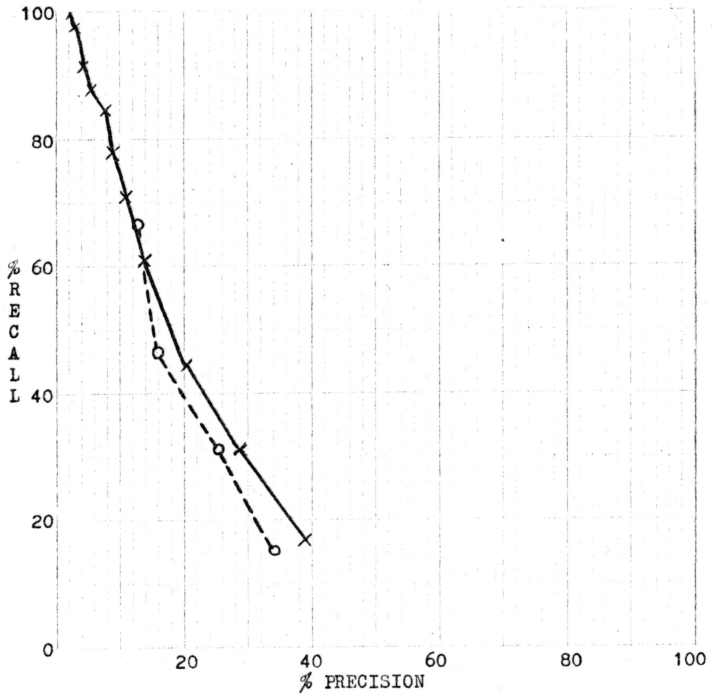
TABLE 37 RECALL AND PRECISION RATIOS FOR QUESTION 230 FROM TABLE 35, CALCULATED BY THE DOCUMENT OUTPUT CUTOFF AT ELEVEN CUTOFF POINTS.



RELEVANT DOCUMENTS CUTOFF

DOCUMENT OUTPUT CUTOFF

TABLE 38 PLOT OF RECALL AND PRECISION RATIOS COMPARING THE RELEVANT DOCUMENTS AND DOCUMENT OUTPUT CUTOFFS, AVERAGING THE RATIOS.



DOCUMENT OUTPUT CUTOFF

CORRELATION COEFFICIENT CUTOFF

TABLE 39 PLOT OF RECALL AND PRECISION RATIOS COMPARING THE DOCUMENT OUTPUT AND CORRELATION COEFFICIENT CUTOFFS, AVERAGING THE NUMBERS.

	5+		4+		3+		2+		1+	
	R	N-R	R	N-R	R	N-R	R	N-R	R	N-R
1. Cumulated Figures	1	2	3	10	5	29	6	75	6	110
2. New Total Retrieved	3		10		21		48		34	
3. New Relevant Retrieved	1		2		2		1		-	
4. Simulated Ranks	1-3		4-13		14-34		35-82		83-116	
5. Ranks of Relevant	2		8,9		23,24		58		-	

TABLE 40 RESULTS OF A SEARCH QUESTION EMPLOYING FIVE CO-ORDINATION LEVELS CONVERTED TO SIMULATED RANKED OUTPUT