THE

Journal of Documentation

VOLUME 25 NUMBER 1 MARCH 1969

THE PARAMETRIC DESCRIPTION OF RETRIEVAL TESTS*

PART I: THE BASIC PARAMETERS†

S. E. ROBERTSON

Research Department, Aslib

Some parameters and techniques in use for describing the results of tests on *IR* systems are analysed. Several considerations outside the scope of the usual 2 × 2 table are relevant to the choice of parameters. In particular, a variable which produces a 'performance curve' of a system corresponds to an extension of the 2 × 2 table. Also, the statistical relationships between parameters are all-important. It is considered that precision is not such a useful measure of performance (in conjunction with recall) as fallout. A more powerful alternative to Cleverdon's 'inevitable inverse relationship between recall and precision' is proposed and justified, namely that the recall-fallout graph is convex.

1. INTRODUCTION

A FAIR AMOUNT of testing of information retrieval systems has now been performed; the output of literature on the subject increases year by year. But it is clear that many of the results so far produced are not really useful: the methodology is not sufficiently advanced for the conclusions to be of general application. The output of literature purely on methodology also grows (for full references see Bourne[1] and Rees[2]). Rees[3] has said of the Cleverdon–WRU test: 'The great value of Cleverdon's contribution lies in the area of test methodology rather than in the experimental results.' The same could probably be said of a large number of tests, if indeed they have any value at all.

* This paper is based on a thesis presented for a Master's degree in Information Science at the City University, London.

† Part 2: 'Overall measures' is due to appear in the next issue.

All of these tests involve some numerical description of the way the system operates; conclusions are then drawn on the basis of the figures. Clearly, the manipulation of the figures that comes between the primary results and the conclusions is an important part of the methodology. This is the part with which I am concerned in this paper.

The parameters now in use seem to have evolved in a somewhat haphazard manner. Now that so many results of tests have been published, it is possible to get a fairly good idea of the relationships between the variables and to analyse what is required of the parameters. It appears that there are many desirable and undesirable properties to be considered; some of these have been considered by other authors, but mostly in isolation. In this paper I attempt to give a reasonably unified analysis of these requirements, and to apply them to various parameters in use or proposed. These comments on known parameters are of immediate interest, but I hope the requirements themselves and the arguments which lead up to them will prove useful in other situations where new parameters are used. An example of what I have in mind is the test by King *et al.*[4] on accuracy of indexing; although the situation is rather different from the usual IR test, there are distinct similarities in the forms of the parameters used.

Appendix A is a summary of the parameters actually used in published tests results. This list is taken for the most part from the two excellent reviews of the field (Bourne,[1] Rees[2]); readers are referred back to these reviews for full references concerning the tests.

2. NOTATION AND TERMINOLOGY

The notation and terminology defined in this section are used throughout the paper. Any other symbols used are defined at the time of use.

The primary parameters, taken directly from the results of the experiment, are represented as follows: R is the number of relevant documents that are retrieved, C is the total number of relevant documents, L the total number retrieved, and N the size of the collection. These quantities and others that are calculated from them are normally given in the form of the 2×2 contingency table of Table 2.1.

TABLE 2.1

	Relevant	*Not relevant*	(*totals*)
Retrieved	R	$L-R$	L
Not retrieved	$C-R$	$N-C-L+R$	$N-L$
(totals)	C	$N-C$	N

The same table with different notation is given in Table 2.2. This notation is occasionally used when it is necessary to stress the mathematical symmetry of the table.

TABLE 2.2

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

The secondary (derived) parameters used are the well-known ones: recall (proportion of relevant documents that are retrieved), precision (proportion of retrieved documents that are relevant), fallout (proportion of non-relevant documents that are retrieved), and generality (proportion of documents in the whole collection that are relevant). Symbols for these parameters, formulae, and alternative names are given in Table 2.3.

TABLE 2.3

Name	*Symbol*	*Formulae*	*Other names*
Recall	M	$\frac{R}{C}=\frac{a}{a+c}$	Sensitivity (Goffman and Newill[5]) Conditional probability of a hit (Swets[6,7])
Precision	P	$\frac{R}{L}=\frac{a}{a+b}$	Relevance
Fallout	F	$\frac{L-R}{N-C}=\frac{b}{b+d}$	Discard (Farradane *et al*[8,9]) Conditional probability of a false drop (Swets)
Generality	G	$\frac{C}{N}=\frac{a+c}{a+b+c+d}$	

I think the only non-standard notation is the symbol M for recall (R is already in use). I normally (in equations etc.) consider these parameters to be pure ratios with maximum values of 1; when quoting actual figures, however, it is more convenient to use percentages. Table 2.4 gives some similar parameters.

TABLE 2.4

Formula	*Description*	*Names*
$1-P$	Proportion of retrieved documents that are not relevant	Noise factor
$1-M$	Proportion of relevant documents that are not retrieved	Conditional probability of a miss (Swets)
$1-F$	Proportion of non-relevant documents that are not retrieved	Specificity (Goffman and Newill) Conditional probability of a correct rejection (Swets)

3. THE 2 × 2 CONTINGENCY TABLE

The earliest tests on *IR* systems (see Appendix A) were normally conducted on the following basis. A source document was given to a user, who thought up a question to which the source document was an answer; the question was put to the system, to see whether it would produce the source document. The parameter normally computed from these tests was the proportion of source documents (of different questions) retrieved by the system. This parameter is often equated with the recall ratio; Cleverdon[10] gives an argument justifying this assumption. The argument is a somewhat dubious combination of statistics and common sense, which would be extremely difficult to formalize, and which is scorned by some people (e.g. Farradane *et al*[8]). But the most serious objection to the use of this parameter seems to be that it equates *relevant* documents with documents *that could have inspired the question*. Since the definition of relevance is now considered to be one of the most difficult problems in the evaluation of *IR* systems (see e.g. Rees[11]), this is highly unsatisfactory. However, the inadequacies of this parameter have long been recognized, and it is seldom used nowadays, so I shall go no further into the problem.

Later tests of retrieval systems have almost all been based on a reversal of the above procedure: a question is posed, and some or all of the documents in the collection are assessed for relevance to this question by the questioner or a subject specialist. Then the results of the test are described in the following terms. The relevance property separates the collection into two parts: relevant and non-relevant documents. The *IR* system also separates the collection into two parts: documents that are retrieved and those that are not. The resulting numbers of documents are put in the form of a 2 × 2 (contingency) table, as in Table 2.1. A variety of parameters are derived from these quantities; some are given in Table 2.3.

Fairthorne[12] says of the 2 × 2 table: 'Such a table is completely determined by any four quantities associated with it, if they are independent.' This is not quite true—in fact the four parameters he proposes later in the paper do not determine it completely (see §4). But the table has four degrees of freedom—that is, four parameters are necessary and can be sufficient to determine it. The bottom line of the table (N, C, $N-C$) has two degrees of freedom, and is normally specified by N and G (the generality). Since it is independent of the actual retrieval operation, it can be said to describe the experimental conditions rather than the actual results. Thus one should specify the experimental conditions when reporting the results of an experiment; but normally one hopes to get results that are as far as possible independent of these conditions. I return to this point in §9.

The rest of the table is then described by two further parameters—usually M and P or M and F are chosen (M=recall, P=precision, F=fallout). Swets[6,7] and Rees[11] argue that M and F 'contain all the information in the

table' and are therefore a better pair to use; this is not true, they *use* all the information, which is different. Provided that the experimental conditions are given as well, either pair specifies the table completely; neither pair does alone. There is, however, one qualification to this statement; in the exceptional case where $R=O$, M and P (with G and N) do not specify the table completely: the distribution of non-relevant material between sets b and d is indeterminate. There are no such exceptional cases if M and F are used.

4. FAIRTHORNE'S PARAMETERS

Fairthorne[12] is concerned with the basic symmetry of the 2×2 table. He says that a parameter measuring how a collection is separated into two parts should not depend on which way round the parts are—i.e. parameters should be invariant (except for a possible change of sign) under interchange of the rows or columns of the 2×2 table. So he proposes the following parameters: N, $C(N-C)$, $L(N-L)$ and $RN-CL$. He has four in order to specify the table completely; but as I mentioned in the last section, these four do not in fact do so. For example,

N	C	L	R		N	$C(N-C)$	$L(N-L)$	$RN-CL$
8	4	3	2	both give	8	16	15	4
8	4	5	3					

so that set of values for his parameters does not determine the table completely.

He then says that if ratios are sufficient the number of parameters can be reduced to three. (I return to the problem of whether ratios are in fact sufficient in §8.) The three ratios he chooses are: $C(N-C)/\frac{1}{4}N^2$, $L(N-L)/\frac{1}{4}N^2$, and $(RN-CL)/C(N-C)$ *or* $(RN-CL)/L(N-L)$, the 'or' being exclusive. 'Given one of the alternative expressions, we can obtain the other by multiplying, or dividing, by the ratio of the first quoted ratios above,'

$$\text{i.e. } \frac{C(N-C)}{\frac{1}{4}N^2}\times\frac{(RN-CL)}{C(N-C)}=\frac{L(N-L)}{\frac{1}{4}N^2}\times\frac{(RN-CL)}{L(N-L)}$$

'Thus tests are not completely described in terms of the alternatives and the other. . .' This is not quite true; *any* three of these four parameters serve equally well to describe the tests.

He points out that if C and L are small relative to N, the four ratios approximate to C/N, L/N, R/C and R/L. He makes the observation, presumably on the basis of the above quoted statements, that R/C and R/L (recall and precision) are 'not mathematically independent'. This is ambiguous, and is only true in a rather restricted sense: see §6.

He claims at the beginning of the paper: 'What [the paper] attempts is to derive from documentary, rather than mathematical, considerations those parameters that display fundamental retrieval functions.' In fact he derives

his parameters from the purely mathematical consideration of the symmetry of the 2×2 table; he subsequently attempts to interpret the parameters in documentary terms, not very convincingly in my opinion.

However, I wish to reconsider his requirements of invariance for these parameters. It seems to me that he has applied the rules too rigidly, and thus has been led to parameters that are too complicated. I shall therefore relax the rules a little. He requires that his parameters should be invariant except for a possible change of sign under the said transformations (interchange of the rows or of the columns of the 2×2 table). To put this in a symbolic form, if x is a parameter and t one of the said transformations, he requires that x satisfy one of the following:

$$t(x)=x \qquad (1)$$
$$t(x)=-x \qquad (2)$$

I shall also allow the following possibilities:

$$t(x)=1-x \qquad (3)$$
$$t(x)=y \text{ and } t(y)=x \qquad (4)$$

My requirement (3) is a slight extension of his requirement (2); one can always return to (2) by replacing x by $x-\frac{1}{2}$. My requirement (4) says that one can consider a pair of parameters together rather than separately—the pair can be invariant under the transformation even if the individual parameters are not. This means that if the parameters are plotted on a graph, the shape of the plot is invariant under the transformation—the graph is merely reflected in the diagonal.

To take a concrete example, consider the set of parameters N, G, M, and F. If t_1 is the interchange of rows and t_2 the interchange of columns, we have:

$$t_1(N)=N \quad t_1(G)=G, \quad t_1(M)=1-M \quad t_1(F)=1-F;$$
$$t_2(N)=N \quad t_2(G)=1-G, \quad t_2(M)=F \quad t_2(F)=M.$$

Thus N, G, and the pair (M, F) satisfy my requirements, nothing as complicated as Fairthorne's parameters is required. Though several other sets of parameters satisfy these requirements, a recall-precision graph does not.

5. EXTENSIONS OF THE 2×2 TABLE

The 2×2 table has become such a well-known representation of the results, that it is often not realized that in many situations it is a simplification—an extension of the table is implicitly being used. The first explicit form of this extension was considered by Swets;[6] his model of the retrieval process can be described as follows. The system attributes to each document a value of a (linear) variable z, which describes how well the document specification fits the question specification. An actual retrieval process involves a choice of cut-off value z_0; all documents with $z \geqslant z_0$ are retrieved, those with $z < z_0$

are not. Swets then considers the distributions of relevant and non-relevant documents with respect to z. I consider some implications of this model in §7 and in Part 2; but it is easy to show that it is equivalent to an extension of the 2×2 table. Consider the Cranfield II experiment,[13,14] where there is a specific interpretation of the variable z: the 'level of co-ordination'. This is a discrete variable, taking values 1, 2, ..., n. The retrieval operation should now be represented by the $2 \times (n+1)$ table of Table 5.1.

TABLE 5.1

z	*Relevant*	*Not relevant*	*(totals)*
	R_n	$L_n - R_n$	L_n
n			
	R_{n-1}	$L_{n-1} - R_{n-1}$	L_{n-1}
$n-1$			
.	.	.	.
.	.	.	.
.	.	.	.
2			
	R_1	$L_1 - R_1$	L_1
1			
	$C - \Sigma R_i$	$N - C - \Sigma(L_i - R_i)$	$N - \Sigma L_i$
(totals)	C	$N - C$	N

(*sums are over the range* $i = 1, 2, \ldots, n$)

Here L_j is the number of documents retrieved at level of co-ordination exactly j. So at level of co-ordination $j+$, the process is represented by the 2×2 table of Table 5.2, derived by contracting Table 5.1.

TABLE 5.2

	Relevant	*Not relevant*	*(totals)*
Retrieved	ΣR_i	$\Sigma(L_i - R_i)$	ΣL_i
Not retrieved	$C - \Sigma R_i$	$N - C - \Sigma(L_i - R_i)$	$N - \Sigma L_i$
(totals)	C	$N - C$	N

(*sums are over the range* $i = j, j+1, \ldots, n$)

In the case of the (more general) continuous variable z, the full table cannot be displayed explicitly as there are infinitely many cut-off points; but the principle is the same.

Several points can be made from a consideration of this idea:

a All parameters in common use can be derived from the contracted form of

the table (Table 5.2); no-one has used parameters that can only be taken from Table 5.1. For example, it might be worth considering R_j/L_j, by analogy with $P=\Sigma R_i/\Sigma L_i$ (summed over $i=j, j+1, \ldots, n$). This would be a more sensitive measure than P.

b This expression for P stresses the fact that the dependence of precision on level of co-ordination is far from simple: both numerator and denominator increase with decreasing j. By contrast recall ($=\Sigma R_i/C$, $i=j, j+1, \ldots, n$) and fallout ($\Sigma(L_i-R_i)/(N-C)$, $i=j, j+1, \ldots, n$) have a much more straightforward dependence on level of co-ordination. I return to this point in §10.

c The only convenient method that has emerged for dealing with this situation is to take two parameters from the contracted table, and plot one against the other; the successive contracted tables defined by successive cut-off points give a series of points on this graph, which are joined to form the 'performance curve' of the system. Clearly this can only be done if two parameters are sufficient to represent the system, i.e. if one makes use of the fact that G and N can be regarded as experimental conditions rather than results. I therefore stress this fact again. A set of three or four interconnected parameters (like Fairthorne's) cannot be used for this purpose.

d Several other variables connected with *IR* systems can be regarded in the same way. For example, the effect of introducing an index language device, such as those studied in Cranfield II, can be shown by a 2×3 table (at one level of co-ordination).

There is one variable which was studied in Cranfield II, which cannot be represented in this fashion. This is the 'relevance level': documents were assessed at various levels rather than just 'relevant' or 'non-relevant', and the effect of including different levels in set C was analysed. It is fairly obvious, by analogy with the above vertical extension of the 2×2 table, that this variable corresponds to a horizontal extension in the same way. Thus the comment *b* above must be reversed to apply to this extension: P is simply related to the variable 'relevance level', but its relationship to M or F is more complicated. Since several authors have considered such a variable but not used it much (see Appendix A), I follow up this idea in §7.

Probably the treatment of any of these variables as linear is a simplification. Farradane *et al*[9] have three distinct ways of increasing recall (conceptual browsing, generic browsing, and using a shortened version of the question), which clearly cannot all be described in terms of one linear variable. Similarly, there are probably several distinct ways in which a paper can be relevant to a question—for example, neither paper A nor paper B might answer a question, but the two taken together might do so. However, the use of these variables represents a great step forward from the simple 2×2 table.

6. RECALL AND PRECISION

By far the most commonly used parameters (see Appendix A) are recall (M) and precision (P). They are also often criticized; in this and later sections I shall attempt to analyse some of the criticisms and some of the advantages claimed for them.

Fairthorne[12] says: '(Their use) misleads . . . if displayed as a plot of R/C against R/L, because this removes R from the scene, leaving only a curve showing the relation between the reciprocals of L and C, the scale of the diagram being stretched locally in proportion to R in a way that does not allow recovery of the value of R.' In the first place, of course one cannot recover the value of R; neither can one recover the values of C or L. The point of using ratios is to give the results more general application, to make it possible to interpret them in conditions other than those of the particular experiment. But if the conditions of the experiment are given as they should be, including the values of G and N, then one can recover the values of R, C, and L.

In the second place, the fact that the value of R cannot be recovered directly from the graph does not imply that R has been 'removed from the scene'. On the contrary, R is very much in evidence: if one fixes C and L but increases R, the whole curve moves away from the origin. This is best indicated mathematically by expressing the same graph in polar co-ordinates: $r=R\sqrt{(C^2+L^2)}/CL$, $\theta=\tan^{-1}(L/C)$, so R appears only in the expression for r (distance from origin), not in that for θ (elevation from origin). Indeed, Fairthorne's argument would imply that it was useless to plot any two variables in polar co-ordinates, since the variable r would be 'removed from the scene'.

It is however true that the relationship between these variables might make it difficult to interpret a graph of one against the other. For example, if a graph of recall against fallout shows any definable pattern, one can reasonably attribute this pattern to the effect of the system. In the case of recall and precision, such a deduction would not necessarily be valid: the fact that if precision is zero so is recall is a mathematical property of the parameters, not a property of the system. This is presumably what Fairthorne means when he says that recall and precision are 'not mathematically independent' (see §4). However, the usual interpretation of such a statement would be stronger: that given one of the parameters the other can be calculated. This is clearly false in general.

Fairthorne makes a further observation: if there are no relevant documents to a question, $C=0$ and therefore $R=0$, and the recall ratio is not defined. Similarly if $L=0$, precision is not defined. Such difficulties are almost bound to occur if ratios are used—and there is no hope of comparing results if ratios are not used. Certainly Fairthorne has not solved the problem with his own parameters. But I would like to distinguish between the

two cases. The case $C=0$ refers to a particular type of question, and does not depend on the test results. If such questions are used to test systems, they can be treated separately from the rest. But the case $L=0$ might occur in answer to any question; to leave such cases out of the averages would be to distort the results. So, if precision is used, a method must be found to deal with such cases. There are such methods, but they can be fairly elaborate, as Keen's consideration of them shows.[15] They can also arouse controversy: see for example the comments by Farradane *et al*[8] on Cleverdon's practice of extrapolating the recall-precision curve to $M=0$, $P=1$. One of the reasons Cleverdon and Keen[15] chose the 'micro' method of averaging (see §8) was the difficulty in dealing with such cases. Therefore it would seem that on this count recall and fallout are the easier pair to use. A similar case could of course occur for fallout: $N-C=0$ (but it is clearly unlikely, it means that all the documents in the collection are relevant to the question); but the same remarks as for recall apply.

Farradane *et al*[8] make the following criticism of the use of recall and precision. In connection with their measure of effectiveness $Q=(ad-bc)/(ad+bc)$, they give the equation $P=\dfrac{1+Q-2QM}{2Q+(1-Q)/G-2QM}$.

They then say: 'the precision ratio can be seen to be a function of the recall ratio if, and only if, Q and G are constant. Theoretical curves for different values of Q (with which practical results could be compared) can be drawn on the recall-precision graph only if G is constant.' They take this as an argument against the use of recall and precision; it could equally well *a priori* be taken as an argument against the use of Q as a measure of effectiveness, since the form of the equation depends on the definition of Q; another effectiveness measure (and there are plenty) would give a different equation. The problem of the *empirical* dependence of parameters on generality is a much more complicated one to which I return in §9.

While I do not subscribe to all the criticisms of recall and precision neither do I agree with all the interpretations of their value. Rocchio[16] says: 'Clearly, recall . . . is a measure of the inclusiveness of the set L with respect to the set C, while relevance [precision] is a measure of the exclusiveness of L with respect to $N-C$.' (I have used my notation.) That is a reasonable description of recall, but the equivalent to recall which measures the exclusiveness of L with respect to $N-C$ is surely fallout. If precision is to be described in these terms, it must be said to measure the inclusiveness of C with respect to L; but since C is not variable in this sense, it seems absurd to try to describe precision in these terms at all.

Cleverdon *et al*[13] classify index language devices as 'recall' or 'precision' devices. A recall device is one which tries to enlarge the set a at the expense of set c, and thus its effect is suitably measured by M. It might also accidentally enlarge b at the expense of d. A precision device, on the other hand, is

designed to enlarge *d* at the expense of *b*, but might accidentally enlarge *c* at the expense of *a*. To measure the beneficial effect of such a device with *P* is clumsy, since *P* is also affected by the adverse effect of this device (on *a*). Similarly *P* is a clumsy measure of the adverse effect of a recall device. By analogy with recall, it would surely be better to measure the beneficial effect of a 'precision' device with fallout, and to rename it accordingly.

While these comments on recall and precision do not provide conclusive arguments against their use, they do indicate that perhaps recall and fallout are a more satisfactory pair for the purpose of describing results. I therefore continue with a consideration of recall and fallout. Some more comments on precision are made later in the paper.

7. RECALL AND FALLOUT

Parameters similar to recall and fallout were first considered for use in *IR* tests by Mooers.[17] Equivalent parameters (derived in the same way from a 2×2 table) are used in many other situations, as both Swets[6] and Rees[11] point out; a particular case is the test by King *et al*[4] on accuracy of indexing. As I have pointed out, recall and fallout appear superior to recall and precision on several counts. Nevertheless, these parameters have not in fact been used very much (see Appendix A): recall and precision are used almost universally.

I now consider two properties of the recall-fallout graph which appear to make it particularly suitable for plotting performance curves. The normal practice when plotting the performance curve of a system is to join the points (representing for example levels of co-ordination) either by a series of straight lines or by a smooth curve. It is often not clear what exactly these lines mean; in the latter case, the curve is often drawn through all points except those showing obvious discrepancies, though this method is highly subjective unless there is a standard form of performance curve. I now show that *a*, if the points on the recall-fallout graph are joined by straight lines, these lines have a real and useful meaning; and *b*, there is a standard form of smooth curve which appears to fit the results very well, and offers some other important advantages.

a Consider for example a system similar to the Cranfield one, by which sets of documents are retrieved at a series of levels of co-ordination, each level represented by a point on the recall-fallout graph. If the documents retrieved at level of co-ordination exactly 4 are given a random order, this will generate a series of cut-off points between point 5 and point 4; on the recall-fallout graph (though not on the recall-precision graph) these points will on average form a straight line. Thus the points on the straight line joining points 4 and 5 are actually attainable by the system (with the help of random ordering); the performance curve formed in this way truly represents the value of the system.

b The standard form of performance curve is that suggested by Swets[6,7] (see also Brookes[18]). Swets proposes that recall should be plotted against fallout on double probability graph paper. His model then predicts that the points will lie on a straight line; according to one version of the model this line is at 45°. In his second paper, he tests these predictions on a number of published results; the first prediction is surprisingly accurate,

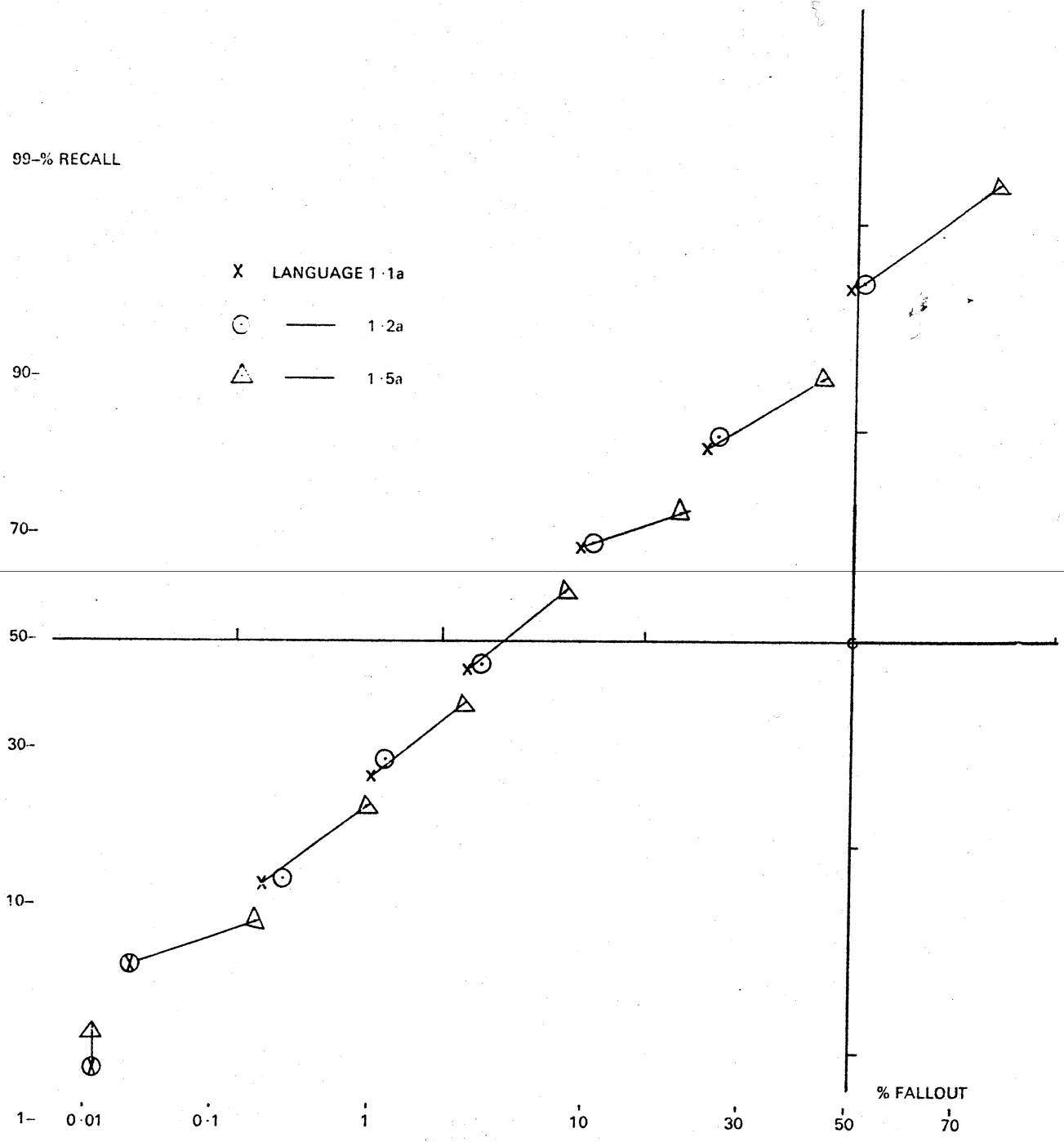
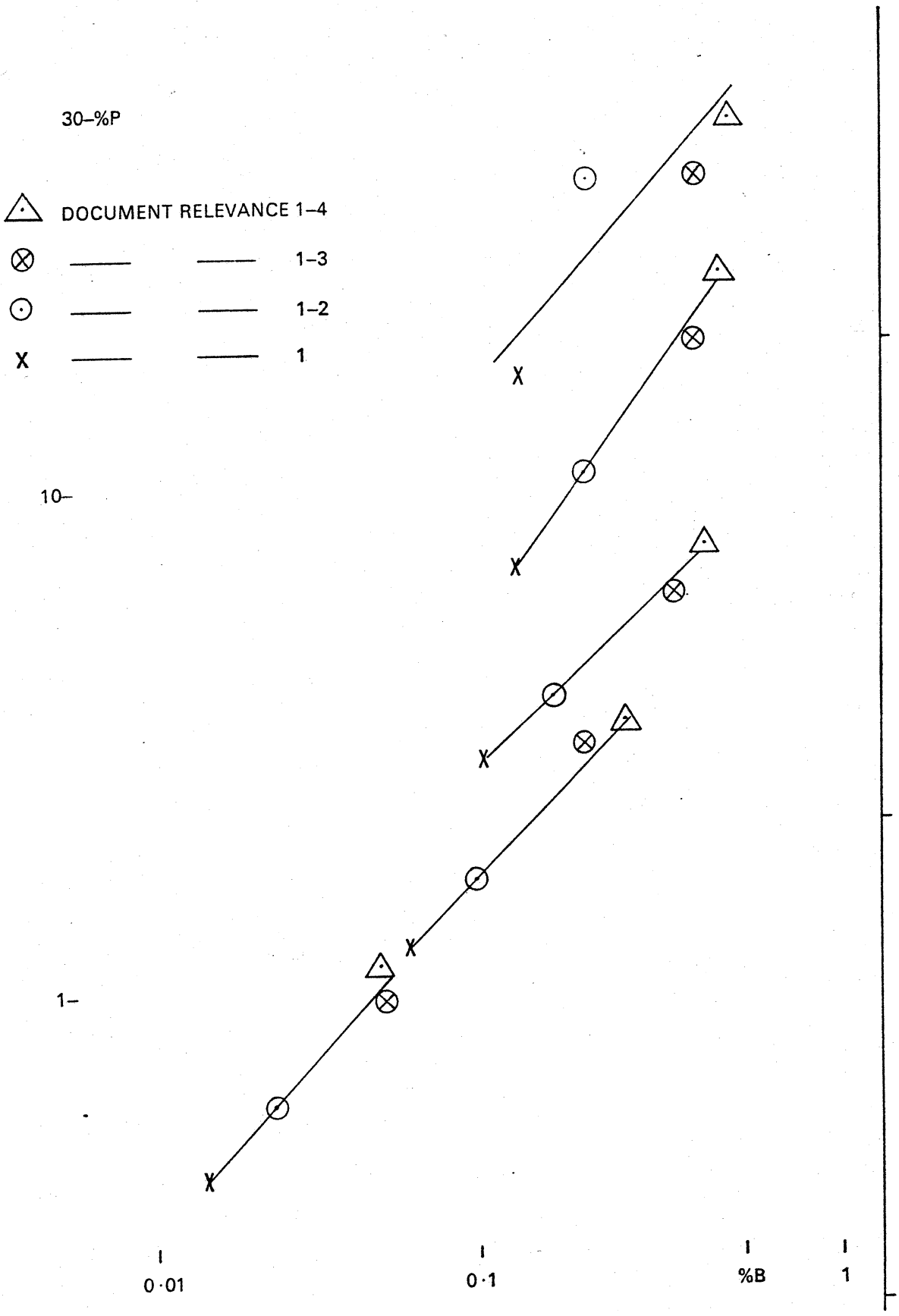


FIG. 7.1

though the second is not (this has repercussions in his attempt to find an overall measure of system effectiveness, see Part 2). This straight line, then, is the standard form, and can be used to iron out any random irregularities in the curve.



Levels of co-ordination (from top to bottom) : 8, 7, 5, 3, 1.
Data from Cranfield II[-4], figures 4.610T–4.613T.
P = a/(a + b) (precision) and B = c/(c + d)

FIG. 7.2

But this type of graph has a great many more advantages. It appears to show up the effect of certain variables much better than the conventional type. For example, Fig 7.1 shows the Swets graphs for three sets of Cranfield results, those for languages 1.1a (single terms; natural language), 1.2a (single terms; synonyms), and 1.5a (single terms; synonyms and quasi-synonyms). It is clear that one could draw a straight line for each language, and that there would not be much difference between them. But it is perhaps more interesting to indicate the effect of confounding synonyms and quasi-synonyms at each level of co-ordination, as I have done in Fig. 7.1. It appears that the total effect is approximately independent of the level of co-ordination, i.e. this form of graph isolates this effect (the effect of confounding synonyms alone is too small to be of much significance). This is more than can be said for the conventional recall-precision graph, or the recall-fallout graph on linear scales.

A related advantage is that a graph which is something like a straight line at 45° shows up the results much better than a curve which is very close to the axes at each end. Cleverdon and Keen[14] apply a similar principle when they plot recall-fallout graphs on semi-log paper; but they still get a curve of irregular shape. Keen[19] says: '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.' But according to Fig. 4.102T in Cranfield II[14] (representing the best language 1.3a), precision will necessarily be below 5% if a recall over 50% is to be obtained. Also the other end of the curve, which is probably more important, suffers from the same problem.

There is another situation in which the technique might be applied. As I suggested in §5, while level of co-ordination and other variables can be regarded as vertical extensions of the 2×2 table, there is one variable, relevance level, which corresponds to a horizontal extension. By analogy with recall and fallout, it might be worth measuring the effect of this variable with $a/(a+b)$ (precision) and $c/(c+d)$, which I shall call B. Fig. 7.2 shows a graph of P against B on double probability graph paper, for different levels of relevance at five different levels of co-ordination (Cranfield II data). Once again the results are surprisingly regular, they show a pattern which is approximately independent of the level of co-ordination.

8. METHODS OF AVERAGING

Everything I have said so far could apply to a test in which only one question was used. But a brief glance at any set of results shows that an individual question is useless for evaluation purposes: if any sense is to be made of the results, averages must be taken over a large number of questions.

There has been some controversy over the two possible ways of averaging a secondary parameter over a set of questions: the 'macro' and 'micro' methods (Rocchio's terminology).[20] For the macro method, a parameter (say recall) is calculated for each question, and the average is then taken. For the micro method, the entries in the 2×2 table (say R and C) are totalled over the set of questions, and the ratio of the totals calculated to give the parameter. These methods sometimes give different results. The reason, as Rocchio points out, is the variation in the denominator (say C) of the parameter. Thus if there is a (linear) correlation between the parameter and its denominator, the methods will give different results; if there is none the results will be approximately the same. There is a marked correlation between P and L, so the results for precision by the two methods are different. There appears to be no appreciable correlation between M and C (but see §9), so for recall the results are approximately the same. Similar remarks apply to fallout, although the existence of a correlation in this case would not have much effect, as there is in practice very little variation in the denominator $N-C$. (Results from Cranfield II).[14]

A number of authors (e.g. Farradane *et al*)[8] claim that the micro method is the 'correct' one and should be used at all times. Cleverdon and Keen[14] say that it does not matter which you use so long as you specify which you are using. I suggest a more careful analysis is required.

The fact that the two methods give different results for precision indicates, as I have said, a strong correlation between P and L. That is to say, if L is larger than average, P is likely to be smaller (other things being equal). Suppose, then, that question 1 gives $R=2, L=5$, i.e. $P=40\%$, and question 2 gives $R=5, L=15$, i.e. $P=33\%$. If you assume that precision is a reasonable basis for comparison between these sets of figures, you can say that (in this respect) the result for question 1 is better than that for question 2. But it is now known that such a result is to be expected—i.e. that a system of given effectiveness is expected to give high P-values corresponding to low L-values, and vice versa. In this case, it is no longer reasonable to say that the first result is better than the second. Hence precision is not a reasonable basis for comparison. So it is not reasonable to average precision over a number of questions at all—it is not clear to which kind of question the resulting value applies, and it certainly does not apply to all of them.

It can be seen that I have now introduced a new principle. This is that the choice of parameters should depend on the existence of empirical relationships between the variables concerned. In this case I am concerned with the non-linearity of the relationship between R and L, which implies the existence of a correlation between P and L. However, it appears from the Cranfield II results that the relationship between R and C is (on the average) linear, also that between $L-R$ and $N-C$ (but see §9). So the choice of parameters M and F is justified by this principle; that of P is not.

There are, however, many more difficulties in averaging. Cleverdon and

Keen[14] consider six possible methods of choosing which point on each individual performance curve to use to give a point on the average curve. They finally opt for a simple level of co-ordination match, irrespective of the number of starting terms of the questions. But later on, in order to calculate their normalized recall, they average by a 'document output cut-off' method, i.e. by values of L (the recall-precision curves for the two methods are different shapes). The results of tests on the SMART system are usually averaged by values of recall; here there is a difficulty in the choice of extrapolation and interpolation methods (see Keen).[15]

Here I would like to make two points. The first is that the interpretation placed on the final curve must depend on the averaging method used. For example, if a user is confronted with a level of co-ordination curve he can say: 'If I ask for all documents at level of co-ordination 5+, then I can expect to get so much recall and so much precision (fallout).' On the other hand, if he is presented with a document output and cut-off curve, he can only say: 'If I get so many documents out, I can expect to get so much recall and so much precision'—a statement which, with the Cranfield system, can only be applied *after* he knows how many documents the system is going to give him.

The second point is that the above principle of making use of statistical relationships should apply. The statistical relationships between the variables being averaged clearly depend on the choice of which points on the curves to average; a study of these relationships would be of great use.

In general, I think the dependence of parameters on controls over which they are averaged (correlation), and the distribution of parameters with respect to controls (variance) should be investigated, in order to get a more accurate picture of how *IR* systems can be expected to work. For example, it might be worth studying how far the Swets model applies to individual questions, and to develop averaging methods based on such an analysis. However, here is clearly a subject for a major statistical study, which unfortunately cannot be attempted here.

A related problem is the statistical significance of the results obtained. The only testers to have attempted analyses of statistical significance are the SMART system workers (see Lesk).[21] I hope this will become more widespread. It is worrying when, for example, Cleverdon and Keen[14] present a table ranking their index languages according to normalized recall, and the first ten languages cover a range of only 1·4%. It is even more worrying when they say: 'It is impossible to state here what is a significant difference; most people who have been consulted agree that anything less than 1% is probably of doubtful significance, but that a difference of 3% or 4% almost certainly represents a significant change in performance.' This of course is a statistical problem—the question can only be answered by an analysis of the sampling distribution of normalized recall.

9. GENERALITY AND COLLECTION SIZE

I have already described G and N as experimental conditions—data which should be given but are not part of the results proper. This idea should be considered in more detail.

Any scientific experiment is conducted under certain conditions. The scientific method requires that these conditions be reported in full. But that is not enough—since field conditions are invariably different from laboratory conditions, the experimenter should ideally try to make his results applicable outside the laboratory. That is, he should either present his results in a form in which they are independent of the conditions, or he should chart the effect of the conditions on his results. This is particularly relevant to *IR* systems, since the conditions under consideration (particularly the generality of a question) are not under direct control.

So I am concerned with the empirical dependence of parameters on G and N. I require that (ideally) the parameters used should be independent of G and N; or that the dependence should be analysed. It is well known (and to be expected) that precision depends on G (see for example Cleverdon and Keen).[14] It is also generally thought that recall and fallout are independent of G. Thus Cleverdon and Keen, when comparing situations of different generality, use recall and fallout or recall and 'adjusted precision'. The adjusted precision is simply the precision to be expected if recall and fallout are independent of generality; it is calculated from the recall, the fallout and the new generality. They do not, however, appear to carry this idea back far enough: they take average precision values over a set of questions of different generality values, and they also average generality values. It would surely be more logical to average adjusted precision figures for some fixed generality, if precision is what is wanted (though they could not do this with the 'micro' method of averaging which they use).

Since adjusted precision is based on recall and fallout, it seems more reasonable to calculate average M and F values, and then (if necessary) to calculate the precision implied by these values. This is particularly so because the non-dependence of recall and fallout on generality is not definitely established, and is certainly only approximate. I now review the evidence on these relationships.

As I observed in the last section, the fact that the macro and micro methods of averaging give very nearly the same results for recall is a strong indication that there is no significant correlation between recall and C, i.e. M is independent of C and therefore of G. The same reasoning applies to fallout, but with the reservation that the variations in $N-C$ are relatively small, so that a correlation would not show up so well.

On the other hand, Cleverdon and Keen compare the performance of a set of questions of low generality (specific questions) with that of a set of questions of high generality (general questions). The effect on individual

parameters is hard to ascertain (the two types of question do not appear to have the same number of starting terms), but on the recall-fallout graph the specific questions show somewhat 'better' performance. The inverted commas are to indicate that I regard this difference as an effect of the conditions on the parameters, rather than an indication that the system works better for specific questions. Keen[15] reports a similar result on the SMART system tests. But neither try a statistical correlation analysis. Farradane *et al*[9] report no correlation between recall and generality.

Another mechanism leads to a change in generality: the acceptance of a different level of relevance for relevant documents. Cleverdon and Keen also study this kind of change (see also §7); once again, on the recall-fallout graph, the low-generality results (only high-relevance documents are accepted) are slightly better.

The problem of the dependence of parameters on collection size is a much more difficult one, as it inevitably involves consideration of the subject area of the collection, which will probably affect results. Cleverdon and Keen study the performance of a set of questions in a specialized document collection smaller than the total collection, but still containing all documents relevant to the questions—i.e. C remains the same, N and G change. Here again the recall-fallout graph shows better results in the low-generality case (large collection).

But one can imagine another kind of change in the conditions in which recall and fallout are not affected. I follow Fairthorne's suggestion:[22] 'Consider this "Gedankenexperiment". A retrieval test has yielded so many acceptable items, so many unacceptable. We now throw out of the window some of these, and return the rest to the collection. We repeat the test exactly, with the same request and the same criteria for selection, rejection, and acceptance. Unless someone goofs, this new test must yield the same items less the items thrown out of the window. In general this will alter most accepted measures of performance, some of them drastically. But no retrieval characteristics have changed.' I do not think that this as it stands is a valid operation to take. We are testing the validity of a decision—the decision on the part of the system whether or not to retrieve a given document. If we wish to test its value with respect to an entire population, we must take a random sample of that population—at least, it must be a random sample as far as the decision-making process is concerned. If we then throw out of the window some of the documents for which the decision is positive, without throwing out a corresponding number of those for which the decision is negative, we are biasing the sample, so we can no longer hope to get the same results.

If on the other hand we take some of the sets C, $N-C$, N and throw them out of the window, without knowing whether they would be retrieved or not—i.e. if we randomly alter the sample—then we hope not to affect whatever parameters are used, since we are changing the conditions and not

the results directly. But now we are alright, as we would not (on average) affect recall and fallout, although precision will probably be affected.

It seems, then, that these relationships need further study. But whatever the dependence of recall and fallout on the conditions of the experiment, they are certainly not as drastically affected as precision. This in my opinion makes them a more suitable pair of parameters for describing in general terms the performance of a system. It is often claimed that recall and precision are the parameters that are most easily interpretable in user terms. I take this to mean that a user can calculate from them how much relevant and irrelevant material to expect in answer to his question. This is not true, as his question (in his library) probably has a different generality from that of the experiment. The precision figure from the experiment is therefore not applicable.

10. EMPIRICAL RELATIONSHIPS

I have already considered a number of statistical relationships—e.g. the dependence of parameters on generality, etc. In this section I consider a rather different kind of relationship: an example is Cleverdon's 'inevitable inverse relationship between recall and precision'. The particular case under consideration is the case where the controlling variable is the level of co-ordination, or some such manifestation of Swets' variable z; but my results can be applied to other variables which are equivalent to vertical extensions of the 2×2 table. Rocchio[23] points out that the relationship in this case is to be expected; I consider it from a rather different viewpoint.

In the first place, it is a compound relationship, not a simple one. The first part is the recall-level of co-ordination relationship. This is negative (inverse), deterministic (it applies to all questions all the time), and obvious. The second and more interesting part is the precision-level of co-ordination relationship. This is positive, but it is of quite different character to the previous one. It is a statistical relationship—it is not in general true for individual questions, only for the average. This difference between the two parts is in itself a reason for not plotting recall against precision in this case—they are not strictly compatible.

Consider now the recall-fallout graph. Here the relationships are (in the first instance) quite straightforward: there is a negative recall-level of co-ordination and a negative fallout-level of co-ordination relationship. Both are deterministic; the combination yields the familiar and universal positive relationship between recall and fallout. I now make a hypothesis concerning a *statistical* property of the graph; it is assumed that the points (including (0,0) and (1,1)) are joined by straight lines, as suggested in §7.

Hypothesis: *The recall-fallout graph representing the average of a set of questions is convex.* That is, if any of the straight lines is extrapolated to the left, as in Fig. 10.1, it will pass above the next point to the left.

Proposition 1: *If the hypothesis is not true of a system, then a change can be*

made in the system whereby (a) nothing fundamental to the system is altered; (b) the performance is improved; and (c) the hypothesis is made true.

The proof of this proposition depends on the interpretation of the straight line joining two points, mentioned in §7. If the graph is not convex, there is a point such as *U* in Fig. 10.2 which lies below the straight line *TV*. So there is a point *W* on *TV* that has both higher recall and lower fallout than *U*. The point *W* can be reached by taking the documents retrieved at point (level of co-ordination) *T*, together with a random sample of the requisite proportion of the extra documents retrieved at *V*. Then if the system is changed to include *W* and exclude *U*, 1. nothing fundamental is altered—only the way in which the retrieved documents are presented; 2. the performance is improved—performance curve *TWV* is better than *TUV*; and 3. the hypo-

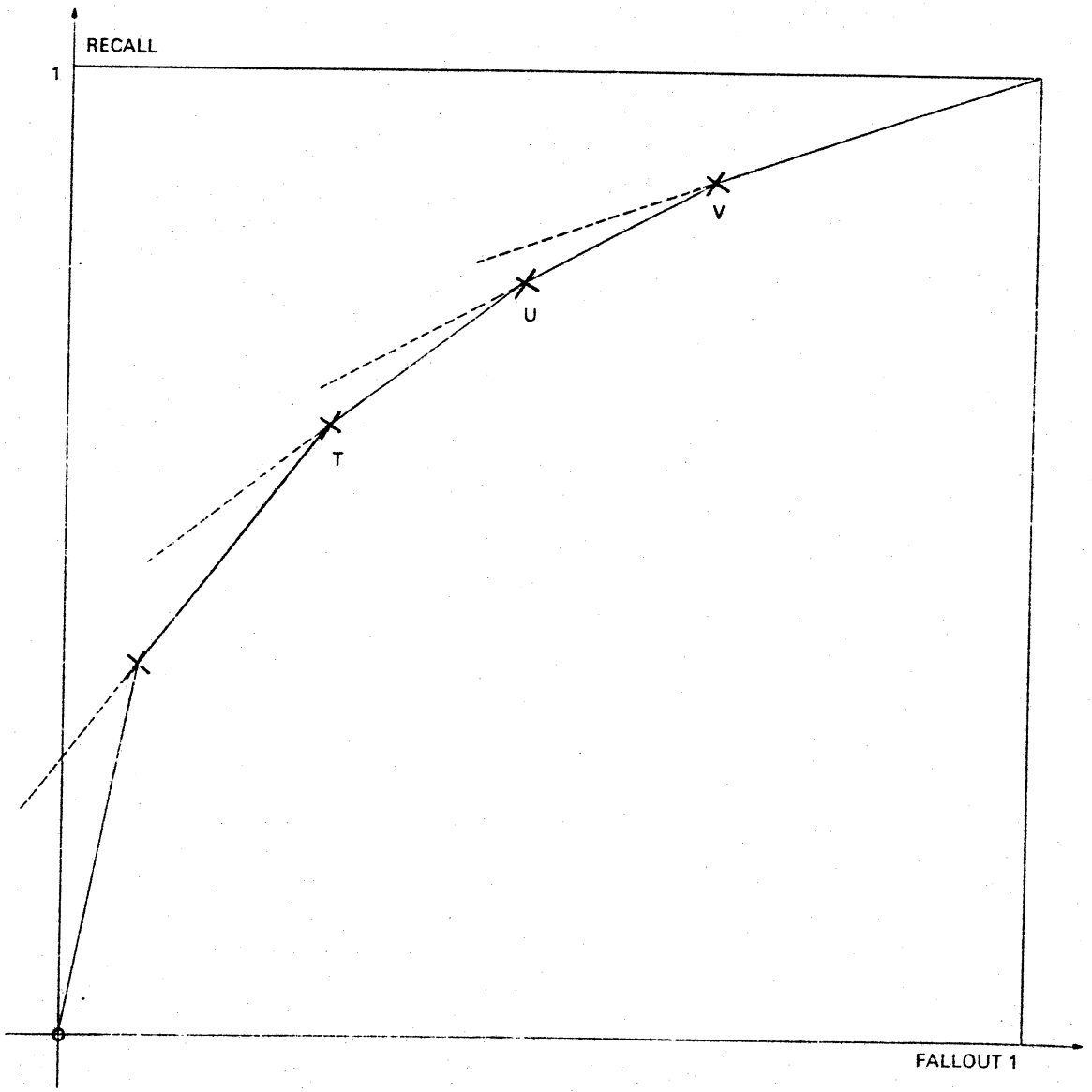


FIG. 10.1

thesis is now true, at least in that area of the graph (it may be necessary to repeat the process in other areas).

That such a trivial improvement should be possible is obviously unlikely. If for example the points represent levels of co-ordination, a non-convex graph would imply that it was better to take a random sample of the documents retrieved at level of co-ordination 4 than to take those at level 5; if this were so, one would clearly deduce that something was drastically wrong. It therefore seems reasonable to accept the validity of the hypothesis.

Proposition 2: *My hypothesis implies Cleverdon's hypothesis, that precision is positively related to level of co-ordination.*

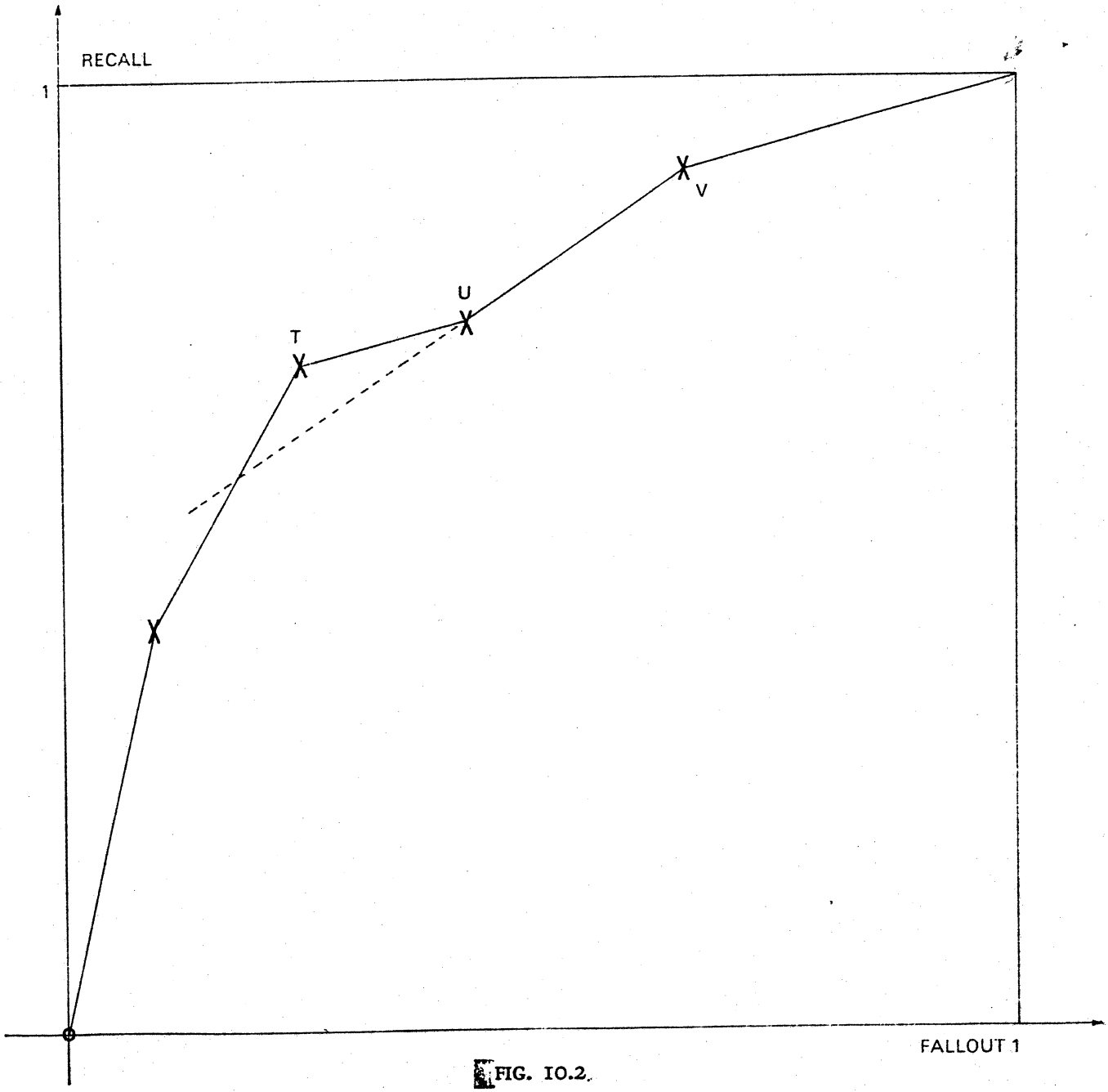


FIG. 10.2.

The mathematical relationship between recall, fallout, precision, and generality can be expressed as follows:

$$M=\frac{P(1-G)}{G(1-P)}\times F$$

Hence, at a given generality, the constant precision lines on the recall fallout graph are straight lines through the origin, of gradients which increase with P. The change in precision between the points U and V of Fig. 10.1 is indicated by the fact that the line UV crosses some of the constant-precision lines. U is the higher level of co-ordination of the two; clearly, there is an increase in precision from V to U if the line VU extended passes above the origin, and a decrease if it passes below. But my hypothesis implies that it passes above the point T, and hence above each successive higher level of co-ordination, including the origin. Hence proposition 2.

In fact, my hypothesis is considerably more powerful than Cleverdon's—it is easy to imagine curves that satisfy Cleverdon's but not mine, e.g. Fig. 10.2. It also implies that $B=c/(c+d)$ is positively related to level of co-ordination (on average). This shows not only that Cleverdon's rule is to be expected, but also that there are other similar relationships to which all reasonable *IR* systems are subject. It might be reasonable sometimes to check that such rules are being obeyed. For example, Farradane *et al*[9] found that with their first value of C, the jump from 'question as indexed' to 'conceptual browsing' (set 112 to set 212) not only increased recall, but also increased precision slightly. Or, as I would prefer to put it, the recall fallout graph for these two points together with the origin was not convex. This implies that the questions as given and the sets of answers as given did not match—the questions arrived at by conceptual browsing were *on average* closer to the answers than the original questions. In fact they made this deduction for other reasons, and had some documents reassessed for relevance; the resulting sets 122 and 222 showed the usual increase in recall and decrease in precision. I would like to stress the fact that they could have made this deduction on purely numerical grounds.

Another possible application of the hypothesis is as follows. The Swets line—a straight line on double probability graph paper—corresponds of course to a smooth curve on the linear scale paper. This curve is only convex if the Swets line is at 45°—otherwise one end of the curve bends the other way. The closer the slope is to 45°, the shorter is the non-convex region. This suggests that if a large enough set of questions is used, the Swets line might in fact be at 45°. The suggestion is reinforced by an observation by Swets:[7] the results for individual questions also give remarkably straight lines, but with much more variation in gradient. As I said in §8, these individual variations deserve further study.

This kind of relationship, then, can be regarded as a true property of *IR* systems; it forms a step between the purely deterministic relationships

mentioned above and the approximate statistical ones, such as the straightness of the Swets line, or the non-dependence of recall on generality. I hope it is also a useful rule.

II. CONCLUSIONS

The usual reasoning behind the choice of a parameter appears to be something like this: 'the set of A documents is included in the set B; in the ideal case, A coincides with B. The parameter A/B equals 1 if this ideal is achieved, less if it is not, and therefore is a measure of how well the ideal is achieved.' In particular, most parameters are based on the 2×2 table representation of the search results under certain conditions.

Such reasoning is not enough. Many other points must be considered if the results are to be described as fully as possible in terms as general as possible. In this paper I have considered some traditional parameters from a number of points of view: in particular, from the point of view of the performance curve (extension of the 2×2 table) by which a variable is introduced into the system; and also from the point of view of the statistical relationship between the parameters. From these considerations have emerged some desirable properties of parameters.

I hesitate to be categorical, but it appears that although recall and fallout satisfy most of these requirements adequately, precision (or the recall-precision graph) does not. The following particular criticisms apply:

a Recall and precision (with generality and collection size) fail to describe the results completely in one case ($R=0$); they do not satisfy my version of Fairthorne's requirements; precision is not in fact the best measure of the functions which it is normally used to measure.

b The recall-precision graph is not so easy to interpret as the recall-fallout graph; there is no satisfactory standard form of performance curve.

c Precision is not independent of L, nor of generality; these facts make it difficult to interpret precision figures in conditions outside those of the experiment.

I must therefore come to the conclusion that precision is in general not such a useful measure of retrieval performance as fallout (in conjunction with recall). My hypothesis of §10 indicates that the recall-fallout graph can show as much as if not more than the recall-precision graph, so nothing is lost by using recall and fallout.

I would like to stress that I consider the principles proposed in this paper to be as important as the resulting comments on particular parameters. As the methodology of testing retrieval systems advances, new quantities will be considered; I hope that the above principles will provide some guidelines on the problem of manipulating these quantities.

ACKNOWLEDGEMENTS

The work was carried out under a grant from the Science Research Council. I am grateful to B. C. Brookes for some helpful comments.

APPENDIX

Following is a table indicating which parameters have been used in various tests. For more details and references, readers are referred to the two reviews of the field, by Bourne[1] and Rees.[2] The 'source documents' column is marked if the questions put to the system were based on particular documents which the system was then required to retrieve. The 'relevance levels' column indicates how many levels of relevance other than non-relevant were used in the assessment of documents; W indicates that the documents were given relevance weights, and R indicates that they were ranked for relevance. The next two columns (*R* and *L*; *C*) indicate how much of the 2 × 2 table was counted for each question. The symbol * in the *C* column means that the values of *C* were estimated by some means. The next three columns (*M*, *P*, *F*) indicate which secondary parameters were used; I have not distinguished between fallout and specificity, these being complementary. The dates are dates of publication.

Date and organization	Source documents	Relevance levels	R and L	C	M	P	F	Other parameters
1954								
Royal Aircraft Establishment	✓		✓					
1955								
R.A.E.—Cranfield	✓							
1956								
Documentation Inc.			✓					
1960								
Netherlands Armed Forces			✓					
Thompson Ramo Wooldridge		W	✓	✓				Swanson's measure
1961								
Cranfield—English Electric	✓							
1962								
Cranfield I	✓		✓			✓		
1963								
University of Pittsburgh	✓							
U.S. Patent Office			✓	✓				
Western Reserve University		2	✓	✓	✓	✓		
W.R.U.	✓	2	✓					
W.R.U.		2	✓			✓		
1964								
W.R.U.		4	✓	✓	✓		✓	
U.S. Bureau of Ships			✓	*	*	✓		
U.S.A.F.			✓	✓				Borko's measure
U.S. Patent Office		2	✓					
Mitre Corporation			✓			✓		
(SYNTOL)			✓	✓	✓	✓		
W.R.U.			✓					
1965								
Union Carbide			✓			✓		Productivity; relative call
MEDLARS			✓	✓	✓	✓		
University of Texas			✓					
E.I. du Pont			✓	*	*	✓		
Harry Diamond Laboratories		3	✓	✓	✓	✓		
IBM Watson Research Centre			✓	✓				
Science Information Exchange			✓	✓	✓	✓		
1966								
American Society for Metals			✓			✓		Usefulness
Euratom			✓	✓	✓	✓		
National Library of Medicine			✓	*	*	✓		
E.I. du Pont			✓	✓	✓	✓		
U.S. Patent Office								Measures for accuracy of indexing
Cranfield II		4	✓	✓	✓	✓	✓	Normalized recall
Arthur D. Little		W	✓	✓	✓			Normalized sliding ratio
SMART system (Harvard/		R	✓	✓	✓	✓		Normalized

REFERENCES

1. BOURNE, C. P. Evaluation of indexing systems. In: Cuadra, C. A. (editor), *Annual Review of Information Science and Technology*, vol. 1, Interscience, New York, 1966, p. 171–90.
2. REES, A. M. Evaluation of indexing systems and services. In: Cuadra, C. A. (editor), *Annual Review of Information Science and Technology*, vol. 2, Interscience, New York, 1967, p. 63–86.
3. REES, A. M. The Aslib Cranfield test of the Western Reserve University indexing system for metallurgical literature: a review of the final report. *American Documentation*, vol. 16, no. 2, April 1965, p. 73–6.
4. KING, D. W. Evaluation of co-ordinate index systems during file development. *Journal of Chemical Documentation*, vol. 5, no. 2, May 1965, p. 96–9.
   KING, D. W. *and* MCDONNELL, P. M. Evaluation of co-ordinate index systems during file development. Part II: an application. *Journal of Chemical Documentation*, vol. 6, no. 4, November 1966, p. 235–40.
5. GOFFMAN, W., *and* NEWILL, V. A. A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval*, vol. 3, no. 1, August 1966, p. 19–25.
6. SWETS, J. A. Information retrieval systems. *Science*, vol. 141, 19 July 1963, p. 245–50.
7. SWETS, J. A. *Effectiveness of information retrieval methods.* Report no. AFCRL-67-0412, Air Force Cambridge Research Laboratories, Bedford, Mass., 1967 (47p.).
8. FARRADANE, J., DATTA, S., *and* POULTON, R. K. *Research on information retrieval by relational indexing. Part 1: methodology.* The City University, London, 1966 (60p.).
9. FARRADANE, J., DATTA, S., *and* HILLS, J. *Research on information retrieval by relational indexing. Part 2: test results.* The City University, London, (forthcoming).
10. CLEVERDON, C. W. *Interim report on the test programme of an investigation into the comparative efficiency of indexing systems.* Aslib Cranfield Research Project, College of Aeronautics, Cranfield, Bedford, 1960 (84p.).
11. REES, A. M. *The evaluation of retrieval systems.* Technical report no. 5, Comparative Systems Laboratory, Center for Documentation and Communication Research, School of Library Science, Western Reserve University, Cleveland, Ohio, 1965 (21p.).
12. FAIRTHORNE, R. A. Basic parameters of retrieval tests. In: American Documentation Institute, *Parameters of information science: Proc. 27th Annual Meeting, Philadelphia, Pa., 5–8 October 1964.* Spartan Books, Washington, D.C., 1964, p. 343–5.
13. CLEVERDON, C. W., MILLS, J., *and* KEEN, E. M. *Factors determining the performance of indexing systems, vol. 1 design.* Aslib Cranfield Research Project, College of Aeronautics, Cranfield, Bedford, 1966 (Part 1, text, 120p.; Part 2, appendices, 377p.).
14. CLEVERDON, C. W., *and* KEEN, E. M. *Factors determining the performance of indexing systems. Vol. 2: test results.* Aslib Cranfield Research Project, College of Aeronautics, Cranfield, Bedford, 1966 (299p.).
15. KEEN, E. M. Evaluation parameters. In: *Information storage and retrieval.* Report no. ISR-13, Department of Computer Science, Cornell University, Ithaca, N.Y., 1968, p. II-1 to II-67.
16. ROCCHIO, J. Performance indices for document retrieval systems. In: *Information storage and retrieval.* Report no. ISR-8, Computation Laboratory of Harvard University, Cambridge, Mass., 1964, p. III-1 to III-18.
17. MOOERS, C. N. *The intensive sample test for the objective evaluation of the performance of information retrieval systems.* ZTB 132, Zator Co., Cambridge, Mass., 1959 (20p.).
18. BROOKES, B. C. The measure of information retrieval effectiveness proposed by Swets. *Journal of Documentation*, vol. 24, no. 1, March 1968, p. 41–54.
19. KEEN, E. M. *Measures and averaging methods used in performance testing of indexing systems.* Aslib Cranfield Research Project, College of Aeronautics, Cranfield, Bedford, 1966 (59p.).

20. ROCCHIO, J. Evaluation viewpoints in document retrieval. In: *Information storage and retrieval.* Report no. ISR–9, Computation Laboratory of Harvard University, Cambridge, Mass., 1965, p. XXI–1 to XXI–10.
21. LESK, M. E. SIG—the significance programs for testing the evaluation output. In. *Information storage and retrieval.* Report no. ISR–12, Department of Computer Science, Cornell University, Ithaca, N.Y., 1967, p. II–1 to II–22.
22. FAIRTHORNE, R. A. (Review of Farradane *et al.*[8]). *Journal of Documentation*, vol. 24, no. 2, June 1968, p. 127–31.
23. ROCCHIO, J. Document retrieval systems—optimization and evaluation. Thesis published as: *Information storage and retrieval.* Report no. ISR–10, Computation Laboratory of Harvard University, Cambridge, Mass., 1966 (163p.).