THE

Journal of Documentation

VOLUME 25 NUMBER 2 JUNE 1969

THE PARAMETRIC DESCRIPTION OF RETRIEVAL TESTS*

PART II: OVERALL MEASURES

S. E. ROBERTSON

Research Department, Aslib

Two general requirements for overall measures of retrieval effectiveness are proposed, namely that the measure should be as far as possible independent of generality (this is interpreted to mean that it can be described in terms of recall and fallout), and that it should be able to measure the effectiveness of a performance curve (it should not be restricted to a simple 2 × 2 table). Several measures that have been proposed are examined with these conditions in mind. It turns out that most of the satisfactory ones are directly or indirectly related to Swets' measure *A*, the area under the recall-fallout curve. In particular, Brookes' measure *S* and Rocchio's normalized recall are versions of *A*.

I. INTRODUCTION

IN THE FIRST PART[1] I considered some parameters which are used to describe the results of tests on *IR* systems, in particular the simple parameters which are derived directly from the 2×2 table. I concluded that there were several considerations outside the scope of the 2×2 table which are relevant to the choice of parameters. In particular, *a*, a variable such as 'level of co-ordination', which produces a performance curve of the

* It has been pointed out to me that a number of unfortunate errors occurred in Part I of my paper.[1] I should like to take this opportunity to correct them.

I incorrectly attributed to Mr Cleverdon the practice of extrapolating the recall-precision curve to $M=0$, $P=1$. I apologize for the error. On page 16, line 15, for 'document output and cut-off curve' read document output cut-off curve. In the Appendix, the Documentation Inc. test should have been called ASTIA—Documentation Inc. It was first reported in 1953. The first R.A.E. test should have been called RAE–Cranfield, and the first W.R.U. test should have been Cranfield–WRU. S.E.R.

system, corresponds to an extension of the 2 × 2 table; and *b*, there appear to be several important statistical relationships between the variables (the 2 × 2 table is only concerned with the strict mathematical relationships).

A large number of more complicated parameters have been proposed to measure the overall 'effectiveness' of the *IR* system under test. Several of these can be faulted on the basis of two simple requirements derived from the above two points. However, many deserve further consideration; and here one comes up against the problem that they are all described in different terms, and therefore cannot be compared directly. In this paper, therefore, I try to describe a number of overall measures in the same terms; and I find to my surprise that most of them are very closely related. Now it is possible to compare their various properties directly; I hope this will prove useful to those engaged in the testing of *IR* systems, even if they disagree with my conclusions.

2. BASIC REQUIREMENTS

The most important of the statistical relationships that I considered in part 1 is probably the dependence of parameters on generality. It is vital that any parameter used should be as far as possible independent of the particular conditions of the experiment (e.g. generality); otherwise comparisons between different results are meaningless. From the results in part 1, recall and fallout appear to be approximately independent of generality, whereas precision seems to have a marked dependence. So any parameter which can be expressed in terms of recall and fallout alone will also be approximately independent of generality. For any others, individual studies would have to be made to establish whether ot not they depend on generality. I therefore require that a measure of effectiveness should be expressible in terms of recall and fallout, or that the tester should make the above study. Since no tester to my knowledge has made such a study, the first part of this requirement is at present the relevant one.

My second requirement is that the measure should in some way be able to cope with a performance curve—that is, it should be definable for a series of points as well as for a single point. Most measures that have been proposed are defined from the simple 2 × 2 table—i.e. from one point only. The only way such a measure can describe the effectiveness of a system with several points is for the same value to describe all points—i.e. the performance curve must coincide with the constant effectiveness curve as defined by this measure. This does not occur with most measures. There are however some measures that are defined in terms of the performance curve rather than the individual points—these I consider in due course.

Some examples of measures serve to illustrate these points. (The notation is that used in part 1; briefly, the 2 × 2 table is described by *R*, *L*, *C*, *N*, or *a*, *b*, *c*, *d*; *M* is recall, *P* precision, *F* fallout, and *G* generality). Verhoeff *et al*[2] propose the measure $a-b-c+d$ (the terms can be weighted); Good[3] con-

siders a somewhat more complicated non-linear function of these terms. Swanson[4] uses $M'-w(L-CM')$, where M' is weighted recall and w is a weight to be chosen as required; Borko[5] modifies this to $M'-w(I-P)$. Giuliano and Jones[6] use a 'normalized sliding ratio', which is derived from the extensions to the 2×2 table (considered in part 1), but which reduces in the simplest case to recall if $L \geqslant C$ and precision if $L < C$.

None of these measures satisfies the first requirement of being expressible in terms of M and F. All but the last are simple one-point measures, whose constant effectiveness curves do not appear to be close to the typical performance curves. The last, although defined in terms of the performance curve, is only designed to measure the effectiveness at one point—to compare performance curves you have to compare several values of the measure. It would seem easier in this case to compare curves directly (visually).

Farradane *et al*[7] propose a measure

$$Q=\frac{ad-bc}{ad+bc}=\frac{M-F}{M+F-2MF}$$

This satisfies my first requirement. For the second, a constant Q curve is fairly close to a typical performance curve, though it does not appear to fit the results as well as the straight line of the Swets model (see §3). Goffman and Newill[8] use the measure $M-F$ (the terms can be weighted). Again this satisfies my first requirement, but does not generate the performance curves. However, it turns out to be related to a more general measure (see §6).

Most of the measures discussed below satisfy my first requirement (are definable in terms of M and F). They are all designed to satisfy my second requirement (to deal with performance curves).

3. SWETS' MEASURES

Swets[9,10] proposes the following model of the retrieval process (see also Brookes[11] for a good description). He considers a variable z corresponding to a vertical extension of the 2×2 table, e.g. level of co-ordination. He makes the following hypothesis: the probability distributions of relevant and non-relevant documents with respect to z are normal. This leads him to plot recall against fallout (for different values of z) on double probability graph paper, i.e. to plot the 'normal deviate' of recall against that of fallout. If these distributions are normal, the performance curve will be a straight line; if in addition the two normal distributions have the same variance, this line will be at 45°. He proposes a measure E', which for a one-point system is the difference between the normal deviate of M and that of F.

Thus this measure is specifically designed to meet my second requirement—if the model is valid, and the performance curve is a straight line at 45°, then E' is the same for all points on the line. E' for a line can now be

measured by means of a scale marked on the negative diagonal (see Fig. 3.1) —the point at which the line crosses the scale gives the value of *E'* for that line.

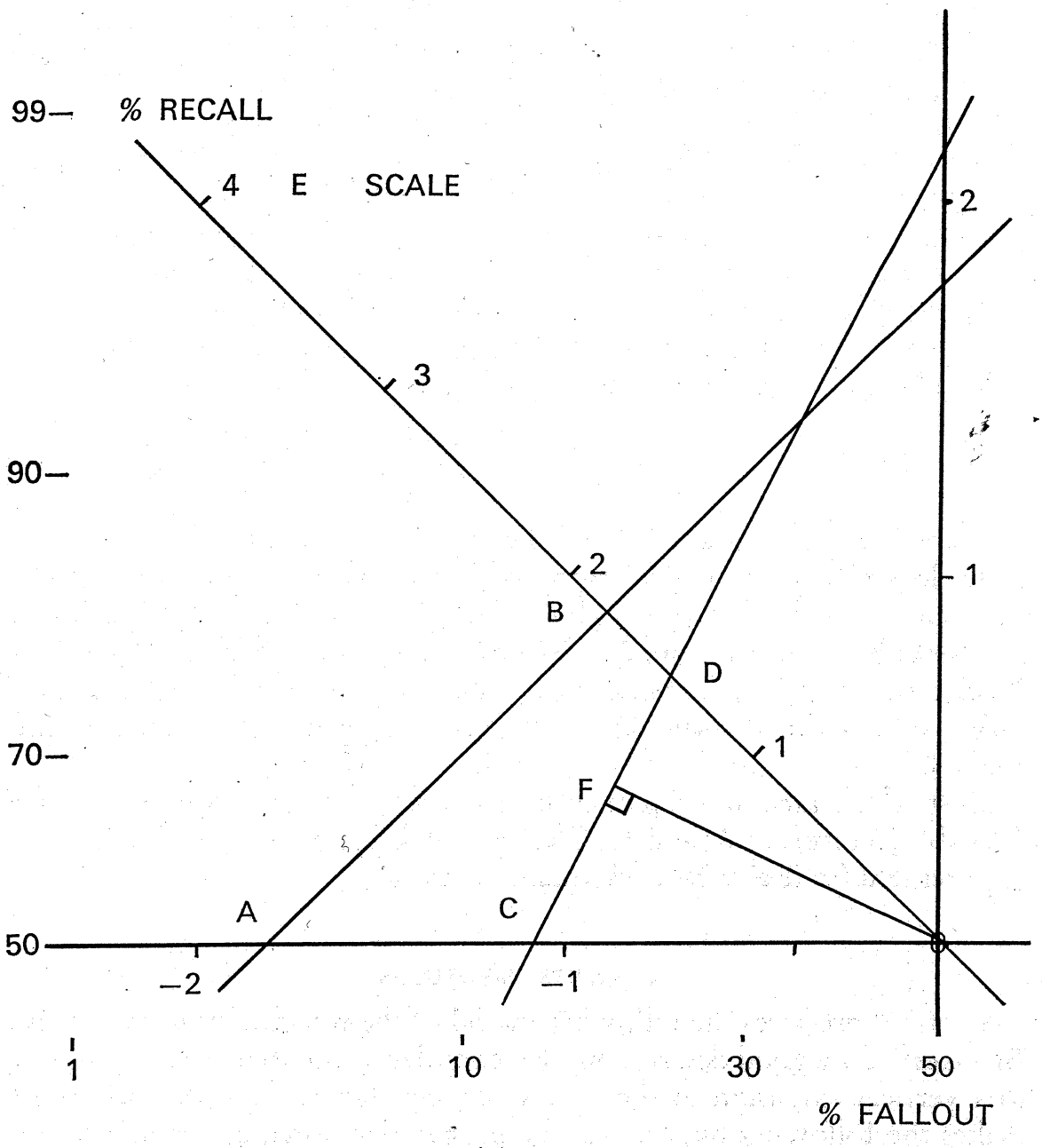


FIG. 3.1 *The Swets measure E*

In his second paper, Swets tests his theory on a number of published results. The first part, that the performance curves will be straight lines, is surprisingly accurate; the second, that the lines will be at 45°, does not hold so well. Swets now proposes to continue using the scale of Fig. 3.1 to measure the performance of a system; I call this measure *E*. Swets does not distinguish between *E* and *E'*, but it is clear that they are of rather different types: *E* cannot be used for one-point systems, only for systems with enough points to form the straight lines. Also a value of *E'* defines a per-

formance curve; a value of E does not. Thus he now has to have a second parameter to describe the performance: s, the slope of the line.

Swets also suggests a measure of a rather different type: A, the area under the recall-fallout graph (on linear scales). Following his practice of describing all parameters in probabilistic terms, he interprets A as the probability that the system will distinguish correctly between two items, one taken at random from C (the set of relevant documents), and the other from $N-C$ (non-relevant documents).

4. BROOKES' MEASURE

Brookes[11] considers that Swets' scale will be biased towards those lines with slope far from unity (45°). E is proportional (factor $\sqrt{2}$) to the length from the origin to the line along the negative diagonal; Brookes proposes instead the perpendicular length S from the origin to the line. Thus in Fig. 3.1, AB has $E=OB\times\sqrt{2}$, $S=OB$; CD has $E=OD\times\sqrt{2}$, $S=OF$. He relates the measure to the parameters of Swets' model; if the probability distribution of non-relevant documents with respect to z is normal with mean μ_1 and variance σ_1^2, and that of relevant documents has mean μ_2 and variance σ_2^2, then

$$E=\frac{\mu_2-\mu_1}{\frac{1}{2}(\sigma_1+\sigma_2)}, \text{ and } S=\frac{\mu_2-\mu_1}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}$$

As he says, the factor $(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}$ is much more satisfactory statistically than $\frac{1}{2}(\sigma_1+\sigma_2)$.

In Appendix A, I prove that under the assumptions of the model (both distributions normal), Brookes' measure S is equivalent to Swets' measure A, in particular S is the normal deviate of A.

Brookes considers that the individual parameters μ_1, σ_1, μ_2, σ_2 describe basic characteristics of the *IR* systems. So he wishes to find the values of these parameters for a system, and at the same time to test directly the hypothesis that the distributions are normal. This he does by plotting recall against level of co-ordination on normal probability × linear graph paper; if the result is a straight line, the distribution of relevant documents with respect to level of co-ordination is normal, and μ_2 and σ_2 can be obtained from the position and slope of the line. Similarly for fallout. He gives a graph of some Cranfield II results plotted in this way; both lines (for recall and fallout) are approximately straight, though not perhaps as good as the Swets lines.

It should be noted that the variable z has now taken on a much more fundamental significance: the distributions are with respect to z, so the exact interpretation of z is important. This is not so in the Swets method, where the actual values of z serve only to connect pairs of values of recall and fallout, and then drop out. Brookes interprets z simply as level of co-ordination; this introduces some problems, notably that z is continuous but

the level of co-ordination is discrete. He notes this, and says: 'However, for the present analysis, it suffices to imagine that underlying the discrete variable "level of co-ordination" there is a continuous variable z which conveniently assumes the value 1.00..., 2.00..., 3.00..., and so on, as the level of co-ordination takes the values 1, 2, 3,...' But this would imply that the distributions were concentrated on the integral points of the line z, and would certainly not be normal.

The alternative assumption would appear to be that the system does use a continuous variable, but the only cut-off points allowed are the integral ones. Thus a document retrieved at level of co-ordination 4 but not at level 5 might have a value of z anywhere in the range $4 \leq z < 5$; but you are not allowed to ask for all documents at level of co-ordination $4\frac{1}{2}$. This makes it clear that the discreteness of the level of co-ordination is a restriction on the system which must affect its value, though it does not affect the Swets line. It also suggests that this continuous variable z has a real meaning which might be revealed by further analysis. Until then, it is difficult to see exactly what the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ mean.

Apart from the Cranfield results, Swets also tests his theory on a number of results based on document output cut-off L, rather than level of co-ordination (e.g. the SMART system tests). It is not clear how Brookes' idea can be applied to such a system, since one clearly cannot interpret z simply as L. Yet the Swets lines are still straight. Clearly the nature of z requires further investigation.

However, if S is defined as the perpendicular distance from the origin to the Swets line, it is independent of the exact interpretation of z. Also, the fact that S (under this definition) is equivalent to A remains true.

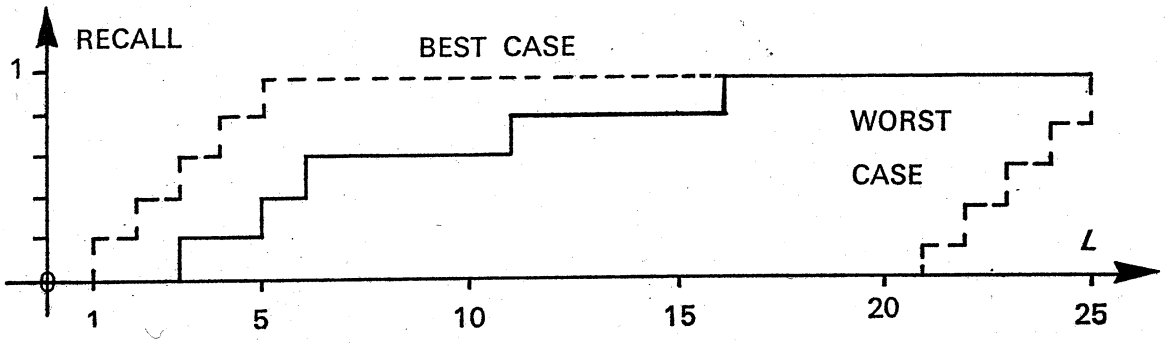
5. THE SMART SYSTEM MEASURES

Some overall measures proposed by Rocchio[12,13] have been used extensively in tests on the SMART system under Salton. The SMART system produces references in a rank order (thus it has a virtually continuous performance curve), and the user decides on a cut-off value of L (number of documents retrieved). Rocchio's measures are designed to be independent of cut-off, i.e. to satisfy my second requirement.

The first measure, normalized recall K, is calculated as follows. A graph is drawn of recall against L for a question; also given are the curves for the best and worst possible cases (for example see Fig. 5.1). K is then the area between the actual case and the worst case, as a proportion of the area between the best and worst cases. Its range is the 0 (worst case) to 1 (best case); random ranking gives a value of $\frac{1}{2}$.

Fig. 5.2 is the recall-fallout graph for the same question, on a distorted scale. The similarity between Figs. 5.1 and 5.2 is immediately clear; if each stratum of Fig. 5.1 is shifted until the best case is vertical, Fig. 5.2 is

obtained. It then becomes apparent that $K=A$, the area under the recall-fallout curve. This result is proved in Appendix B. Thus this measure also satisfies my first requirement.



L is document output cut-off; in this example, $N=25$, $C=5$, and the relevant documents are retrieved at ranks 3, 5, 6, 11, 16.

FIG. 5.1 *Calculation of normalized recall*

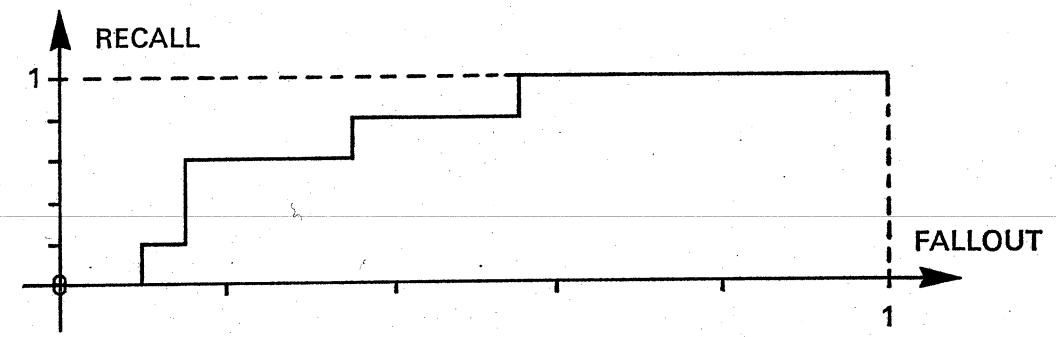


FIG. 5.2 *Calculation of A*

The above argument applies only to one question; K is normally calculated for each question and then averaged, whereas A is more usually calculated from some form of average performance curve. A and its variations are considered in more detail in §6.

The second of Rocchio's measures, normalized precision J, is calculated from the similar but more complicated graph of precision against L (see Fig. 5.3). This graph can be re-interpreted in various ways: *a*, a graph of recall against log L (Fig. 5.4), and *b*, a graph recall against log P (Fig. 5.5).

Fig. 5.4 shows that J is a measure of a distortion of the area measured by K, laying more emphasis on the high-precision end of the curve. Fig. 5.5 shows that J is related to the recall-precision curve; this relationship is, however, complicated by the fact that the worst case is not a simple curve, and also by the slightly curious interpolation of constant precision between the peaks of the curve. J does not satisfy my first requirement, that is it is not expressible in terms of recall and fallout alone.

It will be noted that both K and J, in spite of their names, are *overall*

measures, and as such very different from the ordinary recall and precision. It is therefore something of a surprise to find Rocchio proposing a 'normalized overall measure' which is defined as $5K+J-4$. I do not understand the purpose of this measure—I suppose it is designed to have less emphasis on

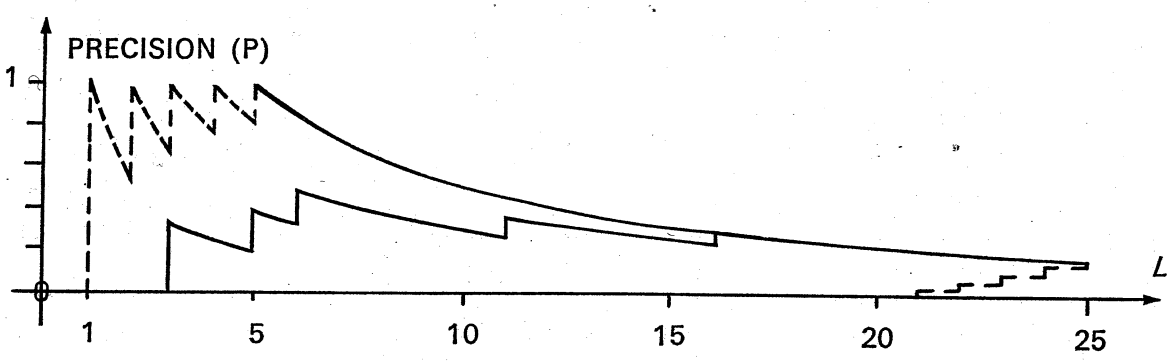


FIG. 5.3 *Calculation of normalized precision*

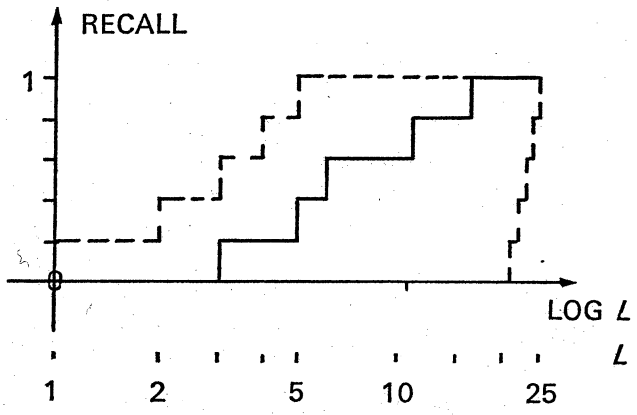


FIG. 5.4 *Normalized precision, second calculation*

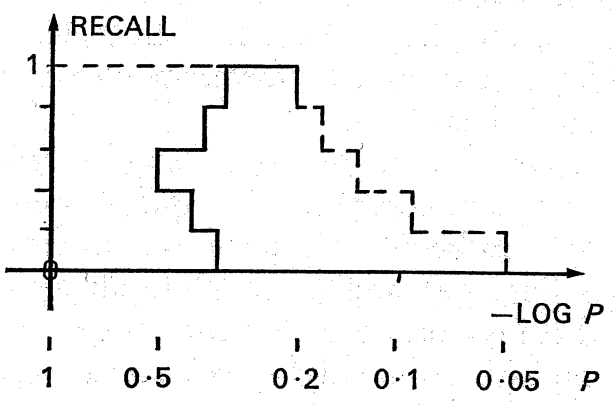


FIG. 5.5 *Normalized precision, third calculation*

the high-precision end of the curve than J (K emphasizes both ends equally), but there are obviously ways of doing that directly. It does not satisfy my first requirement.

Cleverdon and Keen[14] use another variation on normalized recall. It is calculated from a distortion of the *M-L* graph, similar to that for *J*, but simpler. There are several complications, notably that it was necessary to use a formula for randomly ordering the documents retrieved at a given level of co-ordination. There is a simpler way of doing this if a form of *A* is to be used (see §6). Again, my first requirement is not satisfied.

Rocchio proposes two further measures: rank recall, and log precision (not to be confused with log *P*, the logarithm of precision, which I used above). For a given question, rank recall assesses systems in the same order as *K*, but it is not suitable for comparisons between questions—that is to say, Rocchio realizes that this has a marked dependence on generality (but it does not occur to him to ask the same question of *K* and *J*). Log precision bears the same relations to *J* as rank recall does to *K*.

Thus it appears that the only satisfactory measure of the ones considered in this section is *K*; and this is the same as the Swets measure *A*.

6. THE SWETS MEASURE *A*

The results of the last two sections indicate that *A*, the area under the recall-fallout performance curve, is worthy of further consideration.

Swets interprets *A* as the probability that the system will distinguish correctly between two items, one taken at random from *C* and the other from $N-C$. Therefore, he says, the worst result to be expected will be given by a system which randomly orders the documents; then the recall-fallout graph is a straight line at 45° (the diagonal), and $A=\frac{1}{2}$. But this assumes that the system must produce the documents in a rank order; there are some that merely divide the collection in two or more parts. For these systems, *A* interpreted strictly as the probability above is given by the marked area in Fig. 6.1. However, if one assumes that any collection of documents retrieved at one go is randomly ordered, then *A* is given by the area in Fig. 6.2. If on the other hand, one calculates Brookes' measure *S*, this corresponds to the area under an idealized smooth curve, as in Fig. 6.3. Thus we have three versions of *A*: A_1, A_2, A_3.

I have explained (Part 1, §7) that I think the recall-fallout curve made up of straight line segments truly represents the value of the system. This can be demonstrated as follows. Consider a system consisting of points *T* and *V* in Fig. 6.4. Would the addition of point *U* add to the value of the system? I say not, as the system (with the help of random ordering) can already attain point *W* which is better than *U*. Definitions A_2 and A_3 agree with me; A_1 does not. Now consider the same problem in Fig. 6.5. Here I say that point *U* would contribute to the value of the system: without it the nearest the system can get to it is *W*, which is worse. A_2 agrees with me; A_3 does not. So I propose the definition: *A* is the area under the recall-fallout performance curve, where the points are joined by straight lines.

M is recall, *F* is fallout

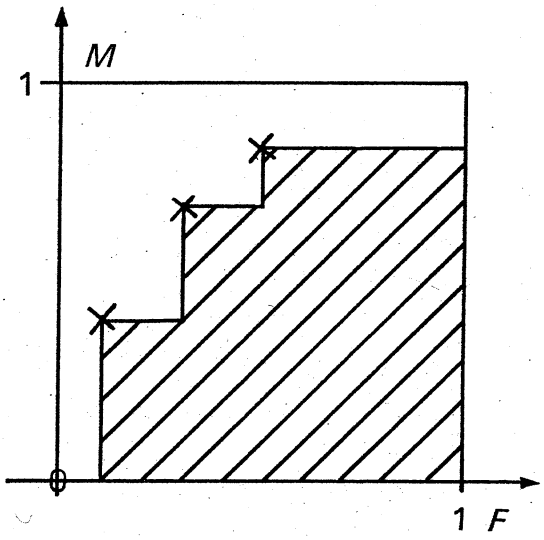


FIG. 6.1 *Calculation of* A_1

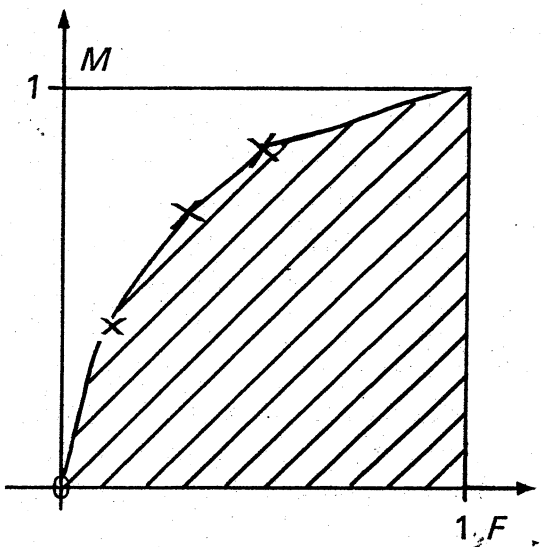


FIG. 6.2 *Calculation of* A_2

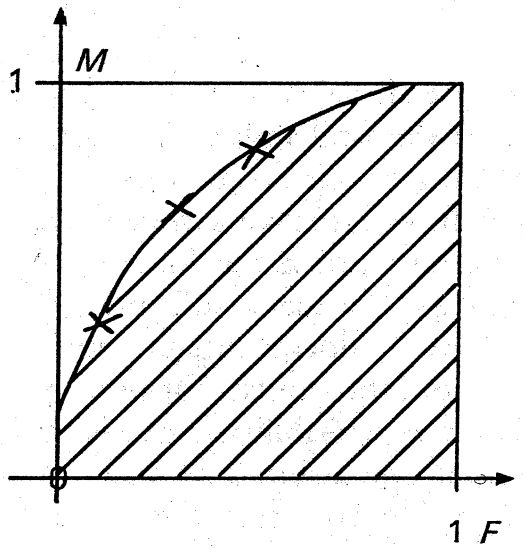


FIG. 6.3 *Calculation of* A_3

For practical purposes, a system with more than a few points will have very similar values of A_2 and A_3. The measure S has some important advantages: as Brookes[11] points out, its sampling distribution is known. Also the values it takes for real systems fall in a much more convenient range than those of A.

However, the difference between A_2 and A_3 is of theoretical significance. The fact that the Swets line is always straight is a strong indication that this line is a real and significant property of the system. Thus the measure S must mean something, though not the same thing as A. What then does it mean? I think this problem could be solved, in part at least, by an investigation into the real meaning, significance and validity of the Swets model.

This would involve examining the validity of the model for individual questions, and the problem of how to average results in order to preserve the parameters of the model. This last point is of particular significance: it does not appear to be possible to calculate S values for individual questions

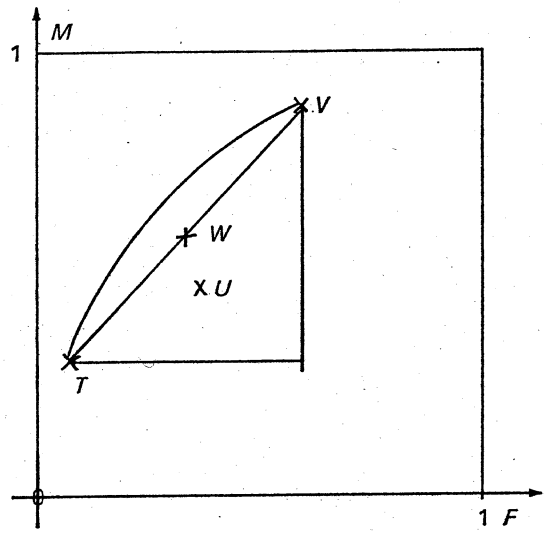


FIG. 6.4 A_1 *versus* A_2 *or* A_3

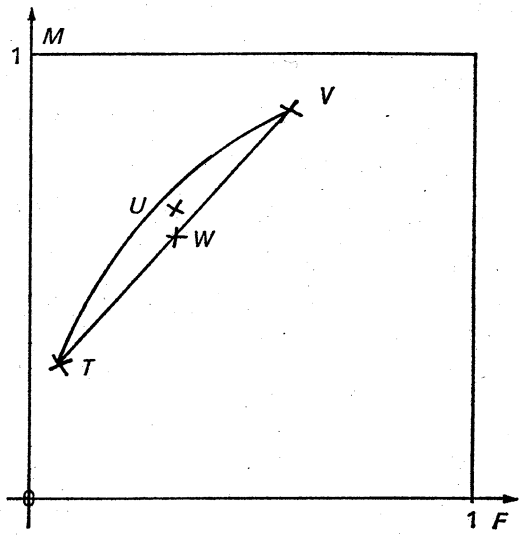


FIG. 6.5 A_2 *versus* A_3

and to average them, at least for the Cranfield II results, as most questions have too few acceptable points to give a valid Swets line (points with, say, $F=O$ do not appear on the Swets graph). But this might on general principles be the more desirable method. These problems require further investigation.

One point emerges from the chosen definition of A. If a system has only one point, i.e. simply divides the collection in two, and the point has values of recall M and fallout F, then a simple calculation gives

$$A=\tfrac{1}{2}(M-F+1)$$

If this quantity is normalized to lie in the range -1 to 1 (random value 0) instead of 0 to 1 (random value $\frac{1}{2}$), it becomes simply $M-F$. This is the basic measure of effectiveness used by Goffman and Newill[8] and others at Case Western Reserve University.

7. CONCLUSIONS

It seems that most reasonably satisfactory measures of effectiveness are directly or indirectly related to Swets' measure A. In particular, Brookes' measure S is a version of A; Rocchio's normalized recall is equal to A; the Case Western Reserve University measure is the version of A for a single-point system. Swets' result[10] that most Swets lines have slopes not far from 45° shows that his measure E is not very different from S. Rocchio's normal-

ized precision is a measure of a distortion of the same area as A; the same remark applies to Cleverdon's normalized recall.

Two of these versions, A and S, appear to be the most satisfactory. There are small differences between them which are of some theoretical significance but probably of no practical significance.

It seems to me, however, that overall measures are in general of doubtful value. It is not usually possible to say without qualification 'system X is better than system Y'; but this is what any overall measure tries to do. It is particularly relevant in this context that the Swets lines are not all at 45°: this means that they cross each other and that therefore system X may be better in one area of the graph, and system Y in another. One could clearly deal with such situations by having other overall measures which lay more stress on one end of the curve (the measure A is clearly capable of development in this respect); but this seems to be a clumsy way of dealing with the problem. It would seem more sensible to use the performance curve itself.

APPENDIX A

Theorem: *Under the assumptions of the Swets model, Brookes' measure S is equivalent to Swets' measure A.*

Let $f_1(z)$ and $f_2(z)$ be the probability distributions of non-relevant and relevant documents respectively, with respect to the Swets variable z. Then for a cut-off value c of z, we have

$$\text{recall} = \int_c^\infty f_2(z)d_z, \text{ and fallout} = \int_c^\infty f_1(z)d_z$$

The area under the recall-fallout graph, A, is then given by

$$A = \int_{-\infty}^{\infty} f_1(z) \int_z^\infty f_2(t)d_t d_z$$

(If document 1 is relevant, and document 2 is not, this formula also gives the probability that $z_1 > z_2$, i.e. that the system will retrieve document 1 before it retrieves document 2. This is Swets' interpretation of A).

The Swets model says that $f_1(z)$ and $f_2(z)$ are normal (gaussian), i.e. that

$$f_1(z) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[\frac{-(z-\mu_1)^2}{2\sigma_1^2}\right]$$

and

$$f_2(z) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[\frac{-(z-\mu_2)^2}{2\sigma_2^2}\right]$$

Then

$$S = \frac{\mu_2 - \mu_1}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}$$

In the above expression for A, make the following transformation:

$$z=\frac{\sigma_1(\sigma_2 v-\sigma_1 u)}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}+\mu_1$$

$$t=\frac{\sigma_2(\sigma_2 u+\sigma_1 v)}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}+\mu_2$$

The integrand $f_1(z)f_2(t)$ now reduces to

$$\frac{1}{\sigma_1\sigma_2\,2\pi}\exp\left(-\tfrac{1}{2}(v^2+u^2)\right)$$

The range of integration is the half-plane bounded by the line $z=t$, which is the line

$$u=\frac{\mu_1-\mu_2}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}=-S$$

Thus the range is: $-S\leq u\leq\infty$, $-\infty\leq v\leq\infty$

The Jacobian is $\sigma_1\sigma_2$

So

$$A=\frac{1}{2\pi}\int_{-S}^{\infty}\exp\left(-\tfrac{1}{2}u^2\right)d_u\int_{-\infty}^{\infty}\exp\left(-\tfrac{1}{2}v^2\right)d_v$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-S}^{\infty}\exp\left(-\tfrac{1}{2}u^2\right)d_u$$

Hence A is a strictly increasing function of S—in particular, S is the 'normal deviate' of A.

This result implies the following: if S is defined as the perpendicular distance from the origin to the Swets line, and A is the area under the recall fallout graph corresponding to this line, then the two measures are related as above. This latter result depends only on the fact that the Swets line is straight—it is independent of the exact form of the distributions $f_1(z)$ and $f_2(z)$.

APPENDIX B

Theorem: *In a system with ranked output, Rocchio's normalized recall as calculated for one question is equal to Swets' measure A.*

Let r_i be the rank of the i'th relevant document. Normalized recall is then defined to be

$$K=1-\frac{\Sigma r_i-\Sigma i}{C(N-C)}\quad\text{(summed over } i=1, 2, \ldots, C)$$

The recall-fallout graph is a series of steps, the vertical lines of which are given by

$$F=\frac{r_i-i}{N-C} \text{ when } \frac{i-1}{C}<M\leq\frac{i}{C}$$

So the area to the right of each of these lines is

$$\frac{1}{C}\left(1-\frac{r_i-i}{N-C}\right)$$

and the total area is

$$A=\sum\frac{1}{C}\left(1-\frac{r_i-i}{N-C}\right) \quad \text{(summed over } i=1, 2, \ldots, C)$$

$$=1-\sum\frac{r_i-i}{C(N-C)}$$

$$=1-\frac{\Sigma r_i-\Sigma i}{C(N-C)}$$

$$=K$$

REFERENCES

1. ROBERTSON, S.E. The parametric description of retrieval tests. Part 1: the basic parameters. *Journal of Documentation*, vol. 25, March 1969, p. 1–27.
2. VERHOEFF, J., GOFFMAN, W., *and* BELZER, F. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the Association for Computing Machinery*, vol. 4, December 1961, p. 557–8, 594.
3. GOOD, I. The decision-theory approach to the evaluation of information retrieval systems. *Information Storage and Retrieval*, vol. 3, no. 2, April 1967, p. 31–4.
4. SWANSON, D. R. Searching natural language text by computer. *Science*, vol. 132, 21 October 1960, p. 1099–1104.
5. BORKO, H. *A research plan for evaluating the effectiveness of various indexing systems.* Report no. FN–5649/000/01, System Development Corporation, Santa Monica, Calif., 1961 (23p.).
6. GIULIANO, V. E., *and* JONES, P. E. *Study and test of a methodology for laboratory evaluation of message retrieval systems.* Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, U.S. Air Force, L. G. Hanscom Field, Bedford, Mass., 1966 (183p.).
7. FARRADANE, J., DATTA, S., *and* POULTON, R. K. *Research on information retrieval by relational indexing. Part 1: methodology.* The City University, London, 1966 (60p.).
8. GOFFMAN, W., *and* NEWILL, V. A. A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval*, vol. 3, August 1966, p. 19–25.
9. SWETS, J. A. Information retrieval systems. *Science*, vol. 141, 19 July 1963, p. 245–50.
10. SWETS, J. A. *Effectiveness of information retrieval methods.* Report no. AFCRL–67–0412, Air Force Cambridge Research Laboratories, Bedford, Mass., 1967 (47p.).
11. BROOKES, B. C. The measure of information retrieval effectiveness proposed by Swets. *Journal of Documentation*, vol. 24, March 1968, p. 41–54.

12. 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.
13. ROCCHIO, J. Document retrieval systems—optimization and evaluation. Thesis published in: *Information storage and retrieval.* Report no. ISR–10, Computation Laboratory of Harvard University, Cambridge, Mass., 1966 (163p.).
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.).