Information Retrieval Systems

Statistical decision theory may provide a measure of effectiveness better than measures proposed to date.

John A. Swets

In the past 5 years, the period of intensive study of information-retrieval systems, ten different measures for evaluating the performance of such systems have been suggested. In this article I review these measures and propose another.

The various measures have much in common. Eight of them evaluate only the effectiveness (accuracy, sensitivity, discrimination) of a retrieval system and are derived completely, in one way or another, from the 2-by-2 contingency table of pertinence and retrieval represented in Fig. 1. The other three measures assess efficiency as well as effectiveness, by including such performance factors as time, convenience, operating cost, and product form.

In some of the measures, in each of the categories just mentioned, the variables considered are combined into a single number. In other measures, of each kind, the separation of two or more variables is maintained, and the numerical value of each is listed. Still other measures consist of a graph showing the relationship between two variables related to effectiveness.

The measure proposed here is one supplied by statistical-decision theory. It compresses the four frequencies of the contingency table into a single number, and it has the advantage that this single number is sufficient to generate a curve showing all of the different balances among the four frequencies that characterize a given level of accuracy. That is to say, this measure, given the validity of the model underlying it, provides an index of effectiveness that is invariant over changes in the breadth of the search query or in the total number of items retrieved. If desired, a second number can be extracted from the fourfold table to characterize the specific balance among the frequencies that results from any specific form of query. This measurement technique has the drawback, at present, that the model on which it is based has not been validated in the information-retrieval setting. The ground for optimism on this score is the fact that the model has been validated in analogous problems of signal detection studied in electrical engineering and psychology.

Before proceeding to the review and proposal, it should be stated that this article deals only with measures of merit—that is, with the dependent variables of an experiment conducted to evaluate one or more retrieval systems. It is not concerned with methodological issues (such as the number of items in the information store, the means of determining relevance, the number and qualifications of judges, and the form and number of queries) in the design of such experiments.

The author is senior scientist at Bolt Beranek and Newman, Inc., Cambridge, Mass., and associate professor of psychology (on leave) at Massachusetts Institute of Technology, Cambridge.

Review

In this review I present the measures to some extent in the terms of their originators and to some extent in common terms which will make it easier to compare and contrast them with the measure proposed here. A common vocabulary is achieved by coordinating the variables expressed in the various terms of different writers with the quantities of the contingency table as they are represented in Fig. 1. These translated quantities always appear in brackets. Thus, $[a]$ represents the number of pertinent items retrieved. A quantity such as $[a/a+c]$ represents the proportion of pertinent items retrieved and may be taken as an approximation to the conditional probability that a pertinent item will be retrieved—a probability denoted $[Pr_P(R)]$. It is convenient at times to refer to these quantities in words, so I define them as follows:

$a/a+c = Pr_P(R)$ = conditional probability of a "hit."

$b/b+d = Pr_{\bar{P}}(R)$ = conditional probability of a "false drop."

$c/a+c = Pr_P(\bar{R})$ = conditional probability of a "miss."

$d/b+d = Pr_{\bar{P}}(\bar{R})$ = conditional probability of a "correct rejection."

The first measure to be considered is one proposed by Bourne, Peterson, Lefkowitz, and Ford (*1*), a measure best described as "omnibus." These writers recommend determining a measure of agreement between an aspect of system performance and a related user requirement for each of approximately 10 to 12 requirements. The measures of agreement are then multiplied by weighting coefficients which represent the relative importance of the requirements, and the products are summed to achieve a single figure of merit.

Bourne and his associates report the results of what they regard as a preliminary investigation, and it is true that the measure has not yet been described in sufficient detail for application. The requirements to be included in the measure have deliberately not been fixed; it is suggested that they will represent such factors as amount of pertinent material missed $[c]$, amount of nonpertinent material provided $[b]$, delay, ease of communication, complexity of search logic accommodated, form in which items are delivered, and degree of assurance that items on a given subject do not exist. Bourne and his associates point out that there are not enough quantitative data available to apply the measure now, but they do present data which demonstrate that users tend to disagree on the relative importance of different user requirements. They do not justify the combining of many different kinds of variables on a single metric.

Bornstein (*2*), in discussing details of experimental design, proposes consideration of the following four variables: (i) the number of pertinent $[a]$, partially pertinent, peripherally pertinent, and nonpertinent $[b]$ responses to each question; (ii) the time spent by the user in examining materials in each of the four categories of pertinence; (iii) the proportion of acceptable substitutes for "hard copy" which are supplied in each of the four categories; and (iv) the actual coincidence and unique-

ness of responses on the four-point scale of pertinence.

Bornstein suggests that analyses of variance be carried out for variables i, ii, and iii for each point on the pertinence scale, the main effects attributed to the different retrieval systems being compared. He suggests that within-cell variance terms and certain interaction effects will also convey valuable information. This procedure yields only relative measures of effectiveness and efficiency, but Bornstein argues that using a known store of information, which is necessary if absolute measures are to be obtained, requires excessive effort and biases the comparison of different retrieval systems.

Wyllys (*3*) derives a single figure of merit for the efficiency of a search tool as employed at a given stage in a sequential search process. He proposes to obtain the product of four variables: (i) the "restriction ratio" (which is the number of items retrieved at stage k divided by the number of items retrieved at stage $k - 1$, or the reduction in the number of potentially pertinent items that is effected by the search tool); (ii) the cost of using the tool; (iii) the number of pertinent documents eliminated from further consideration [c]; and (iv) a loss function [K_2].

Wyllys states that the cost variable should be defined in accordance with the local situation to weight correctly such factors as time, money, inconvenience, and indexing costs, and that the loss function should incorporate the degree of pertinence of the pertinent documents eliminated.

Verhoeff, Goffman, and Belzer (*4*) propose a measure which may be translated as

$$M = [a(V_1) - b(K_1) - c(K_2) + d(V_2)],$$

in which [K_1] and [K_2] are non-negative constants, and for which pertinence is defined by each system user.

These writers would maximize the measure for a given retrieval system, over users with different opinions about pertinence, by defining a critical probability

$$[Pr = \frac{K_1 + V_2}{V_1 + K_1 + K_2 + V_2}]$$

for each item relative to each query, and by retrieving those items having a probability of pertinence greater than the critical probability. This procedure is germane to system design and use rather than to system evaluation, but I mention it because it bears some resemblance to concepts discussed later in this article.

	P	$\bar{P}$	
R	a V_1	b K_1	$a + b$
$\bar{R}$	c K_2	d V_2	$c + d$
	$a + c$	$b + d$	$a + b + c + d$

Fig. 1. The 2-by-2 contingency table of pertinence and retrieval. P and $\bar{P}$ denote, respectively, pertinent and nonpertinent items; R and $\bar{R}$ denote, respectively, retrieved and unretrieved items; a, b, c, and d represent the simple or weighted frequencies of occurrence of the four conjunctions; V_1 is the value of retrieving a pertinent item; V_2 is the value of not retrieving a nonpertinent item; K_1 is the cost of retrieving a nonpertinent item; and K_2 is the cost of failing to retrieve a pertinent item.

Swanson (*5*) has proposed a measure $M = R - pI$, where R is the sum of relevance weights of retrieved items divided by the sum of relevance weights (for a given query) of all items in the store; I is the effective amount of irrelevant material, and is defined as $N - LR$ where N is the total number of items retrieved [$a + b$] and L is the total number of pertinent items in the store [$a + c$]; and p is a penalty [K_1] which takes on arbitrary values.

Borko (*6*) presented a modification of Swanson's measure in which the irrelevancy score I is redefined. In Borko's measure, I is the number of nonpertinent items retrieved [b] divided by the number of items retrieved [$a + b$], or [$Pr_R(\bar{P})$]. Borko also prefers to do without the arbitrary penalty on nonpertinent retrievals so that the measure is bounded by $+1$ and -1.

Mooers (*7*) proposes three ratios: (i) the ratio of the number of crucially pertinent items retrieved to the number of crucially pertinent items in the store; (ii) the ratio of the number of pertinent or crucially pertinent items retrieved to the number of pertinent or crucially pertinent items in the store; and (iii) the ratio of the number of nonpertinent items retrieved to the number of nonpertinent items in the store.

In more general terms, Mooers proposes consideration of three conditional probabilities—of an important hit, of a hit, and of a false drop. He states that, for an excellent system, the first probability should be high and the last should be simultaneously low. He further states that a system is good if the first is near 1.0, and very good if the last is less than .01. The second variable is said to be of less importance than the other two; it should always closely approach 1.0.

Perry and Kent (*8*) consider several quantities found in the contingency table but focus on two: (i) the conditional hit probability, [$a/a + c$] or [$Pr_P(R)$], and (ii) the inverse hit probability [$a/a + b$] or [$Pr_R(P)$].

Cleverdon (*9*) considers the same two variables, which in his terms are the "recall ratio" and the "relevance ratio," respectively. He emphasizes the trading relationship (that is: as the inverse hit probability or relevance ratio increases, the hit probability or recall ratio can be expected to decrease) and suggests that it will be illuminating to plot the recall ratio against the relevance ratio. He expects that a curve will be generated, with an as yet unknown shape, as the restrictiveness of the search query is varied. He has speculated that the curve will look something like the one in Fig. 2. Thus, highly restrictive queries would lead to a relatively high relevance ratio and to a relatively low recall ratio; a less restrictive query would enhance the recall ratio at the expense of the relevance ratio.

Swanson's second proposal (*10*) is much like the last few discussed. From an experiment on automatic text searching he has plotted the percentage of pertinent items retrieved [or the conditional hit probability, $Pr_P(R)$] against the number of nonpertinent items retrieved [b]. His data follow the curve of Fig. 3. The data points forming this curve were generated by a procedure analogous to varying the requirements for proximity, in text, of key words, and by specifying either an "and" or an "or" relationship among a variable number of key words.

The proposal made in the next section, discussed later in detail, is that $Pr_P(R)$ be plotted against $b/b + d$—that is, against $Pr_{\bar{P}}(R)$—rather than against b, as Swanson has done, or against $Pr_R(P)$, as Cleverdon has done. One reason is that the quantities $Pr_P(R)$ and $Pr_{\bar{P}}(R)$ contain all the information in the fourfold table of pertinence and retrieval, because the other two,

$Pr_P(\bar{R})$ and $Pr_{\bar{P}}(\bar{R})$, are simply their complements. The major point, however, is that, for these axes, statistical-decision theory provides a family of theoretical curves, which are similar in appearance to Swanson's empirical curve, along with two parameters. One of the parameters, the one of major interest, reflects accuracy; it is an index of the distance of a curve from the positive diagonal, the latter corresponding to chance performance. The other parameter reflects the breadth of the query; it indexes a point on a curve by the slope of the curve at that point.

Thus, if it can be generally established that varying the breadth of the query in retrieval systems will generate a curve of the type provided by the theory (and this seems likely), then it is possible to combine the hit and false-drop probabilities, $Pr_P(R)$ and $Pr_{\bar{P}}(R)$, into a single number which is an absolute measure of retrieval-system accuracy, independent of the particular balance between the two probabilities that is struck by any particular form of query, and also to combine the two probabilities in another way into another single number to represent any particular balance between them.

Proposal

The relevance of statistical-decision theory to the problem of information retrieval has been observed before. Maron and Kuhns (*11*) have suggested "an interpretation of the whole library problem as one where the request is considered as a clue on the basis of which the library system makes a concatenated statistical inference in order to provide an ordered list of those documents which most probably satisfy the information needs of the user." Wordsworth and Booth (*12*) have applied game theory to the problem of deciding when to stop a search. Here I simply call attention to the possibility that statistical-decision theory will provide a measure of retrieval effectiveness that is preferable to the measures proposed to date.

Decision theory is addressed to the problem of assigning a sample, bearing evidence, to one or the other of two probability distributions—one of two (mutually exclusive and exhaustive) statistical hypotheses is to be accepted. An analogous problem is posed in classical problems of signal detection. The measure taken of the input to a detector, in a particular time interval, must be assigned to one of two events—the detection system reports either that noise (random interference) alone existed or that a specified signal existed in addition to the noise. Similarly, a retrieval system takes a measure of a given item in the store, relative to a particular query, in order to assign the item to one of two categories—the retrieval system rejects the item as not being pertinent or retrieves it.

Decision theory describes the optimal process for making the type of decision with which it deals (*13*). This process description has been translated directly into a functional specification of the ideal signal-detection device (*14*) and it has been found to represent quite accurately the behavior of human observers in a variety of detection and recognition tasks (*15, 16*).

The primary concern here is not with a process, or with system design, but rather with the measurement techniques that accompany the process description. The process model is presented here, though very briefly, because it provides a rationale for the measurement techniques. The model is described in the language of the retrieval problem to display one possible coordination between the elements of the model and the physical realities of retrieval. It is suggested, however, that the measurement techniques may be used to advantage whether or not this particular coordination seems entirely apt.

The model. Let us assume that when a search query is submitted to a retrieval system the system assigns an index value (call it z) to each item in the store (an item can be a document, a sentence, or a fact) to reflect the degree of pertinence of the item to the query. (Maron and Kuhns have described a particular procedure to accomplish this assignment, but let us regard such a procedure, in general, as a feature of all retrieval systems.) Now it may be that for a given need, or for the need as translated into a search query, the items in a given store do in fact vary considerably in pertinence, from a very low value (or no pertinence) to a very high value (or full satisfaction of the need). On the other hand, all of the items may in fact (according to expert opinion or the user's opinion) be either clearly nonpertinent or clearly pertinent to the need. In either case the retrieval system, being imperfect, will view the items as varying over a range of pertinence; indeed, because of the error which will exist in any retrieval system, the value of z assigned to a nonpertinent item will frequently be higher than the value of z assigned to a pertinent item.

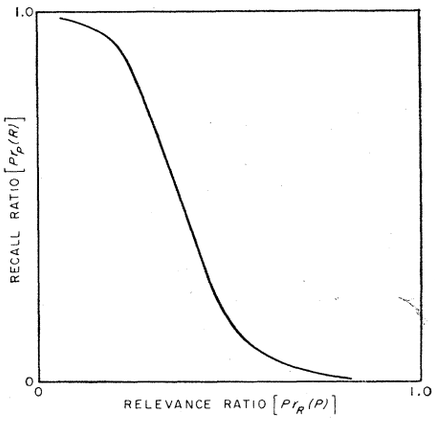

Fig. 2. The plot suggested by Cleverdon (*9*) to show the trading relationship between the proportion of pertinent items which are retrieved and the proportion of retrieved items which are pertinent.

Thus, we assume that the retrieval system assigns a fallible index of pertinence, z, and that there exists, apart from the retrieval system, a knowledge of which items are "in truth" pertinent and nonpertinent. We may speculate that the situation is similar to that depicted in Fig. 4. The abscissa represents the degree of pertinence as indexed by z. The ordinate shows the probability of assignment of each value of z. The lefthand function, $f_{\bar{P}}(z)$, represents the distribution of values of z assigned to nonpertinent items, and the righthand function, $f_P(z)$, represents the distribution of values of z assigned to pertinent items. It should be noted that an umpire can determine

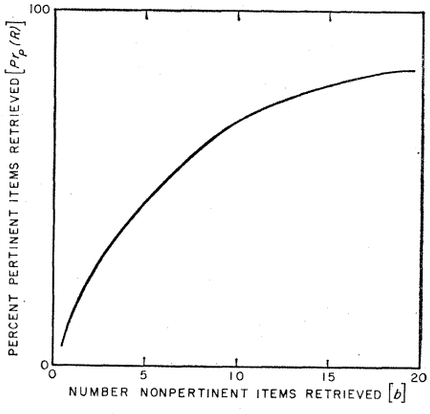

Fig. 3. Curve representing an idealization of data reported by Swanson (*10*).

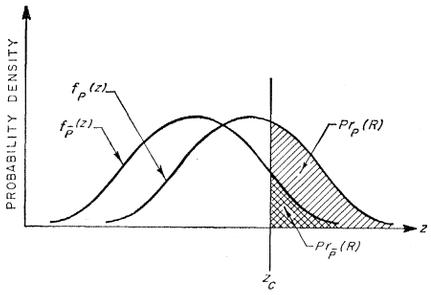

Fig. 4. A representation of the probability distributions, on the index of pertinence z, related to nonpertinent and pertinent items, and of the retrieval criterion z_c.

the condition P or $\bar{P}$ (that is, he can classify all items as being pertinent or nonpertinent) either because, in his opinion, they all fall clearly into one of these two categories or by virtue of selecting arbitrarily a cutoff along a continuum of pertinence.

Figure 4, as described so far, is intended only to portray the assumptions that the values of z assigned to $\bar{P}$ items vary about a mean, that the values of z assigned to P items vary about a higher mean, and that the two distributions overlap. The two distributions are shown as normal and of equal variance. These assumptions, if justified empirically, facilitate the calculation of a measure of effectiveness, but they are not necessary.

It is clear from the representation of the problem in Fig. 4 that if the retrieval system is to accept (retrieve) or reject each item on the basis of the index value z associated with it, a criterion of acceptance must be established by, or for, the system. A cutoff value of z, denoted z_c, must be established such that all items with $z > z_c$ are retrieved and all items with $z < z_c$ are rejected. It may also be seen in Fig. 4 that a trading relationship exists between the conditional probability of a hit, $Pr_P(R)$ [represented by the area under $f_P(z)$, to the right of z_c], and the conditional probability of a false drop, $Pr_{\bar{P}}(R)$ [the area under $f_{\bar{P}}(z)$ to the right of z_c]. If, for example, the acceptance criterion is made more lenient—that is, if z_c is moved to the left —$Pr_{\bar{P}}(R)$ is increased at the expense of increasing $Pr_{\bar{P}}(R)$.

Just where, along the z axis, the cutoff is best set is determined by the values and costs appropriate to a particular retrieval need. If the user is willing to examine a good deal of nonpertinent material in order to reduce the chance of missing a pertinent item, the cutoff should be low. Alternatively, if time or money is an important factor and a miss is not very serious, the cutoff should be high. Similarly, certain a priori probabilities may affect the level of the desired cutoff. If the user has good reason to believe the store contains the item he wants, he may choose to make a relatively thorough search; if he is doubtful that the store contains the item he requires, he may prefer a token search, of only the items most likely to be responsive to his query. In practice, the level of the cutoff may be set, though imprecisely to be sure, by the choice of a form of query. The choice of an "and" or "or" relationship among a number of key terms, and the selection of the number of key terms, are ways of determining the breadth of the query and thus the level of the z-axis cutoff.

In the next section I derive a measure of the basic effectiveness of a retrieval system that is independent of the level of the acceptance criterion. If the general representation of the problem in Fig. 4 is valid, then the measure presented is the only one that serves the purpose. Measures of the proportion of hits, or the proportion of hits and correct rejections, or the proportion of hits minus the proportion of false drops, or other measures of this kind, are not adequate, and simply observing various values of two variables, such as the proportion of hits and the proportion of false drops, is an unnecessarily weak procedure.

Of course, the assumption that a real retrieval system has a constant effectiveness, independent of the various forms of queries it will handle, is open to question. It seems plausible, however, that the sharpness of the retrieval system's query language, and its depth of indexing, and also the heterogeneity of items in the store, will determine a level of effectiveness that is relatively invariant over changes in the form of the query. In any event, the assumption is subject to empirical test, and its importance is sufficient to justify the effort of testing. Again, it may be that retrieval systems vary considerably in their ability to handle different forms of query—that is, to adopt different acceptance criteria—some present systems being relatively inflexible. Differences in this respect come under the heading of efficiency, as opposed to effectiveness, and can be taken into account separately. This kind of flexibility will probably be a standard feature of future retrieval systems.

Derivation of the measure. The proposal made here amounts to a recommendation that retrieval-system performance be analyzed by means of the

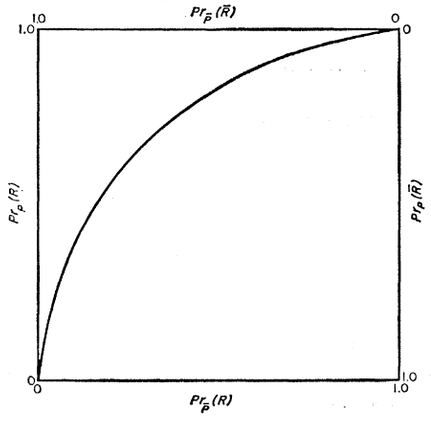

Fig. 5. A typical operating-characteristic curve.

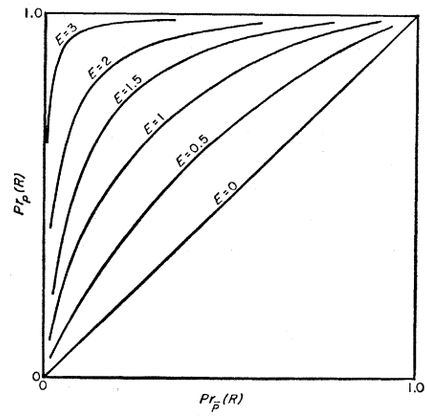

Fig. 6. A family of operating-characteristic curves, based on normal distributions of equal variance, with values of the parameter E.

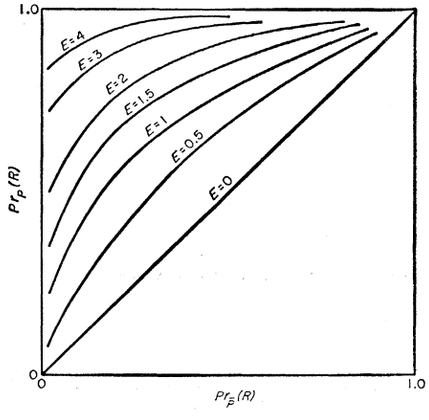

Fig. 7. A family of operating-characteristic curves based on an assumption of increasing variance.

operating characteristic as used in statistics. One form of the operating characteristic is the curve traced on a plot of $Pr_P(R)$ versus $Pr_{\bar{P}}(R)$ as the z-axis cutoff varies. One such operating-characteristic curve, calculated on the basis of the assumptions that the underlying probability distributions (of Fig. 4) are normal and of equal variance, is shown in Fig. 5. The complementary probabilities, of a correct rejection and of a miss, $Pr_{\bar{P}}(\bar{R})$ and $Pr_P(\bar{R})$, are also included in Fig. 5 to emphasize the fact that a complete description of the retrieval system's performance can be obtained from an operating-characteristic curve.

It is evident that a family of theoretical operating-characteristic curves can be drawn on the coordinates of Fig. 5, bounded by the positive diagonal and the upper left-hand corner, that correspond to different distances between the means of the two probability distributions, $f_{\bar{P}}(z)$ and $f_P(z)$. Since the distance between the means of these two distributions reflects the ability of the retrieval system to segregate nonpertinent and pertinent items, the parameter of this family of curves will serve as a measure of effectiveness.

If the curves are normal, they can be characterized by a single parameter —namely, the distance between the means of the two probability distributions divided by the standard deviation of the distribution of nonpertinent items,

$$\frac{{}^{M}f_P(z) - {}^{M}f_{\bar{P}}(z)}{{}^{\sigma}f_{\bar{P}}(z)}$$

It does no harm to adopt the convention that this standard deviation is unity, and then the parameter is just the difference between the means. This measure, which in this context I shall call E, is simply the normal deviate. It is easily obtained from a table of areas under the normal curve. Figure 6 shows a family of operating-characteristic curves and the associated values of E.

It is possible to relax the assumption of equal variance of the two distributions which was made in drawing the curves of Fig. 6. By way of illustration, Fig. 7 shows a family of theoretical curves calculated on the basis of the assumption that the variance of $f_P(z)$ increases with its mean—in particular, that the ratio of the increment in the mean to the increment in standard deviation is equal to 4.0.

I next describe a convenient way of

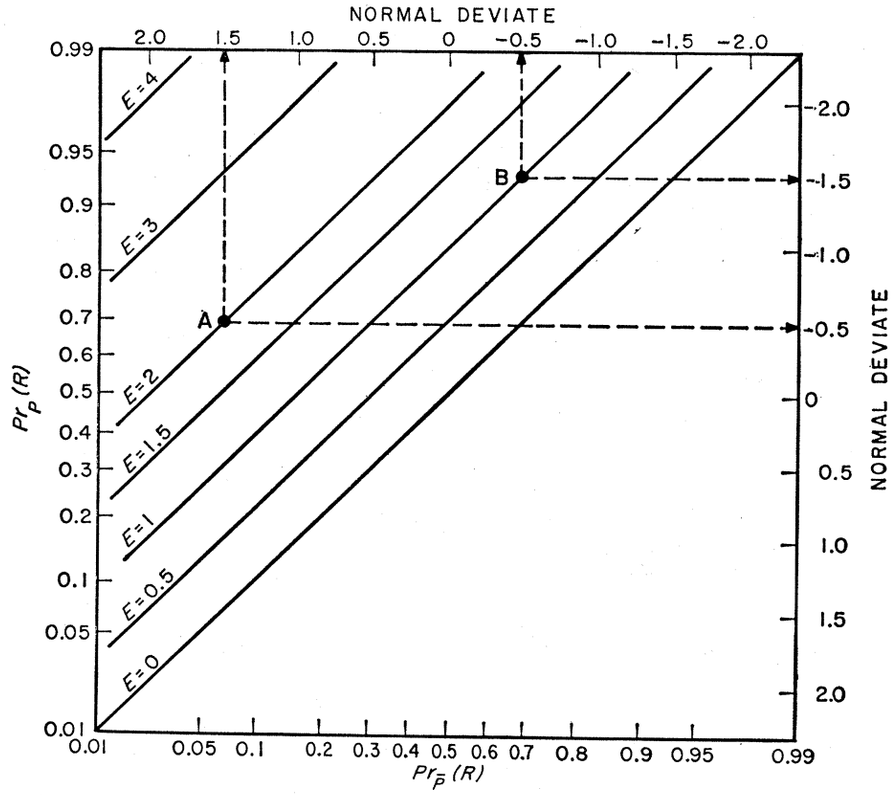

Fig. 8. Normal operating-characteristic curves plotted on double-probability graph paper.

calculating E in practice, and show how an additional parameter may be obtained, if necessary, to represent the ratio of variances. For the present, it is important to note that the assumption of normality will probably be adequate, and that curves based on quite extreme variance ratios differ very little. It will, in fact, be difficult to obtain enough data to reject the normality assumption or to distinguish among similar assumptions about variance. With the variance ratio as a free parameter, a normal operating-characteristic curve can be drawn to fit closely any steadily rising function. That a steadily rising curve will be obtained in practice is a reasonable expectation; Swanson's data (Fig. 3) certainly fit such a curve.

Having seen how a given level of sensitivity or effectiveness can be measured by a single number E, independent of a particular acceptance criterion, let us observe in passing that the slope of the curve at any point will serve as an index of the particular acceptance criterion, and of the breadth of the search query, which yielded that point. Strictly speaking, it is assumed in statistical theory that the z axis of Fig. 4 is a scale-of-likelihood ratio, $f_P(z)/f_{\bar{P}}(z)$, and then the value of the slope of the operating-characteristic curve at any point is exactly the value of likelihood ratio at which the acceptance criterion must be set to produce that point. Moreover, if the a priori probabilities [$Pr(P)$ and $Pr(\bar{P})$] and the values and costs are known, then the optimal setting of the acceptance criterion is at the value of likelihood ratio equal to

$$\frac{Pr(\bar{P})}{Pr(P)} \cdot \frac{(V_2 + K_1)}{(V_1 + K_2)}$$

However that may be, the formal assumptions and the full quantitative power of statistical theory have not been emphasized here; although they may ultimately be found to be of value in the retrieval application, the use of the suggested measures does not depend on this finding. The slope may simply be taken as the measure of the acceptance criterion, empirically, without concern for its precise theoretical basis. It is clear that the difference in the slopes at two points of a steadily rising function is a straightforward measure of the effective change in the breadth of a search query.

In brief, the operation of a retrieval system yields entries in the cells of a

fourfold contingency table and thus yields estimates of four conditional probabilities, two of which are independent. For any given query or form of query, these two probabilities can be plotted as a point in the unit square of Fig. 6.

The parameter of the theoretical curve on which the point falls is a measure of retrieval system effectiveness; the slope of the curve at that point is a measure of query breadth. It is expected that the various fourfold tables which result from systematically varying the breadth of the queries addressed to a given retrieval sytem will generate a steadily rising function similar to the ones shown in Fig. 6.

Specifics of measurement techniques. The most convenient way of converting the conditional probabilities of a hit and of a false drop into a value of E is to plot them on normal coordinates—that is, on probability scales transformed so that the normal deviates are linearly spaced (for example, Codex Graph Sheet No. 41,453). On these scales, as illustrated in Fig. 8, the normal operating-characteristic curve becomes a straight line. The points indicated as A and B illustrate that the difference between the normal-deviate values, one taken from the abscissa and one from the ordinate, is equal to E.

The lines shown in Fig. 8, having unit slope, are based on probability distributions, $f_{\bar{r}}(z)$ and $f_r(z)$, of equal variance. In general, the reciprocal of the slope (with respect to the normal-deviate scales) is equal to the ratio of the standard deviation of $f_r(z)$ to the standard deviation of $f_{\bar{r}}(z)$. Thus, the curves of Fig. 7, if plotted on these scales, would show slopes of less than unity and a decrease in slope with increases in E.

It should be noted that the procedure just given for calculating the value of E—that is, taking the difference between the two normal-deviate values—is not adequate when the operating-characteristic curve has a slope other than unity, for then a single curve will produce different values of E, depending on which point along the curve is chosen. A satisfactory convention often followed in this case is to determine the value of E from the point where the curve crosses the negative diagonal.

Validating requirements. The validity of the decision-theory model and measurement techniques for a given retrieval system can be tested by determining the operating-characteristic curve experimentally. Four or five data points, spread over a range of $Pr_{\bar{r}}(R)$ from approximately 0.10 to 0.90, will establish whether or not the curve rises steadily, and, if it does, these four or five points will establish its slope. Each data point must, of course, be based on a large enough sample to provide a fairly reliable point estimate of a probability. This is admittedly a large number of data, more than many investigators would at present consider economically feasible to obtain. There is, unfortunately, no substitute for adequate numbers of data if retrieval systems are to be evaluated on an empirical basis. Perhaps a small offsetting consideration is the fact that results can be pooled for queries of the same breadth, or of the same logical form.

It is possible to conduct further, and stronger, tests of validity. For retrieval systems that calculate something like the index of pertinence, z, additional information can be obtained by determining directly the probability distributions that underlie the operating characteristic. Other tests exist which may be applied appropriately to retrieval systems that do not provide an index comparable to z. An extensive testing program, originally designed for the study of signal detection in psychology, could be directly translated and applied to retrieval systems. A description of this program may be found elsewhere (*16*). For most purposes, however, a determination of the operating-characteristic curve should be adequate.

If the empirical operating-characteristic curve obtained from a given retrieval system is reasonably well fitted by a linear function on normal-deviate coordinates, the measure E is appropriate to represent the effectiveness of that system. It is to be expected, on rational grounds, that the model will be found to apply generally to a variety of retrieval systems. If this proves to be the case, there will be many current applications of the measure E. This outcome would greatly facilitate performance-cost analysis of available retrieval systems (*17*).

References and Notes

1. C. P. Bourne, G. D. Peterson, B. Lefkowitz, D. Ford, *Stanford Res. Inst. Proj. Rept. No. 3741* (1961).
2. H. Bornstein, *Am. Doc.* **12**, 254 (1961).
3. R. E. Wyllys, *Trans. Congr. Inform. System Sci., Hot Springs, Va., 1st* (1962).
4. J. Verhoeff, W. Goffman, J. Belzer, *Commun. Assoc. Computing Machinery* **4**, 557 (1961)
5. D. R. Swanson, *Science* **132**, 1099 (1960).
6. H. Borko, *System Development Corp., Santa Monica, Calif., Field Note No. 5649/000/01* (1961).
7. C. N. Mooers, *Zator Company, Cambridge, Mass., Tech. Note No. RADC-TN-59-160* (1959).
8. J. W. Perry and A. Kent, Eds., *Tools for Machine Literature Searching* (Interscience, New York, 1958), pp. 3–18.
9. C. W. Cleverdon, *Association of Special Libraries and Information Bureaux, Cranfield, England, Interim Rept.* (1962).
10. D. R. Swanson, paper presented at the Congress of the International Federation of Information Processing Societies, Munich (1962).
11. M. E. Maron and J. L. Kuhns, *J. Assoc. Computing Machinery* **7**, 216 (1960).
12. H. M. Wordsworth and R. E. Booth, *Western Reserve Univ. Tech. Note No. 8, AFOSR-TN-59-418* (1959).
13. A. Wald, *Statistical Decision Functions* (Wiley, New York, 1950).
14. W. W. Peterson, T. G. Birdsall, W. C. Fox, *IRE (Inst. Radio Engrs.) Trans. Information Theory* **4**, 171 (1954); D. Van Meter and D. Middleton, *ibid.*, p. 119.
15. W. P. Tanner, Jr., and J. A. Swets, *Psychol. Rev.* **61**, 401 (1954); J. A. Swets, *Psychometrika* **26**, 49 (1961).
16. J. A. Swets, *Science* **134**, 168 (1961).

17. The preparation of this article was supported by a grant from the Council on Library Resources, Inc. Lewis and Judith Clapp gave helpful advice and greatly facilitated the review of relevant literature.