AFCRL-67-0412

EFFECTIVENESS OF INFORMATION RETRIEVAL METHODS

John A. Swets

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

Contract No. AF19(628)-5065
Project No. 8668

Scientific Report No. 8

15 June 1967

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627, Amendment 2

Contract Monitor: Stanley R. Petrick
Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

AFCRL-67-0412

EFFECTIVENESS OF INFORMATION RETRIEVAL METHODS

John A. Swets

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

Contract No. AF19(628)-5065
Project No. 8668

Scientific Report No. 8

15 June 1967

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627, Amendment 2

Contract Monitor: Stanley R. Petrick
Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

ABSTRACT

Results of some fifty different retrieval methods applied in three experimental retrieval systems were subjected to the analysis suggested by statistical decision theory. The analysis validates a previously-proposed measure of effectiveness and demonstrates its several desirable properties. The examination of a wide range of data in relation to this one metric provides a clear and general assessment of the current state of the retrieval art, and shows that the art is still far from what might be considered a desirable state.

A desirable measure of retrieval performance would have the following properties. First, it would express solely the ability of a retrieval system to distinguish between wanted and unwanted items -- that is, it would be a measure of "effectiveness" only, leaving for separate consideration factors related to cost or "efficiency." Second, the desired measure would not be confounded by the relative willingness of the system to emit items -- it would express discrimination power independent of any "acceptance criterion" employed, whether the criterion is characteristic of the system or adjusted by the user. Third, the measure would be a single number -- in preference, for example, to a pair of numbers which may covary in a loosely specified way, or a curve representing a table of several pairs of numbers -- so that it could be transmitted simply and immediately apprehended. Fourth, and finally, the measure would allow complete ordering of different performances, indicate the amount of difference separating any two performances, and assess the performance of any one system in absolute terms -- that is, the metric would be a scale with a unit, a true zero, and a maximum value. Given a measure with these properties, we could be confident of having a pure and valid index of how well a retrieval system (or method) were performing the function it was primarily designed to accomplish, and we could reasonably ask questions of the form "Shall we pay X dollars for Y units of effectiveness?".

In a previous article I reviewed ten measures that had been suggested prior to 1963, and proposed another (1). None of the ten measures, and none that has come to my attention since then, has more than two of the properties just listed.

Some of them, including those most widely used, have the first two properties, and some of the others have the last two properties. The measure I proposed, one drawn from statistical decision theory, has the potential to satisfy all four desiderata. At the time it was proposed, however, the decision-theory measure had not been applied to any empirical retrieval results, so that its assumptions about the form of retrieval data had not been tested. In the present paper we examine this measure in relation to test results obtained from three experimental retrieval systems with some fifty different retrieval methods. With minor qualifications, the data are uniformly consistent with the assumptions of the decision-theory measure, and quite clearly demonstrate its usefulness. A substantive outcome of the extensive analysis in terms of this measure is a clear appraisal of the current state of the retrieval art. The analysis shows in precise terms how much room for improvement is left by current retrieval techniques. The room for improvement, as we shall see, is large.

Before proceeding to a review of the decision-theory measure and to an examination of the data, let us consider briefly the domain of the measure and a disclaimer about the scope of this paper.

The measure is most clearly applicable to retrieval systems that deal in documents or messages, and it is applied here to systems of this type. Less clearly perhaps, but as well, the measure can be applied to information systems that handle facts, or give answers to ordinary English questions. In both cases queries are addressed to a system and the system's responses to

the queries must be evaluated. Whether the response is a set of documents, or a fact selected or deduced from a collection of writings, is immaterial. Appropriate text must be isolated in either case, to constitute the response or to supply the base from which the response is drawn. The data represented by the decision-theory measure are entries in a two-by-two contingency table: just as documents suited or unsuited to a need may be retrieved or not retrieved, so facts that correctly or incorrectly answer questions may be presented or withheld. For some relatively simple fact systems, of course, such as airline-reservation systems, discrimination or correctness is not a problem; the reference here is to fact systems in which the facts to be retrieved are not all neatly isolated, and in which the questions are not all anticipated in detail.

This measure, like those used most often in the past, is most directly applicable when the entire information store is known, when, in particular, the number of items appropriate as responses to each query is known. This condition is frequently satisfied in experimental systems, which usually contain no more than a few thousand items. If the measure is to be applied to stores large enough to make impractical a complete knowledge of them, three alternatives exist for estimating the required number. One is to select, by some heuristic process or by fiat, that subset of the full store likely to contain almost all of the items appropriate to a given set of queries, and to examine the subset in detail. A second alternative, used in one instance in the following, is simply to sample the large store and to extrapolate from the sample. A third alternative, used in another instance in the following, is to preselect certain items

from the store and to design test queries specifically to retrieve those items.

Application of the decision-theory measure assumes that the "relevance" of any item in the store to a given query, or user's need, can be determined. As the reader will know, or can imagine, the definition of relevance is generally regarded in the retrieval field as a very thorny problem, and even the concept itself has at times come under attack. However that may be, the definition of relevance is an issue separate from the measure under consideration, and is not discussed here. Nor is the concept defended here; I take it for granted that it is essential to the evaluation of retrieval performance and that sooner or later we shall come to terms with it. For our present purposes, we can accept the definitions of relevance adopted by the investigators who collected the data we shall examine, just as we accept for the present purposes other experimental procedures they have followed. It will become clear, by the way, that the decision-theory measure can be applied when judges use several, rather than two, categories of relevance, and that it uses to full advantage the output of a system that ranks or otherwise scales all items in the store according to their degree of relevance to the query at hand.

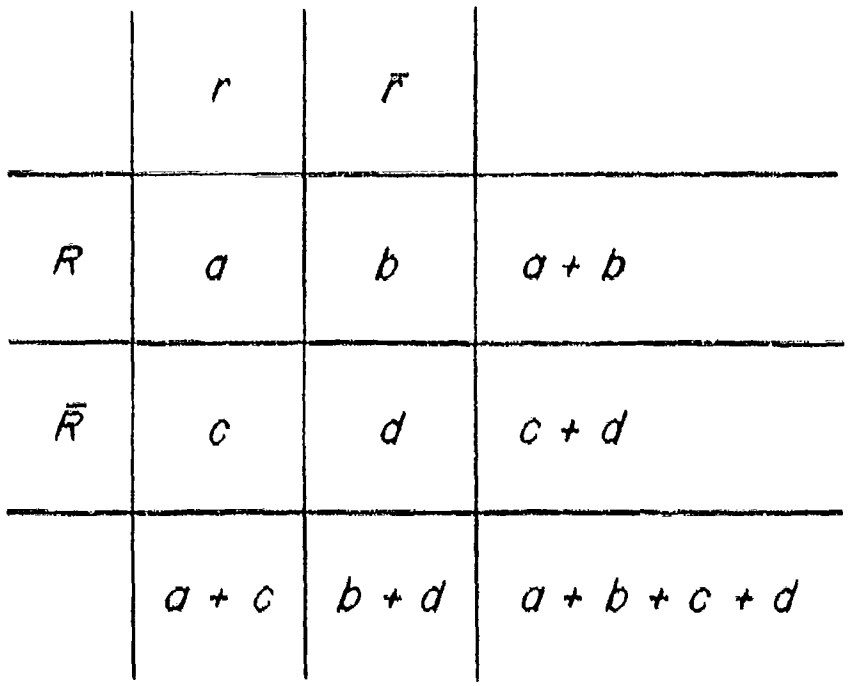
Decision-Theory Measure

A good way to begin in reviewing the decision-theory measure is to consider a measure more familiar in the retrieval context and to note the differences between the two. The measure used far more than any other (2) consists of two quantities

termed the "recall ratio" and the "precision ratio." Like other measures that attempt to assess only retrieval effectiveness, this measure can be described by reference to the relevance-retrieval contingency table shown in Fig. 1.

The recall ratio is defined as $a/a+c$, the number of items both relevant and retrieved divided by the number of items relevant. This ratio, then, is the proportion of relevant items retrieved, and it may be taken as an estimate of the conditional probability that an item will be retrieved given that it is relevant. The precision ratio (formerly called the "relevance ratio") is defined as $a/a+b$, the number of items both relevant and retrieved divided by the number of items retrieved. This ratio is the proportion of retrieved items deemed relevant, and an estimate of the conditional probability that an item will be relevant given that it is retrieved.

Now, if a system's effectiveness is characterized by two numbers, a value of the recall ratio and a value of the precision ratio, we know relatively little about the system, for one reason because we don't know how the two quantities relate to each other. What does it mean, for example, to say that a system yielded a recall ratio of 0.70 and a precision ratio of 0.50? If System A performs this way, and System B yields a recall ratio of 0.90 and a precision ratio of 0.40, is System B more or less discriminating than System A? That is, is a gain of 0.20 in recall and a loss of 0.10 in precision good or bad? Of course, should System B show a gain in both recall and precision over System A, we know B's effectiveness is superior to A's, but, in general,



	r	$\bar{r}$	
R	a	b	$a + b$
$\bar{R}$	c	d	$c + d$
	$a + c$	$b + d$	$a + b + c + d$

Fig. 1. The relevance-retrieval contingency table: $\underline{r}$ and $\underline{\bar{r}}$ denote, respectively, relevant and irrelevant items; $\underline{R}$ and $\underline{\bar{R}}$ denote, respectively, retrieved and unretrieved items; $\underline{a}$, $\underline{b}$, $\underline{c}$, and d represent frequencies of occurrence of the four conjunctions.

the measure consisting of this pair of quantities will give only a partial ordering of different systems, or of different methods employed by one system.

The problem here is that System A's recall of 0.70 and precision of 0.50 represents only one of the many balances between the two ratios that it can achieve. This balance might have occurred when an item had to satisfy five descriptors specified in a query in order to be retrieved. If this requirement is changed, so that now an item has only to satisfy any two of the query's five descriptors, it is very likely that more items will be retrieved, and that recall will go up and precision will go down. But we must know exactly how recall and precision will covary, along with variation in the acceptance criterion, if uncertainties are to be avoided in attempting to rank different systems or methods.

A solution to this problem, one that is sometimes adopted, is to test each system with several acceptance criteria and to present as the measure of a system's effectiveness the empirical curve so generated. Extensive tests have shown (3) that the empirical curve will resemble in form the curve shown in Fig. 2. If System A yields the curve shown while System B yields another curve everywhere above and to the right of the one shown, it is clear that B is superior to A.

However, these curves do not tell us, in general terms, by how many units B is superior to A (we can determine that B's precision is greater than A's by some specific percentage at some specific value of recall, but this number varies widely as

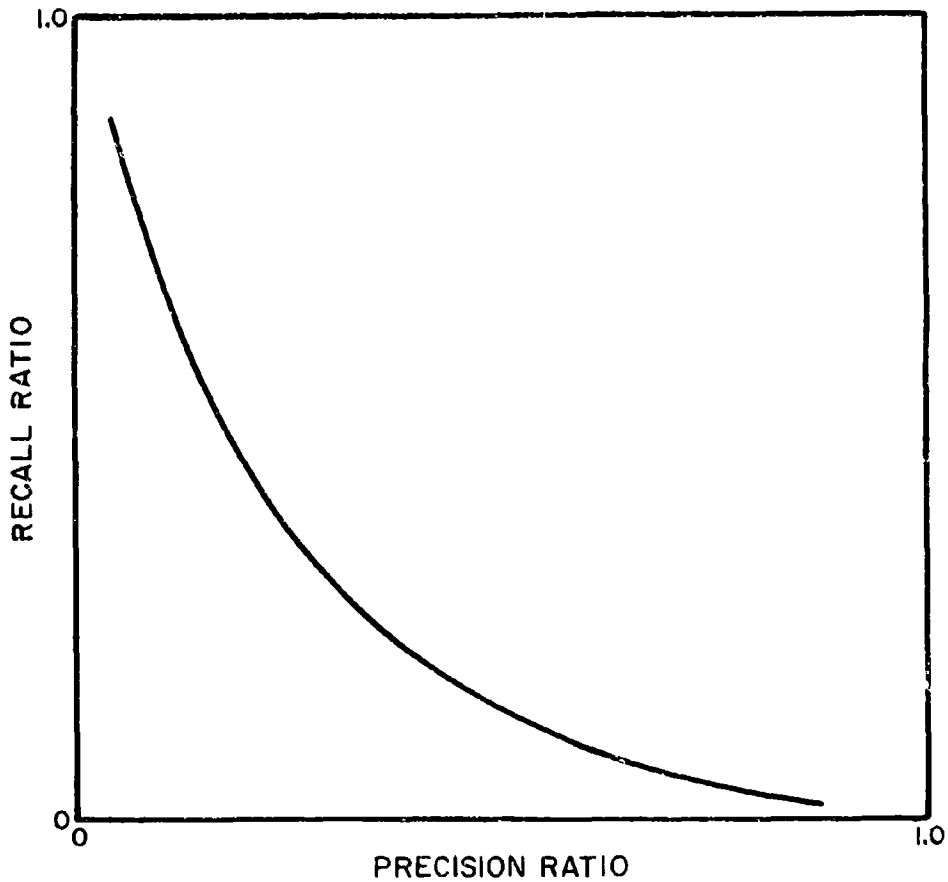


Fig. 2. Idealized example of an empirical recall-precision curve, fanned out by varying the acceptance criterion. For lenient criteria, recall is high and precision is low. Progressively more stringent acceptance criteria increase precision at the expense of recall.

a function of the value of recall selected). Nor can we tell from the curves how good either system is in absolute terms. And, of course, it is relatively awkward (we might say that a large "bandwidth" is required) to transmit and receive a full curve.

A measure that retains the basic information inherent in the recall-precision curve, and at the same time overcomes the drawbacks of using a curve as a measure, would be attained if there is a way to represent completely an empirical curve of this general sort by a single number on a scale with a unit, a true zero, and a maximum. The thrust of my earlier article was that statistical decision theory offers a way -- indeed, several ways. Whether or not we can take advantage of one of them, or to what extent, depends upon the form of retrieval data when analyzed by decision-theory techniques, and that form is the concern of this paper.

Though a way might be found to completely characterize any empirical recall-precision curve by a single number on the type of scale desired, decision theory suggests using the curve that results when another variable is substituted for precision. The variable to be substituted, in the terms of Fig. 1, is $\underline{b}/\underline{b}+\underline{d}$. This quantity is the number of items both irrelevant and retrieved divided by the number of items irrelevant, or the proportion of irrelevant items retrieved, and is an estimate of the conditional probability that an item will be retrieved given that it is irrelevant.

As in the earlier article (1), I refer to the retrieval of an irrelevant item as a "false drop." Also for consistency, the

retrieval of a relevant item is termed a "hit," so instead of the term "recall ratio" I use "the conditional probability of a hit." Some of the notation used here differs from that of the previous article. Here, as seen in Fig. 1, lower-case letters, r and $\bar{r}$, designate relevant and irrelevant items, while upper-case letters, R and $\bar{R}$, designate retrieved and unretrieved items. The two conditional probabilities of principal interest are here denoted $P(R|r)$ and $P(R|\bar{r})$. In the present notation, the curve we shall consider has $a/a+c$ or $P(R|r)$ on the ordinate and $b/b+d$ or $P(R|\bar{r})$ on the abscissa. This curve is a form of the "operating characteristic" used in statistics, or "OC curve."

One consideration in choosing the two variables used in decision theory, which are derived from the two columns of the relevance-retrieval contingency table, is that they contain all of the information in the table; the remaining quantities of the table ("misses" and "correct rejections") are, respectively, their complements. The recall and precision ratios are derived from a column and a row of the table and do not serve to specify the remainder of the table.

A related, but more salient, consideration is that using the two variables of decision theory permits us to draw upon several models of the retrieval process which stipulate different forms that empirical OC curves may take. That is, each of several available models developed within decision theory precisely specifies a given form for a theoretical OC curve. Or rather, each model specifies a family of OC curves having an index of effectiveness as the parameter. Conveniently, the OC curves of all but one of the models devised to date are straight lines or

very nearly straight lines when plotted on linear normal-deviate, or "probability," scales. A single number is adequate as an index of effectiveness, because it is sufficient to generate the entire curve, under those models that assume some fixed relationship between the degree of effectiveness and the slope of the curve. Generality is gained at the cost of a second parameter in one model that permits a variable relationship between effectiveness and slope. Still another model gives a one-parameter fit to data without regard to the slope, or, for that matter, without regard to the general form of the OC curve, but this number is not sufficient to regenerate the curve from which it is taken. We turn now to a description of these alternative models, and then to the retrieval data that will enable us to choose from among them the one or ones that will be useful.

The general decision model. Though the assumption is not essential to their application, I shall assume in describing the alternative decision-theory models that for each query submitted to a system, the system in some manner assigns an index value (call it z) to each item in the store to represent the degree of relevance of the item to the query. Plotting separately for irrelevant and relevant items the probability of assignment of each value of z yields two probability density functions. One form the two density functions might have is depicted in Fig. 3. The left-hand function is associated with irrelevant items, $f(z|\bar{r})$, and the right-hand function is associated with relevant items, $f(z|r)$.

If, as suggested in the figure, any given value of z might be assigned by the system to an item that is relevant or to an item that is irrelevant (as judged by a user or other umpire), then, as

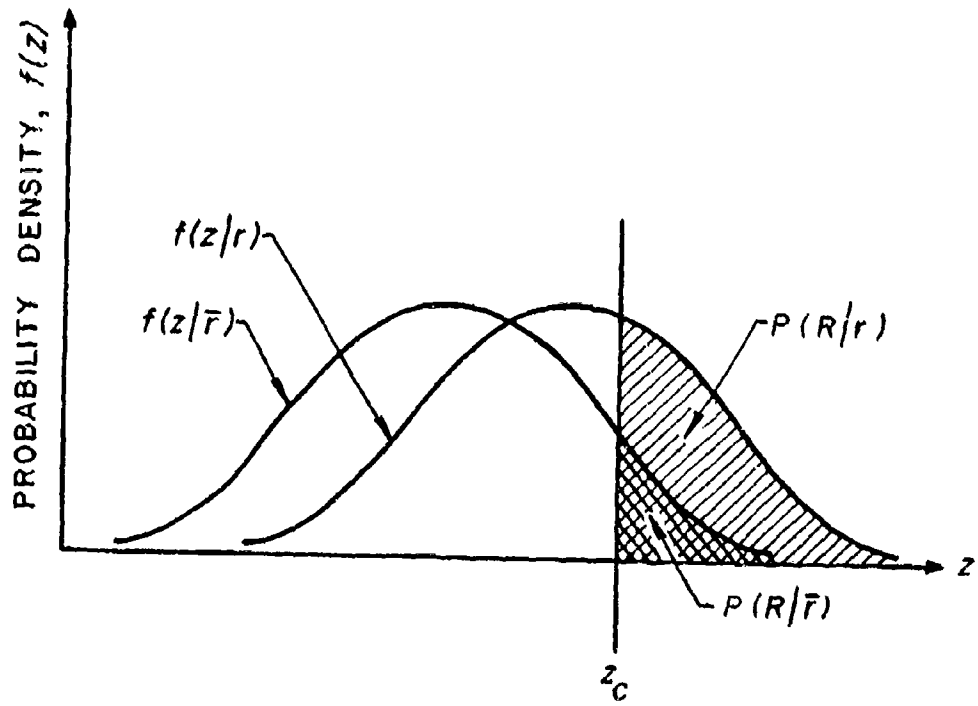


Fig. 3. One possible representation of the density functions for relevant and irrelevant items. The abscissa is the index of relevance, z, assigned by the system to each item. An acceptance criterion is labeled z_c.

shown, some criterion value of z, denoted z_c, should be adopted, such that items assigned values greater than z_c are retrieved while items assigned values less than z_c are not retrieved. The areas under the two density functions to the right of z_c represent the probabilities of retrieving irrelevant and relevant items. They are the coordinates of the OC curve, $P(R|\bar{r})$ and $P(R|r)$.

Any given separation between the two density functions represents a stable retrieval system, with some particular capacity to distinguish between relevant and irrelevant items, or some particular degree of effectiveness. For a fixed separation between the density functions, variation in the acceptance criterion z_c will result in a particular OC curve. Another system or method, with greater or lesser ability to discriminate relevant and irrelevant items, will yield a different OC curve as the acceptance criterion is varied.

The exact form of an OC curve, it is clear, depends upon the shapes of the density functions that underlie it. Various measurement models are generated by hypothesizing density functions of different shapes.

Gaussian, equal-variance model. The density functions shown in Fig. 3 are Gaussian and of equal variance. Given the separation shown, variation in the acceptance criterion will trace the OC curve labeled $E = 1$ in Fig. 4. The measure E is defined as the difference between the means of the two density functions divided by their common standard deviation. If the separation is increased so that the difference between the means is twice as

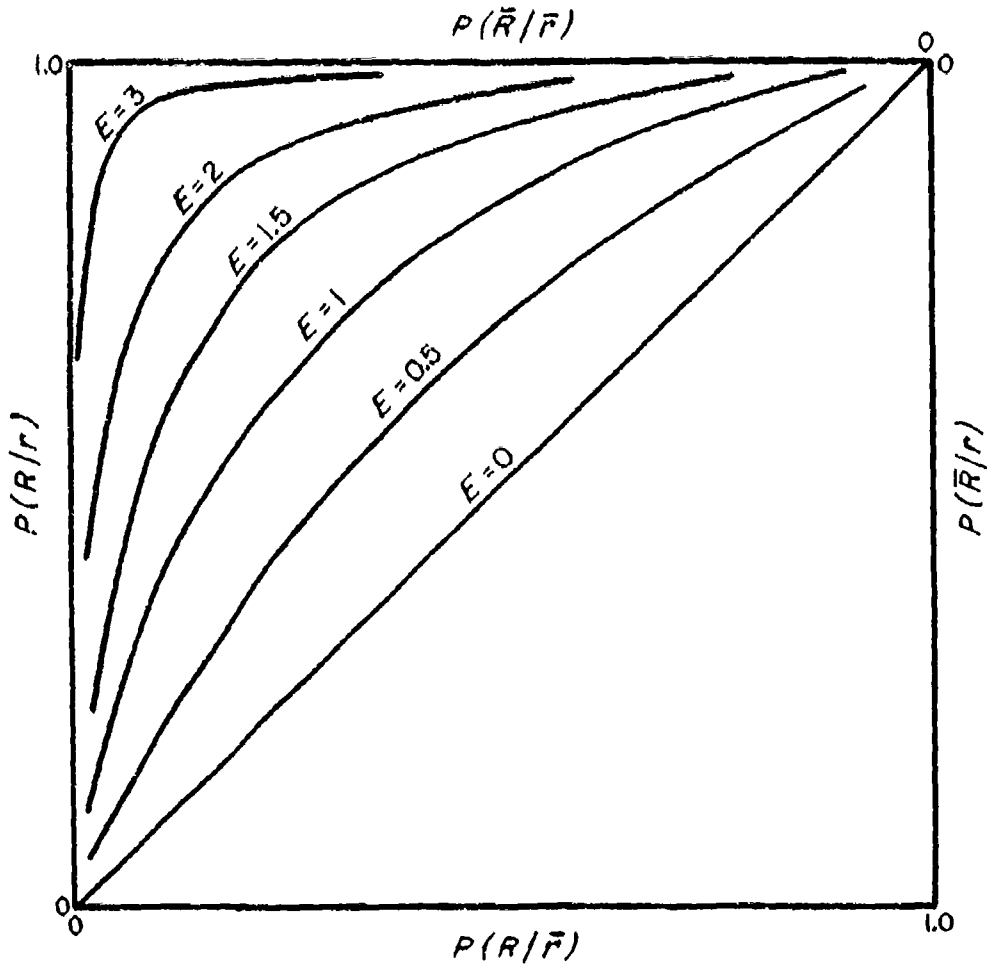


Fig. 4. A family of operating-characteristic curves, based on Gaussian density functions of equal variance, with values of the parameter E. Labels on the upper and right-hand scales indicate that the full relevance-retrieval contingency table can be recovered from the plot.

great as that shown in Fig. 3, then criterion variation will produce the OC curve labeled E = 2 in Fig. 4.

We see that empirical data obtained from a test of a retrieval system could be plotted in the space of Fig. 4. If the data points followed the contour of one of the curves shown, or one of the intermediate curves not shown, the label on that curve would completely describe the effectiveness of the system -- knowing the single number permits reconstruction of the entire curve.

It is more convenient to plot data fitted by the OC curves of Fig. 4 on probability scales, that is, on axes scaled linearly for the normal deviate, for then these OC curves are straight lines with unit slope, as shown in Fig. 5. The measure E for any curve can be read from the normal-deviate scales; one simply subtracts the value on the right-hand scale from the value on the top scale corresponding to any point on the curve. In Fig. 5, E is also scaled along the negative diagonal.

It can be seen that for practical purposes E has a maximum of approximately 5.0 -- though the axes could be extended to show higher values of E, effectiveness is not really at issue for retrieval systems yielding a hit probability greater than 0.99 and, simultaneously, a false-drop probability less than 0.01. There is the additional fact that reliable estimation of such extreme probabilities demands a sample of excessive size.

Gaussian, unequal-variance model. If the density functions are Gaussian, but of unequal variance, the OC curves on the scales

of Fig. 5 will be linear with slopes other than unity. In particular, the slope of the OC curve is equal to the ratio of the standard deviation of $f(z|\bar{r})$ to the standard deviation of $f(z|r)$.

For density functions of unequal variance, E must be redefined, for it was previously defined in terms of a standard deviation common to the two functions. Note that for OC curves of non-unit slope, the value of E obtained by subtracting a normal-deviate value on the right scale from one on the top scale is not constant along the curve. The definition adopted here consists in normalizing the difference between the means of the two density functions by their average standard deviation; this definition is reflected by measuring E at the intersection of the OC curve and the negative diagonal of the OC space.

Now, at least two alternatives are open to us. If we find that the slopes of empirical OC curves vary without regard to E (measured at the intercept of the negative diagonal), two parameters will be needed to fit the curve. Reconstruction of the curve will require reporting the value of the slope, s, in addition to the value of E. It could turn out, on the other hand, that s bears some fixed relation to E, for example, that s increases regularly as E increases. This would be the case if the ratio of the increment in the mean of $f(z|r)$ to a decrement in its standard deviation were a constant. If this constant were a stable property of a given retrieval system, it could be reported once, and then the single value of E would be sufficient to describe the various curves the system produces as a result of changes in one or another independent variable.

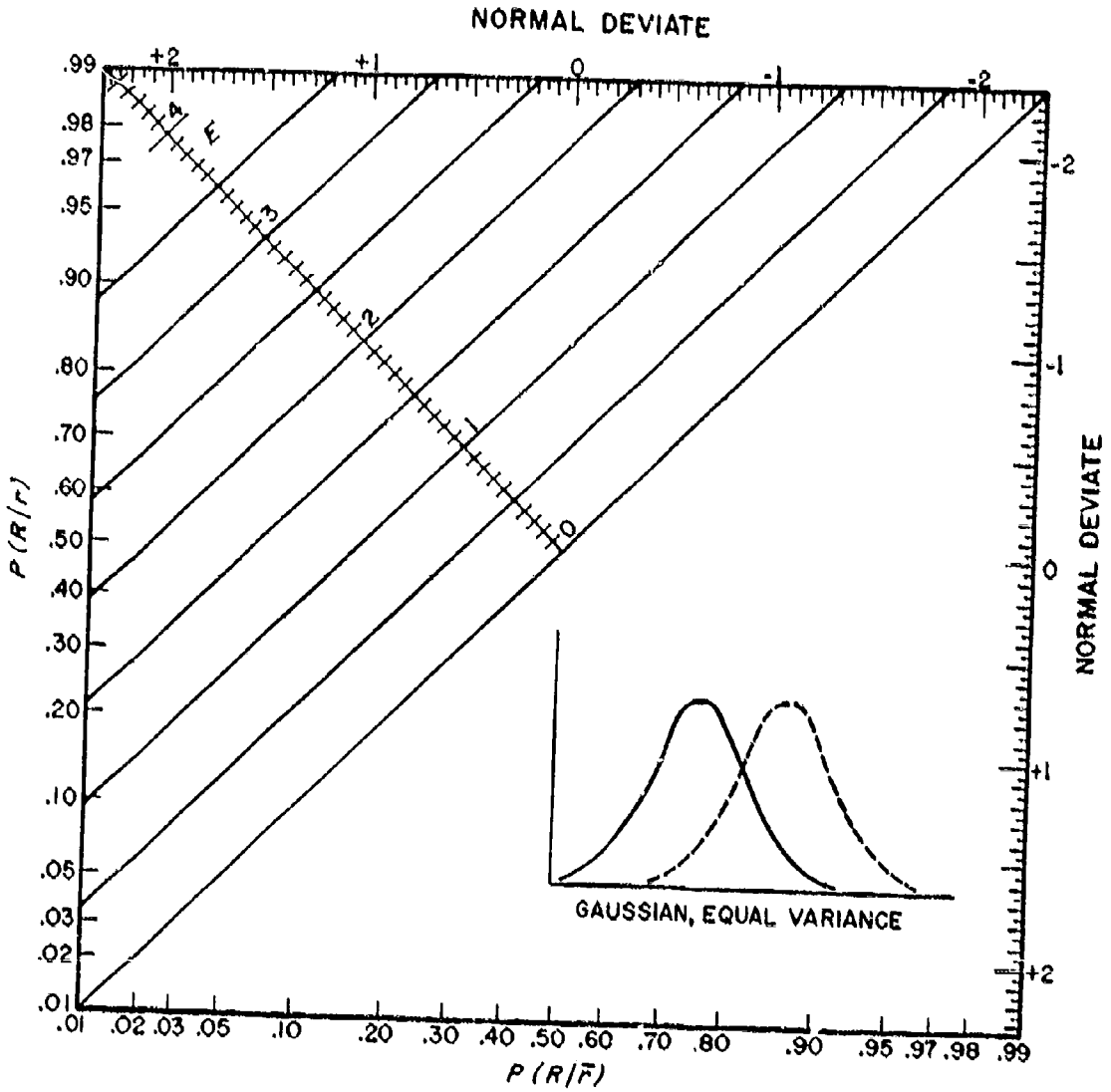


Fig. 5. The operating-characteristic curves of Fig. 4 on probability scales, that is, on axes scaled linearly for the normal deviate. Density functions inserted at lower right identify the basis of these OC curves in Gaussian, equal-variance density functions.

Exponential model. Simply as an illustration of further modelling possibilities, consider hypothesizing that the density functions are exponential in form, as shown at the lower right in Fig. 6. Then, again, the OC curve is essentially linear on probability scales and can be described by a single parameter. The parameter $K = \sqrt{k}$ is defined in the figure; for $k > 1.0$, the OC curves have the property that s decreases regularly as the effectiveness (K) increases.

Distribution-free model. If, after looking at data, hypothesizing some particular form of the density functions, and hence of the OC curve, seems too strong a procedure, we can resort to a measurement scheme that leaves these forms unspecified and free to vary. We can take as the measure of effectiveness the percentage of the area of the OC space that falls beneath any empirical OC curve, when plotted on linear scales (as in Fig. 4). This measure, call it A, will vary from 50% for a curve that follows the positive diagonal, representing equal hit and false-drop proportions or no discrimination, to 100% for a curve that follows the extreme left and top coordinates of the graph, representing a hit proportion of 1.0 at a false-drop proportion of 0.0 or perfect discrimination. The measure A, though a simple summary measure of effectiveness, does not permit reconstruction of the empirical curve from which it is drawn. It has the property useful for conceptual purposes that the value of A is equal to the percentage of correct choices a system will make when attempting to select from a pair of items, one drawn at random from the irrelevant set and one drawn at random from the relevant set, the one that is relevant. As demonstrated elsewhere (4) this equality holds for OC curves of any form.

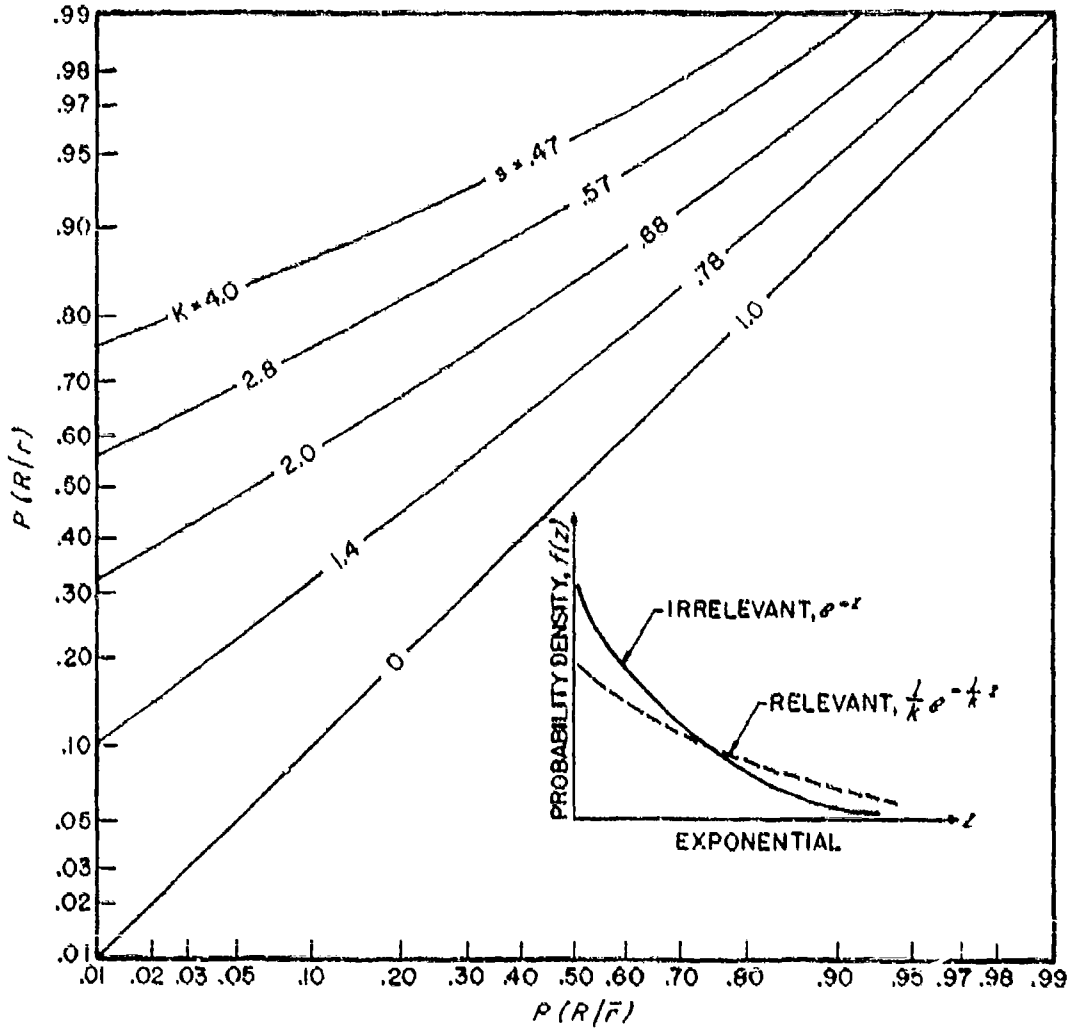


Fig. 6. A family of OC curves based on exponential density functions, plotted on probability scales.

Data

The three sets of data we shall examine were collected, respectively, at the Computation Laboratory of Harvard University by Gerard Salton (now at Cornell University) and Michael Lesk; under the Aslib project at Cranfield, England, by Cyril Cleverdon and Michael Keen; and at Arthur D. Little, Inc., by Vincent E. Giuliano and Paul E. Jones. These data were originally presented in technical reports published in late 1966 (3,5,6).

Salton and Lesk, and Giuliano and Jones, kindly made their raw data available to me so that I could calculate the hit and false-drop proportions. Cleverdon and Keen presented these quantities in their report. Though they are not responsible for the outcome, one or more of the authors of each report discussed with me the problem of measurement and commented on a draft of this paper. Their cooperation was essential, and I am pleased to acknowledge their very helpful advice and criticism.

Plots of data following are identified by the various terms for independent variables used in the original reports, to make possible cross references, but the terms are not defined here. Similarly, our present purposes do not require a description of the procedures of the three sets of experiments. However, a brief characterization of the scopes of the studies will be helpful in evaluating the general conclusions drawn here.

At Harvard, the questions asked experimentally include these: "can automatic text processing methods be used effectively to replace a manual content analysis; if so, what parts of the

documents [titles, abstracts, full text] are most appropriate for incorporation into the analysis; is it necessary to provide vocabulary normalization methods to eliminate linguistic ambiguities; should such normalization be handled by means of specially constructed dictionaries, or is it possible to replace thesauruses by statistical word association methods; what dictionaries can be used most effectively for vocabulary normalization; is it important to provide hierarchical subject arrangements, as is done in library classification systems; alternatively, should syntactical relations between subject identifiers be preserved; does the user have an important role to fulfill in controlling the search procedure" (5, pp. I-3, I-4). The experimental retrieval system, which operated on an IBM 7094 computer, was fully automatic in most applications; content-analysis procedures incorporated into the system processed documents and queries in natural language with no prior manual analysis. Stores of items used consisted of four collections of documents in three subject fields: documentation, aerodynamics, and computer sciences.

Experiments at Cranfield were based on manual analysis of documents. They were conducted to examine several different index languages -- some languages using single terms, others based on concepts, and others based on a thesaurus; the exhaustivity of indexing; the level of specificity of index terms; a gradation of relevance assessments; and the amount of intelligence applied in formulating search rules. Two collections used consisted of documents in aerodynamics and aircraft structures.

The experiments at Arthur D. Little, Inc., evaluated manual and automatic indexing; length of the query; coordinate retrieval methods; and retrieval methods based on statistical word associations, with and without human intervention. The system operated on an IBM 1401 computer, with fully automatic indexing in most applications. All items in the file were abstracts of reports in the aerospace field.

Data from the three sources lead to the same conclusions about the usefulness of a decision-theory measure, so the analyses of the three sets of data will be presented with little evaluative comment prior to a general discussion of results. Each of the OC plots is made on probability scales. Most of the plots summarize the results of one method of retrieval used with a given system; a few of them summarize the results of a single query used with a given method. The first question we ask is whether or not the plots of data are adequately fitted by straight lines. If they are, then we are interested in the slopes of the lines.

Harvard-Cornell data. All of the data I obtained from the Harvard-Cornell project are presented here; this set includes almost all of the data collected under the project before June of 1966, the major exception being some collected toward the end of that time in tests permitting iterative searches under the user's control.

The system at Harvard, called "SMART," assesses the relevance of each item in the store to each query addressed to the system. Print-outs of data containing the relevance index for

each item are, of course, extensive, and are not usually obtained; therefore we can not examine directly the shapes of the density functions. The standard print-out lists for each query the code number of every item relevant to it, and the rank value of each of these items in a list ordered (by the system) according to degree of relevance. Data in this form permit adopting, for purposes of analysis, each of several arbitrary acceptance criteria according to the total number of items considered as retrieved. That is, $P(R|r)$ and $P(R|\bar{r})$ are calculated in turn, for example, for the 5 items ranked highest, the 10 items ranked highest, the 15 items ranked highest, and so forth, terminating at an arbitrary point.

To gain a relatively stable sample, results are combined for all queries used with a single method. One can pool results before calculating $P(R|r)$ and $P(R|\bar{r})$, or alternatively, calculate these quantities for each query and take their average. The first of these procedures was followed in the analyses reported here.

Figure 7 shows the results for the collection of items in the subject field of documentation (called the ADI collection), under each of 6 retrieval methods. As in subsequent figures, in order to conserve space, only a portion of the OC space is shown for each plot; the last panel in the figure reproduces the lines of the previous panels on the full OC space. These lines, in all cases, were fitted to the data by eye.

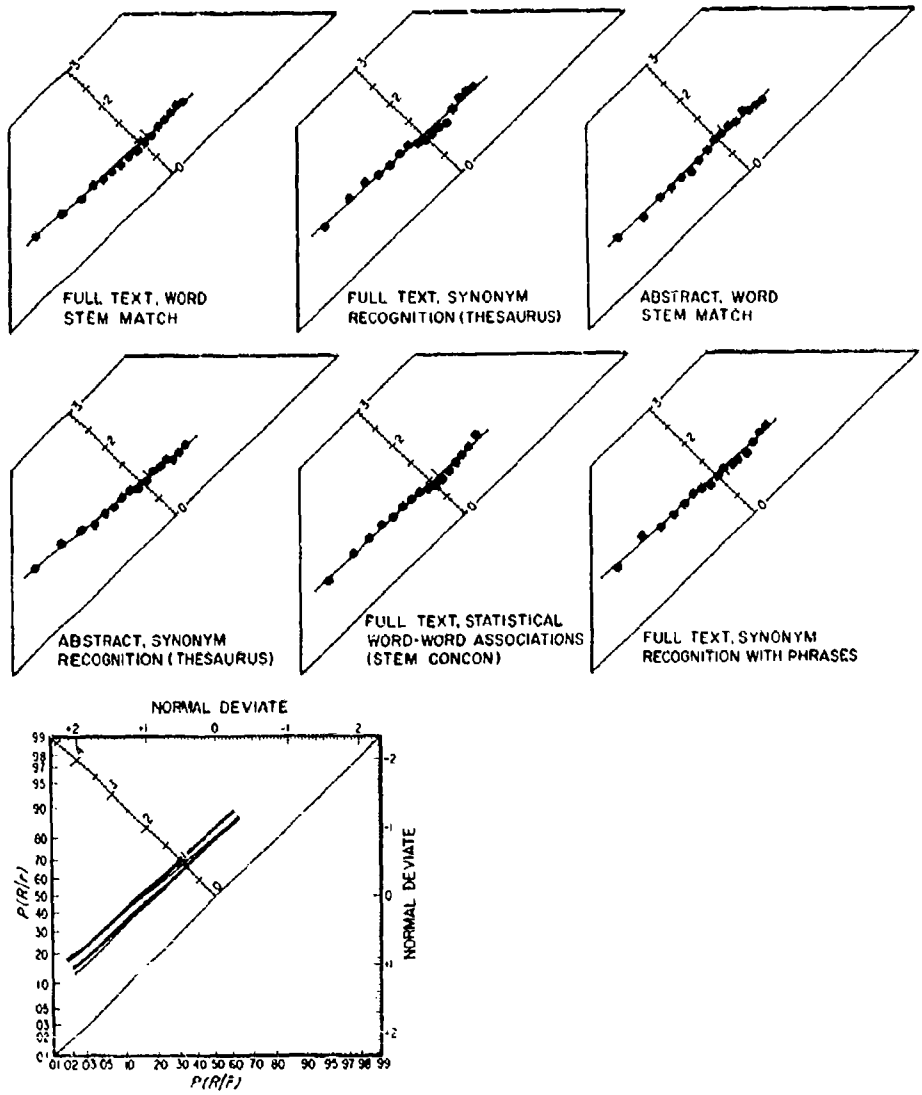


Fig. 7. ADI collection: 6 methods. 82 items, 35 queries. 170 relevant + 2,700 irrelevant = 2,870 total. Criteria: 3, 6, 9, ..., 47 retrievals. Harvard-Cornell: Salton and Lesk.

The data are quite adequately fitted by straight lines in every instance. Indeed, according to standards acquired through experience in other fields (for example, human signal detection and recognition memory) where the decision-theory measure has proved to be useful (4), the fits are fantastically good.

A small staircase effect can be discerned in the data. This effect may be the result of having a relatively small sample (containing an average of 5 relevant items for 35 questions); the procedure used in analysis for defining acceptance criteria forces each successive point a certain distance to the right, and a low density of relevant items would produce irregular upward movement. In any case, the effect is not large enough to be of much concern. We can see also some variation in the slopes of the lines; we shall consider the significance of this variation after all the data have been examined.

Figure 8 shows the results of seven retrieval methods applied to a collection of items on aerodynamics borrowed by the Harvard-Cornell group from the Cranfield project. Again, the straight-line fits exceed reasonable aspirations, and a variation in slopes appears.

Figure 9 represents one of two collections in the subject area of computer science, called IRE 1, and six retrieval methods. Figure 10 shows the second IRE collection and ten methods. Figure 11 shows the second IRE collection with a different set of ten methods.

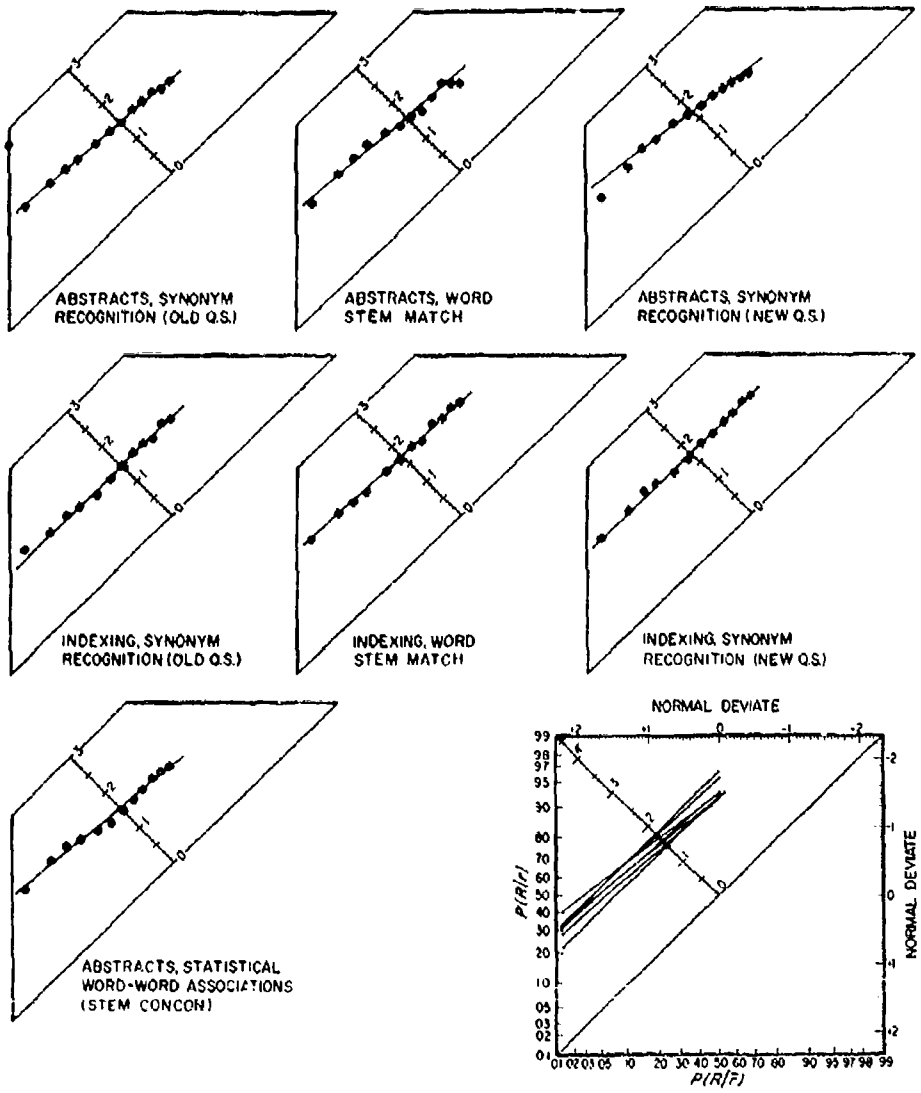


Fig. 8. Cranfield collection: 7 methods. 200 items, 42 queries. 198 relevant + 8,202 irrelevant = 8,400 total. Criteria: 5, 10, 15, 20, 30, 40, ..., 100 retrievals. Harvard-Cornell: Salton and Lesk.

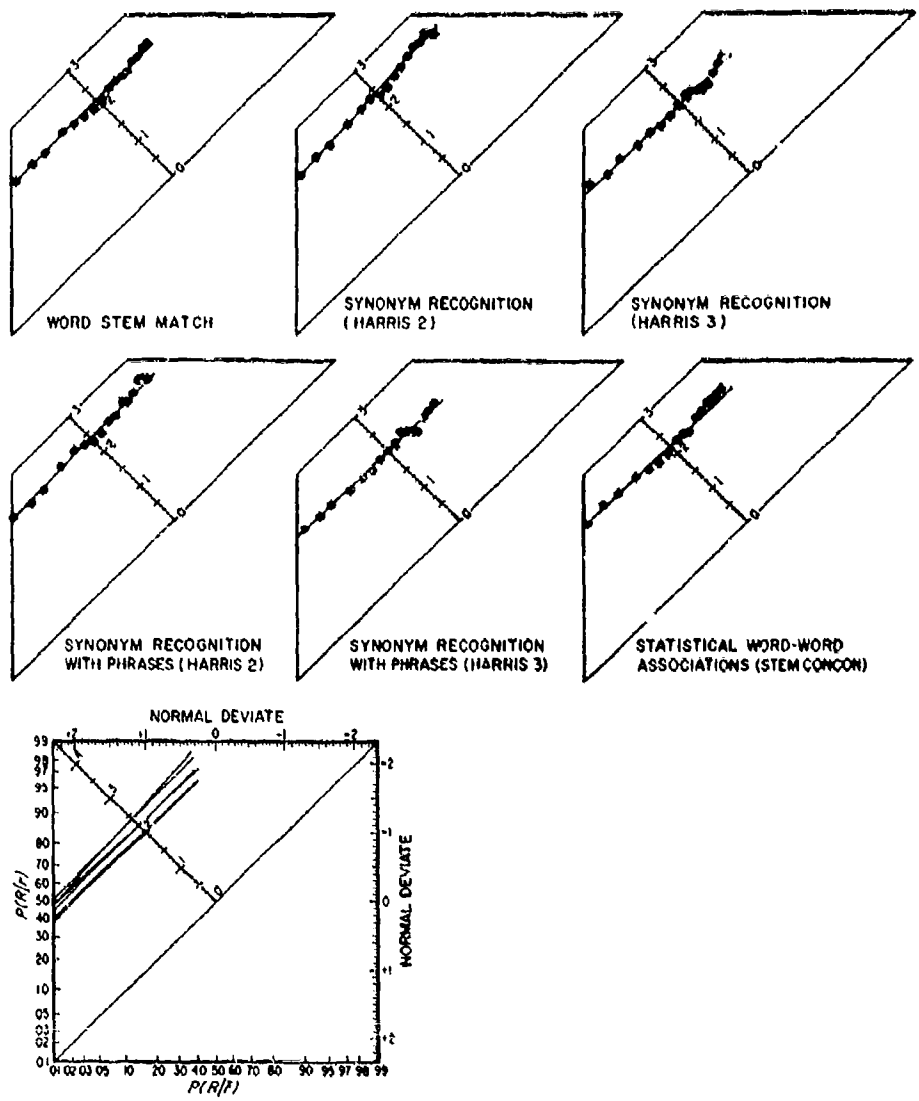


Fig. 9. IRE 1 collection: 6 methods. 405 items, 17 queries. 186 relevant + 6,699 irrelevant = 6,885 total. Criteria: 10, 15, 20, 30, 40, ..., 150 retrievals. Harvard-Cornell: Salton and Lesk.

With the IRE collection we notice a tendency, at higher values of E, for the slopes to be greater than unity. The slopes in Fig. 9 range from 0.95 to 1.12, in Fig. 10 from 0.98 to 1.40, and in Fig. 11 from 1.20 to 1.56. With the ADI collection (Fig. 7) the slopes range from 0.83 to 0.99, and with the Cranfield collection (Fig. 8), from 0.76 to 1.00.

We can't help but observe the substantive result of this analysis that the differences in effectiveness among the various methods are small relative to the differences among collections. The range in E for the six methods applied to the ADI collection is 0.20 (from 0.90 to 1.10); for the seven methods used with the Cranfield collection, 0.35 (from 1.45 to 1.80); for the six methods used with the IRE 1 collection, 0.40 (from 2.00 to 2.40); for the first ten methods used with the IRE 2 collection, 0.55 (from 1.95 to 2.50); and for the second group of ten methods used with the IRE 2 collection, 0.30 (from 2.10 to 2.40). These ranges, on the order of 0.50 or less, can be compared with the range over all collections of 1.60, keeping in mind the scale range of about 5.00 from chance performance to very good performance. The Harvard-Cornell and Cranfield investigators are inclined to believe that the dependency of effectiveness on the collection results both from differences in the "hardness" of the vocabularies of the three subject fields, and from the use of different procedures with the three collections for establishing relevance (7).

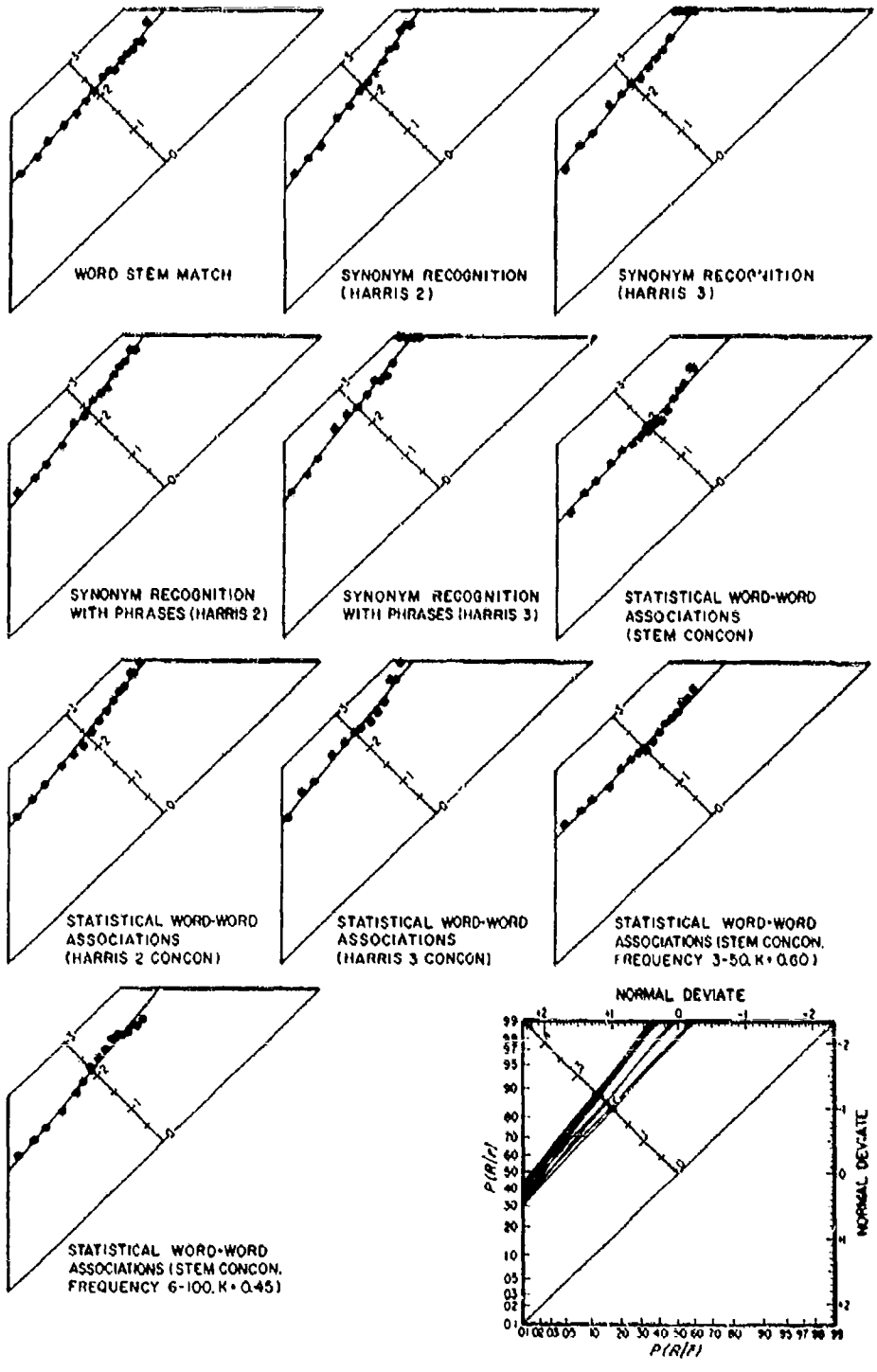


Fig. 10. IRE 2 collection: first set of 10 methods. 380 items, 17 queries. 181 relevant + 6,279 irrelevant = 6,460 total. Criteria: 10, 15, 20, 30, 40, ..., 150 retrievals. Harvard-Cornell: Salton and Lesk.

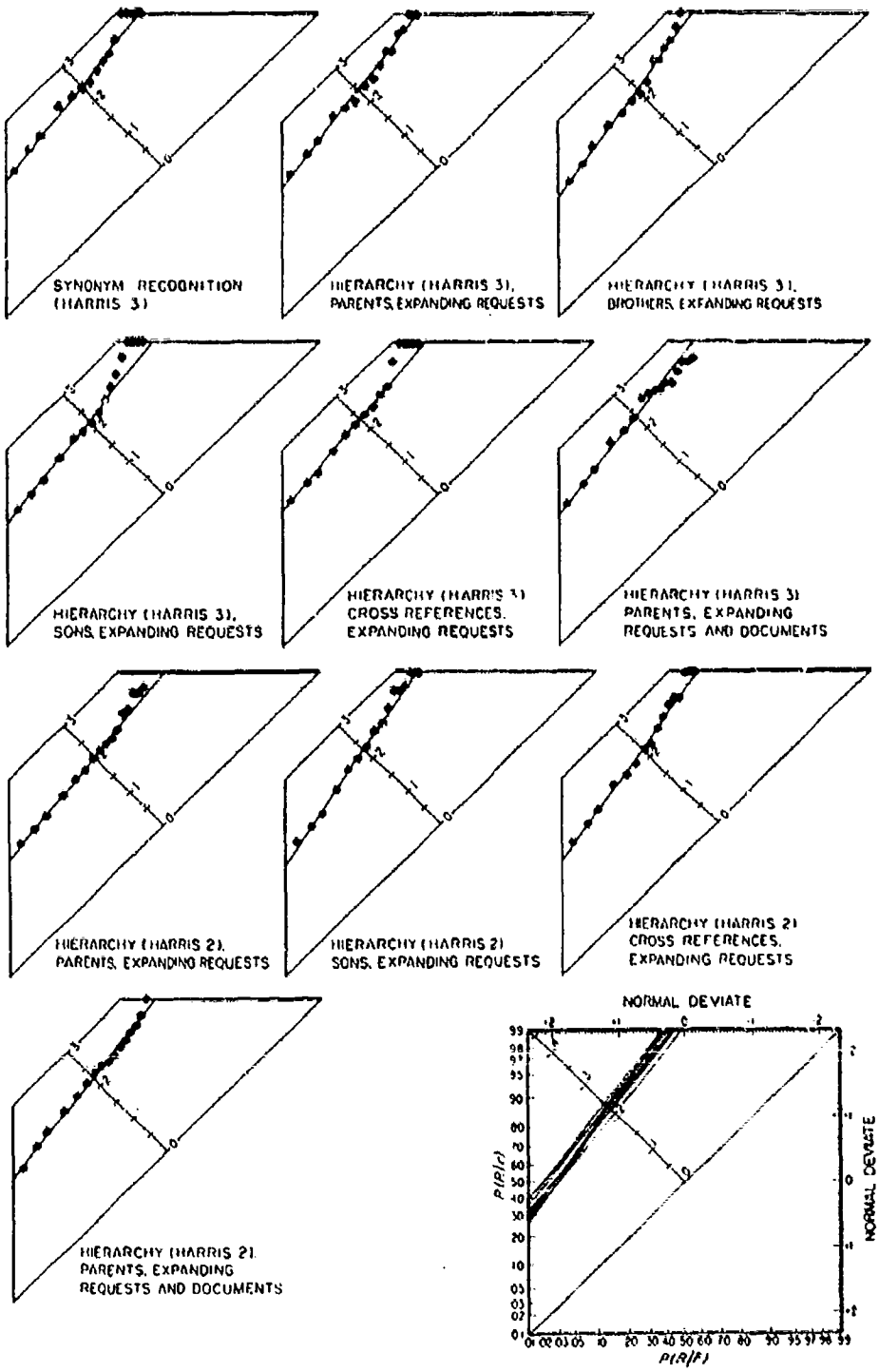


Fig. 11. IRE 2 collection, second set of 10 methods. 380 items, 17 queries. 178 relevant + 6,282 irrelevant = 6,460 total. Criteria: 10, 15, 20, 30, 40, ..., 150 retrievals. Harvard-Cornell: Salton and Lesk.

Cranfield data. The study at Cranfield has been actively pursued for several years, and the last report contains an enormous amount of data. I have plotted only a fraction of the results; however, I am not aware of any particular bias in my casual sampling, and all the plots prepared are included here.

The Cranfield data are distinguished from the Harvard data in being based on a larger file (in most cases 1,400 items, as compared with the largest Harvard collection of about 400 items), and on more questions (approximately 220, as compared with the Harvard maximum of about 40). One consequence is the appearance of lower false-drop proportions, proportions that fall off the graph paper (Codex Graph Sheet No. 41,453) used in the preceding figures. So we use another graph paper (Keuffel and Esser Co. No. 47 8062) that ranges down to a proportion of 0.0001. Though the graphs following have on them scales of the normal deviate, these scales, unfortunately, are not given on the Keuffel and Esser paper available commercially.

In the Cranfield system, a manual one, the relevance of every item to every query is determined by judges, but the system itself does not rank items according to their degree of relevance to the query. Various acceptance criteria are obtained by establishing different "levels of coordination," that is, by varying the requirements on the number of query terms an item must satisfy in order to be retrieved.

Figure 12 shows the results of five retrieval methods that vary in the "recall device" they employ. The slopes are quite uniform, slightly greater than unity, and not many of the points

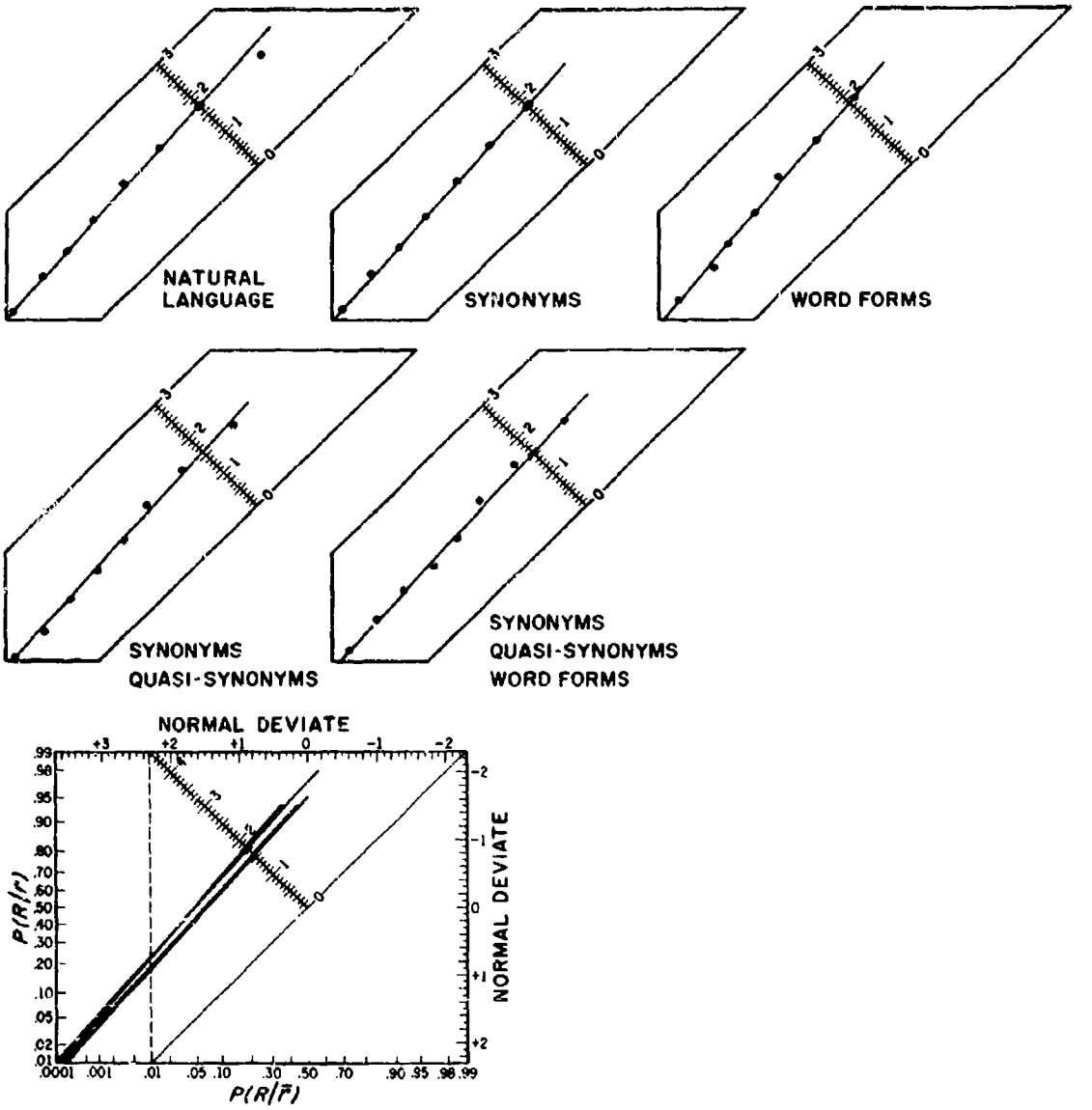


Fig. 12. Recall devices, single-term index language: 5 methods. 1,400 items, 221 queries. 1,590 relevant + 307,810 irrelevant = 309,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

fall off the fitted lines. Essentially the same comments apply to Fig. 13, which shows two levels of indexing exhaustivity for two sets of recall devices. Likewise for Fig. 14, which illustrates the effects of requiring different degrees of relevance for retrieval to be effected. The left panel results when all four categories of judged relevance satisfy the retrieval criterion; moving to the right, the relevance requirement is strengthened, so that in the last panel we have the results when only those items with the highest degree of relevance are retrieved. Figure 15 shows some results obtained with a smaller collection when retrieval is based only on titles and abstracts, or only on titles, and the fits are about as good as before.

In Fig. 15 values of E range from 1.33 to 1.70, and values of the slope range from 0.80 to 0.95. In the three figures preceding, E ranges from 1.58 to 1.86, and s lies between 1.08 and 1.18.

Arthur D. Little, Inc., data. Like the Harvard system, the system constructed at Arthur D. Little, Inc., (ADL) assigns an index value to each item according to its relevance for each query. Again, however, the system did not produce a print-out of data in full enough form to enable us to look directly at the density functions supposed to underlie the OC curves.

The ADL system was used with a still larger store, effectively 4,000 items. I have based arbitrary acceptance criteria, again, on the number of items considered as retrieved. The terminal criterion, in this case, was determined by the ADL investigators; they proceeded through the items according to

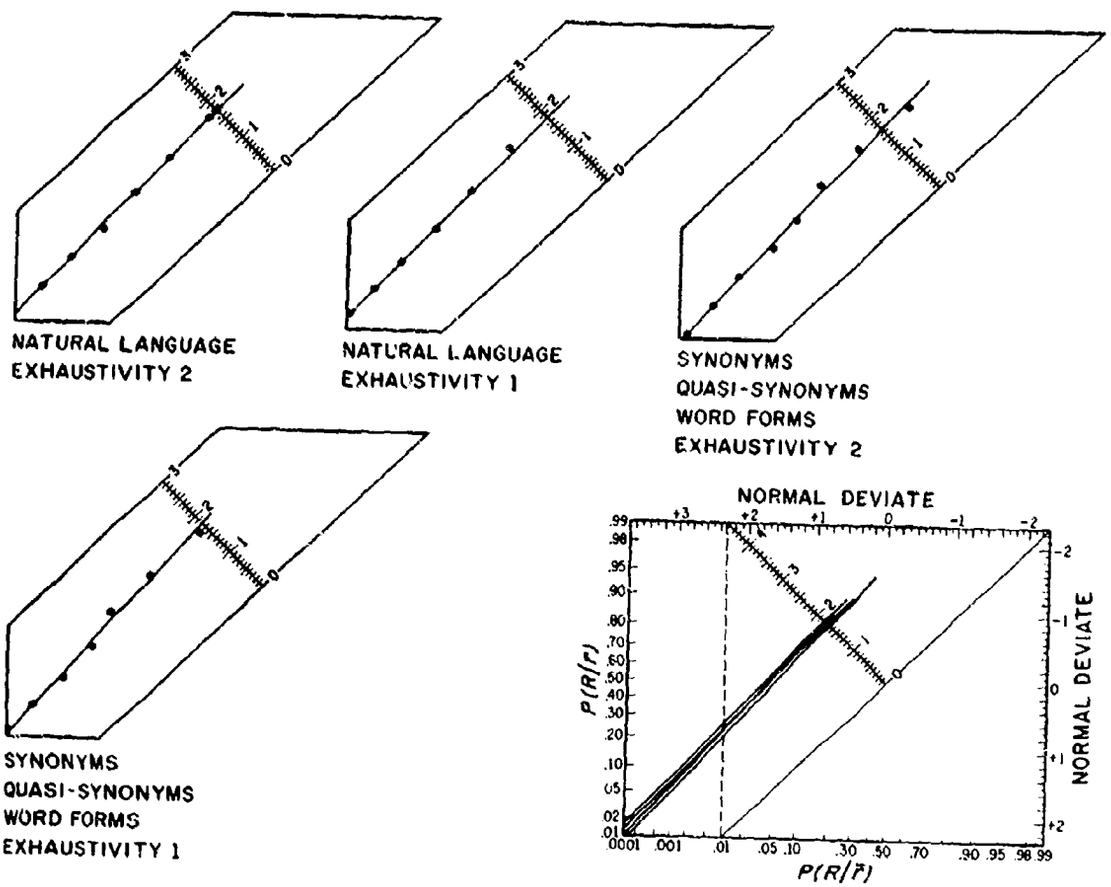


Fig. 13. Indexing exhaustivity, single-term index language: 4 methods. 1,400 items, 221 queries. 1,590 relevant + 307,810 irrelevant = 309,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

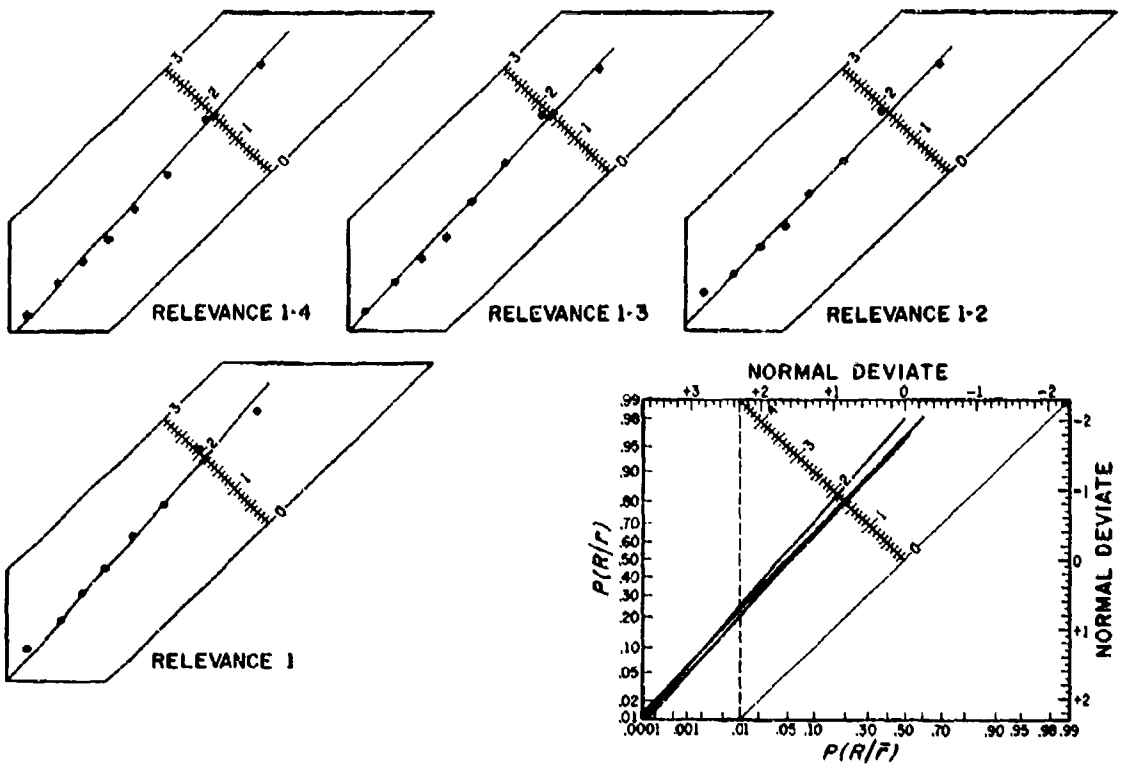


Fig. 14. Document relevance, single-term index language, natural language: 4 methods. 1,400 items, 50 queries.
Relevance 1-4: 361 relevant + 69,639 irrelevant = 70,000 total.
Relevance 1-3: 297 relevant + 69,703 irrelevant = 70,000 total.
Relevance 1-2: 155 relevant + 69,845 irrelevant = 70,000 total.
Relevance 1: 95 relevant + 69,905 irrelevant = 70,000 total.
Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

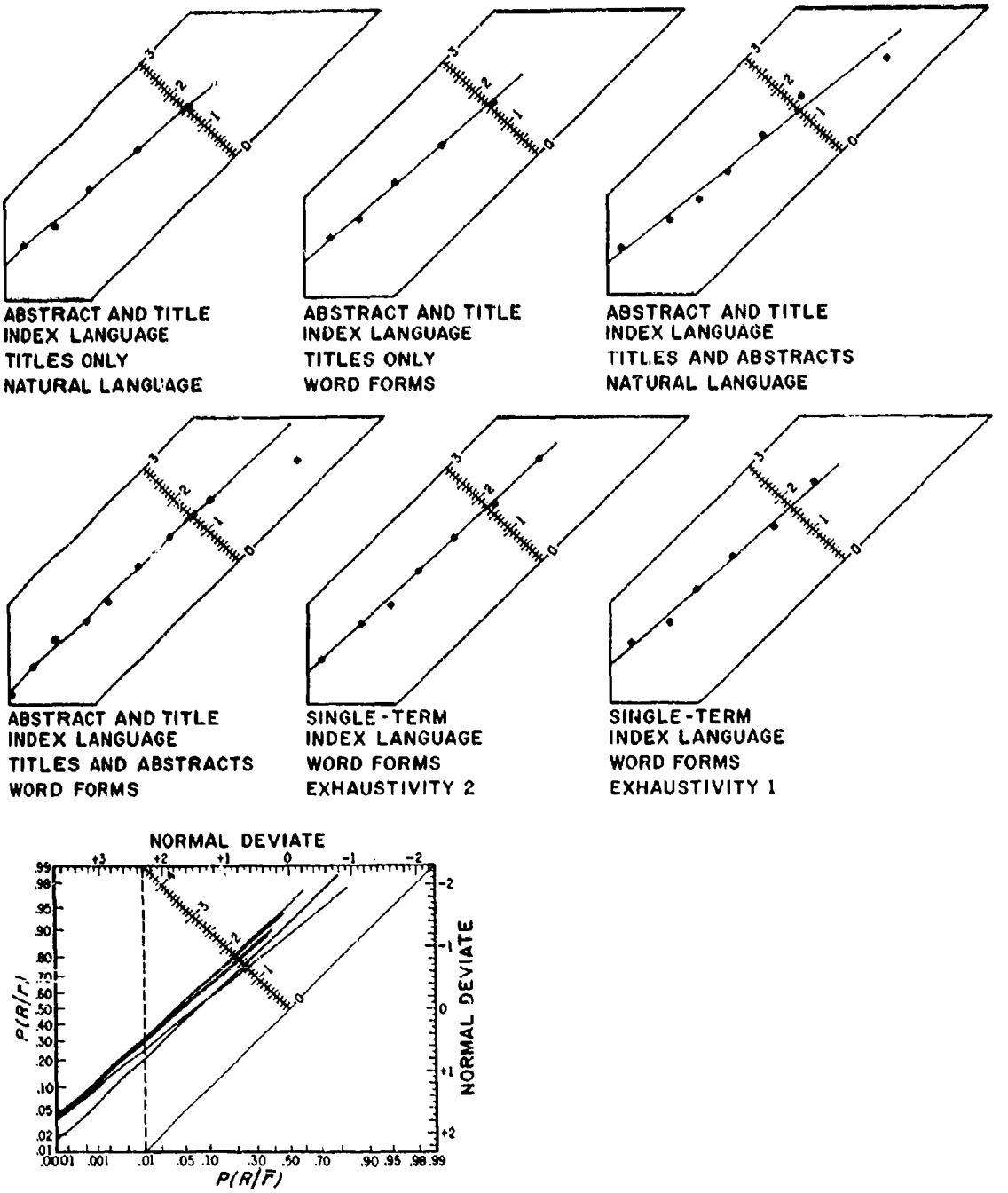


Fig. 15. Abstracts and titles: 6 methods. 200 items, 42 queries. 198 relevant + 8,202 irrelevant = 8,400 total. Criteria: levels of coordination. Cranfield: Cleverdon and Keen.

their rank to judge the relevance of each, and stopped when it seemed that relevant items were turning up on a random basis. In order to determine the recall ratio, or hit proportion, of course, the total number of relevant items for each query had to be established. These numbers were estimated at ADL from a sample of 400 items drawn from the store of 4,000 items.

Included in the following figures are almost all the data, and all the major data, collected at Arthur D. Little, Inc. A difference between these and foregoing plots is that most of these are based on single queries. The data points, surprisingly, do not show much greater scatter about a line, but substantially greater variation in the slopes is evident.

Figure 16 shows the associative retrieval method applied to four queries which consisted of abstracts ("full text queries"). Also shown is the same method applied to briefer forms of the same queries. In the latter case ("CBU queries") the queries consisted of critical word strings selected from the abstracts, designated as "content-bearing units." The full OC plots show the pooled results for queries 1, 3, and 4 for each type of query. Query 2 was excluded from the pooled results because the range of acceptance criteria available for it was relatively limited, and various means of pooling queries with different ranges of acceptance criteria proved unsatisfactory. If the curves of the last two plots are extrapolated to the negative diagonal, values of E are obtained (approximately 1.30 and 2.20) that lie in the range of empirical values noted earlier. The slopes of the lines (approximately 1.00 and 1.50) are also in the range of empirical values noted earlier.

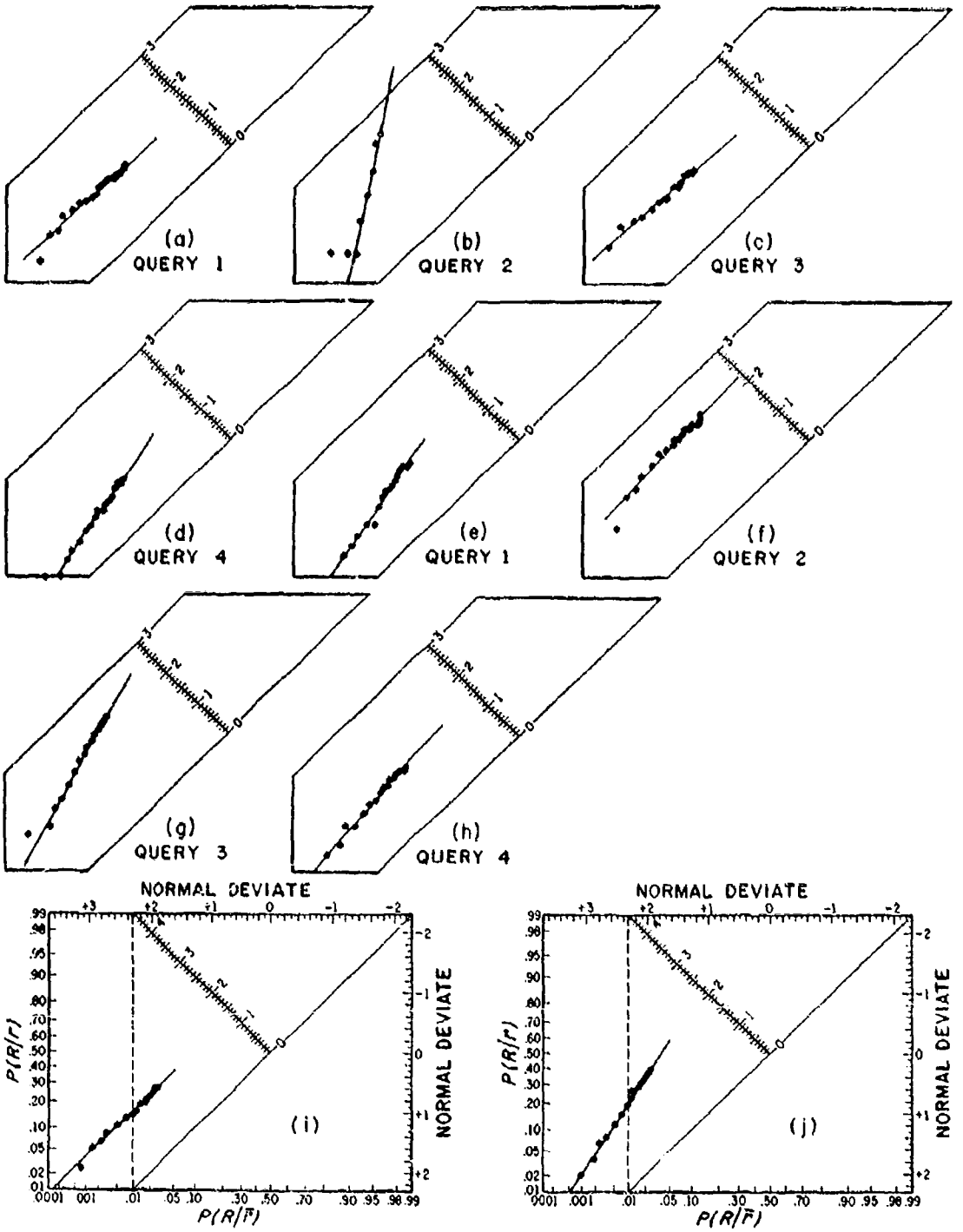


Fig. 16. (a)-(d): Fully automatic associative, 4 full-text queries. (e)-(h): Fully automatic associative, 4 CBU queries. 4,000 items. Number relevant: Query 1, 80; Query 2, 30; Query 3, 70; Query 4, 100. (i), (j): Average of Queries 1, 3, and 4. (i): full-text queries; (j): CBU queries. Criteria: 5, 10, 15, 20, 30, 40, ..., retrievals. Arthur D. Little, Inc: Giuliano and Jones.

Figure 17 shows three different retrieval methods used with the short queries. Figure 18 shows another method and reproduces the fitted lines for the four methods of Figs. 17 and 18 for each query. There is a tendency for the slopes to depend more upon the query than the method. Averaging over methods, the slopes range from about 1.00 for query 4, through approximately 1.45 for queries 2 and 3, to about 1.75 for query 1. Average slopes for the four methods lie between 1.28 and 1.52. The average values of $\underline{E}$ associated with the four methods, by extrapolation, range from 1.60 to 2.10.

We may note that the highest value of $\underline{E}$, 2.10, is obtained with the method called "selected associations," shown in panels (i) through (l) of Fig. 17. It can be seen that, in fitting straight lines to the data obtained with that method, data points falling below the line at the lower false-drop probabilities were virtually ignored in the case of two queries (queries 1 and 4). Clearly, if we were to restrict our interest to low false-drop probabilities -- say, if we were to consider only the left-most half-dozen or so points -- then the slopes for that method would be steeper, and the values of $\underline{E}$ estimated would be higher. In fact, if the four queries are pooled with only the left-most nine points included, the resulting value of $\underline{E}$ is close to 3.0 (and the resulting slope is about 1.8). The "selected-associations" method is one of two methods tried at ADL with user intervention between iterative searches. The other method in which adjustments were made between iterations is the one called "reweighted associative," shown in Fig. 18; in that case all the data points are quite well taken into account in fitting lines, and an $\underline{E}$ = 1.90 is obtained.

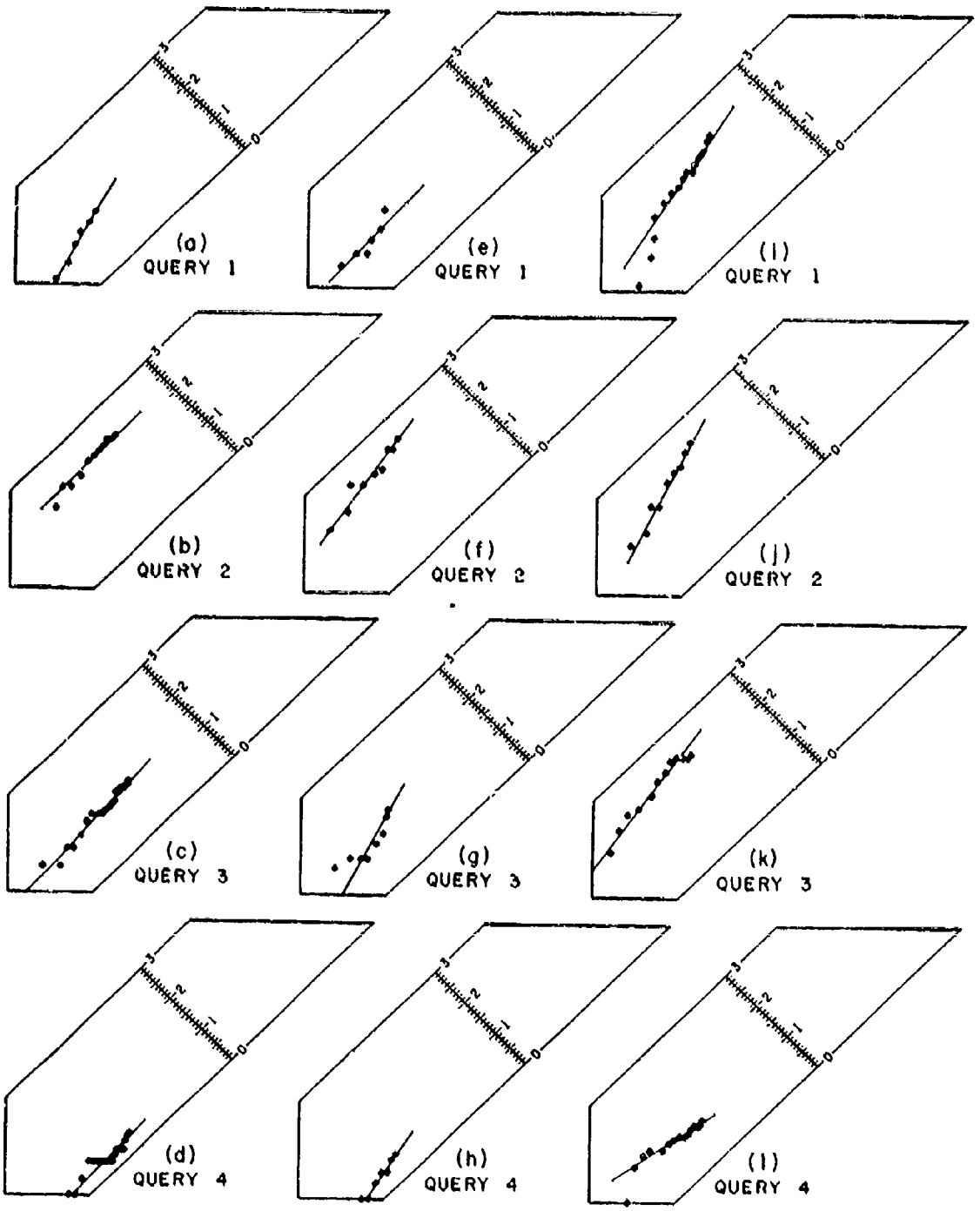


Fig. 17. CBU queries: 3 methods. (a)-(d): Modified coordinate. (e)-(h): Frequency-weighted coordinate. (i)-(l): Selected associations. Number of items, number relevant per query, and criteria as in Fig. 16. Arthur D. Little, Inc: Giuliano and Jones.

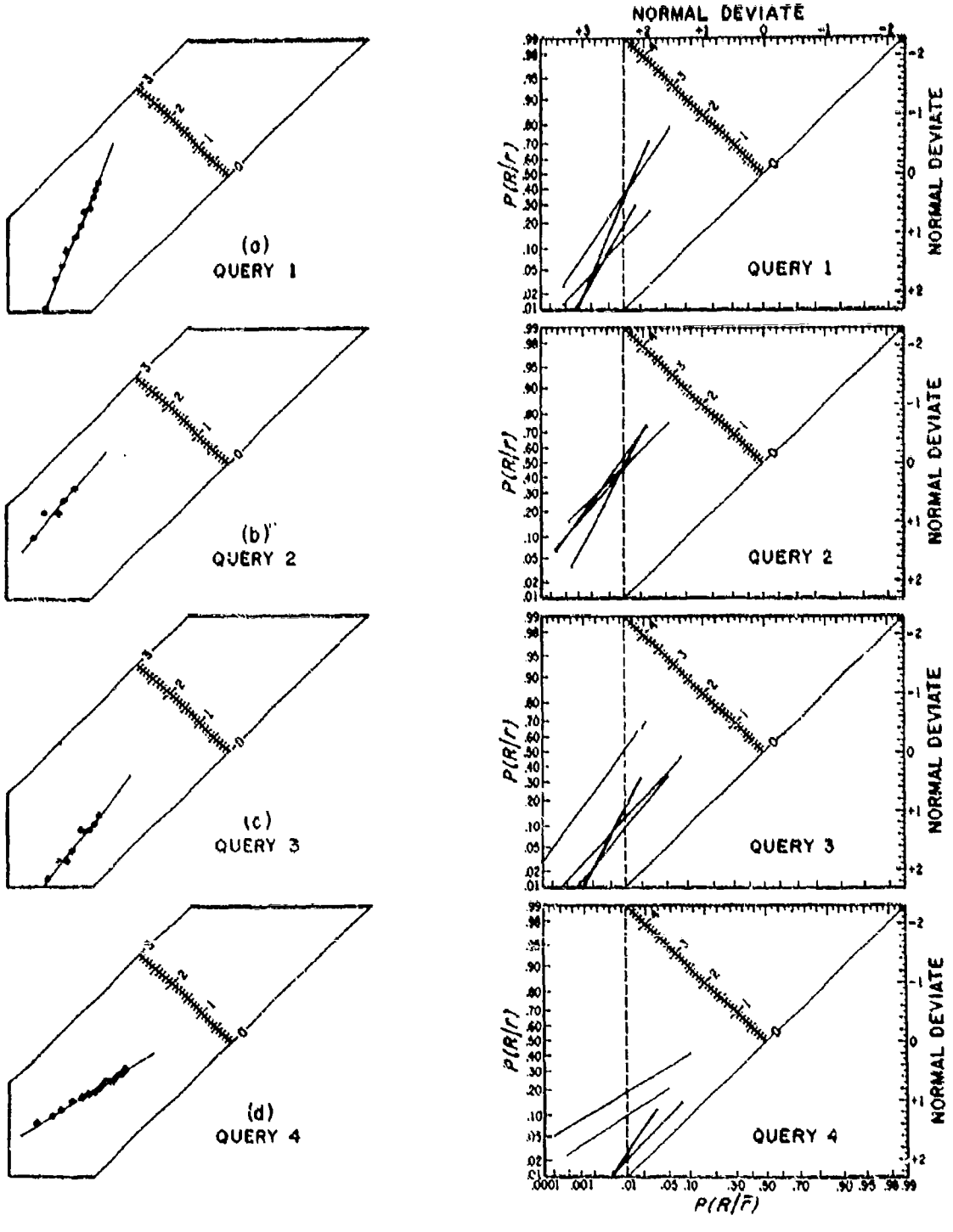


Fig. 18. CBU queries: a 4th method, and summaries of it and the 3 methods of Fig. 17. (a)-(d): Reweighted associative. (e)-(h): results of 4 methods for each query. Number of items, number relevant per query, and criteria as in Fig. 16. Arthur D. Little, Inc: Giuliano and Jones.

Conclusions

The consistent linearity of the empirical operating-characteristic curves confirms that a decision-theory measure can be used to reflect solely the effectiveness of a retrieval system, and effectiveness unconfounded by variation in the acceptance criterion. The apparently irregular variation in the slopes of the curves presents a slight complication relative to achieving a measure that is a single number, but not enough of a complication to impair seriously the usefulness of a decision-theory measure.

Two numbers -- E measured at the negative diagonal of the OC space, and the slope, s -- give an accurate description of the curve representing constant retrieval effectiveness over varying acceptance criteria. Two numbers are not as convenient as one, but these particular two give a considerably more economical description of the performance curve than available previously, and can be reported in cases where conveying information about the full curve is desirable.

The data at hand indicate, however, that for most purposes conclusions about effectiveness can be drawn from the value of E alone, without regard to s. In short, there is little point in concern over small differences in s when differences in E are small. We have seen that when values of s are based on more than a few queries they do not vary enough to obscure a substantial difference in E.

What constitutes a "substantial difference" in E, or a difference of practical significance? An approximate answer derived from the present data is that a difference in E in the neighborhood of 0.30 to 0.50 is a reasonably significant one. Thus, for example, in the Harvard data based on the IRE collections (Figs. 9, 10, 11), a difference between two methods of that magnitude corresponds to a factor of about two in the false-drop probability. (By way of illustration, it can be seen in Fig. 9 that at a hit probability of 0.90 the extreme methods show false-drop probabilities of approximately 0.25 and 0.13; at a hit probability of 0.70 the extreme false-drop probabilities are about 0.13 and 0.07; at a hit probability of 0.50 the extreme false-drop probabilities are about 0.02 and 0.01.) It seems unlikely that a smaller experimental difference would have much practical import.

As discussed earlier, if it should seem worthwhile to have a measure that is both a single number and sensitive to variation in slope, the distribution-free measure A could be used. Let us use the measure A now to get a different view of the observed differences among methods in the present sample, a view that will help us judge how small a difference in E is practically significant. A, it will be recalled, is the proportion of the area of the OC space that lies beneath an OC curve plotted on linear scales (as in Fig. 4), and is equal to the probability of choosing between two items, one drawn at random from the relevant set and the other drawn at random from the irrelevant set, the item that is relevant. Assume for the purpose at hand that all of the OC curves in our sample are of unit slope; this approximation introduces a distortion that is

negligible relative to the point of interest here, and permits a conversion from E to A by means of published tables (8). For the Harvard data, values of A, or values of the probability of a correct choice in a two-alternative forced-choice test, denoted $P_2(C)$, range from 0.74 to 0.78 for the ADI collection (Fig. 7), from 0.85 to 0.90 for the Cranfield collection (Fig. 8), and from 0.92 to 0.96 for the IRE collections (Figs. 9, 10, 11). For the Cranfield data, $P_2(C)$ ranges from 0.87 to 0.91 for the large collection (Figs. 12, 13, 14) and from 0.83 to 0.89 for the small collection (Fig 15). For the data collected at Arthur D. Little, Inc., the range of the four "CBU" methods (Figs. 17, 18), averaged over the four queries, is from 0.87 to 0.93. It might be argued, again, that the differences between extreme methods for any collection, of 0.04 to 0.06, are real differences, but it seems unlikely that differences of less than 0.04 in $P_2(C)$ have material implications.

These values of $P_2(C)$, lying between 0.74 and 0.96, indicate that present retrieval methods leave considerable room for improvement. (Said otherwise, these values of $P_2(C)$, considered along with the competence and diligence with which the experiments here represented were pursued, indicate that information retrieval is a very difficult problem.) On the face of it, choosing the single relevant item from a collection of two items is not a demanding task, and we should hope that our retrieval systems would make the correct choice almost every time, say, with a probability of 0.99 or greater. A more compelling impression, however, of the current state of the retrieval art is gained by taking pairs of hit and false-drop probabilities from the empirical OC curves and converting these probabilities to raw numbers.

Consider an OC curve with $E = 2.5$ and $s = 1.3$. This curve is close to the best of the curves seen in the foregoing, and exceeded by none of them. It passes through the points $P(R|\bar{r})$ and $P(R|r)$ having coordinate values of (0.001, 0.12), (0.01, 0.42), and (0.10, 0.88). Assume a file of 3,000 items and a group of queries to each of which 10 of the 3,000 items are relevant. Now, if we will settle for retrieving, on the average, only 1 of the 10 relevant items per query, we will also receive 3 false drops each time. If we desire 4 of 10 relevant items, we will have to winnow the 4 from 30 irrelevant items. If we should aspire to 9 of 10 relevant items, we would have examine more than 300 items, in response to each query, to find the 9.

These noise-to-signal ratios are dramatically large. The ratio amounts rapidly even for a file as small as 3,000 items: from 3 to 7 to 33 for the three acceptance criteria of the example. For a file of 10,000 items the corresponding noise-to-signal ratios are 10, 25, and 100 plus. It is with these ratios in mind that I earlier suggested dismissing small differences in E and ignoring small variations in s.

The decision-theory analysis can be seen to set the stage clearly for identifying an important advance in retrieval technique. The best of the performances sampled here, in the vicinity of $E = 2.5$ and $s = 1.3$, gives a false-drop probability of approximately 0.10 for a hit probability of 0.90. Assuming the same slope, and taking the same hit probability, an $E = 3.0$ corresponds to a false-drop probability of 0.05, and an $E = 3.6$ corresponds to a false-drop probability of 0.01. An $E = 4.0$

means a false-drop probability of 0.005, or reception of 15 unwanted items along with 9 of the 10 wanted items from a file of 3,000. An $\underline{E}$ = 4.5 means a false-drop probability of 0.001, or reception of 3 unwanted items along with 9 of the 10 wanted items from a file of 3,000.

A belief of several people working in the retrieval field is that a very significant advance in retrieval effectiveness will be achieved in the near future by "on-line" systems, in which the user is given immediate feedback and enabled to progressively refine the search prescription over successive trial searches. It will be informative to apply the decision-theory analysis in experiments on on-line procedures. Will we see values of $\underline{E}$ in the vicinity of 3.0, or 3.5? Might we even find values of $\underline{E}$ about 4.0 -- or will present knowledge of language forms impose a barrier at a lower level of effectiveness?

References and Notes

1. J. A. Swets, Science, 141, 245 (1963).

2. See C. P. Bourne, Chapter 7 of Annual Review of Information Science and Technology (Interscience, New York, 1966), pp. 176-179.

3. C. Cleverdon and M. Keen, Association of Special Libraries and Information Bureaux, Cranfield, England, Volume 2 (1966).

4. D. M. Green and J. A. Swets, Signal Detection Theory and Psychophysics (Wiley, New York, 1966), pp. 45-51.

5. G. Salton, M. Lesk, et. al., Department of Computer Science, Cornell University, Sci. Rept. No. ISR-11 (1966).

6. V. E. Giuliano and P. E. Jones, Arthur D. Little, Inc., Cambridge, Mass., Interim Rept. (1966).

7. Personal communications (1967).

8. J. A. Swets, Ed., Signal Detection and Recognition by Human Observers (Wiley, New York, 1964), pp. 682, 683.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

EFFECTIVENESS OF INFORMATION RETRIEVAL METHODS

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

John A. Swets

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
15 June 1967	47	8

8a. CONTRACT OR GRANT NO

Contract No. AF19(628)-5065

b. PROJECT NO

8668

c. DoD Element 6154501R

d. DoD Subelement n/a

9a. ORIGINATOR'S REPORT NUMBER(S)

BBN Report No. 1499
Scientific Report No. 8

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

AFCRL-67-0412

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

11. SUPPLEMENTARY NOTES

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627, Amendment 2

12. SPONSORING MILITARY ACTIVITY

Air Force Cambridge Research Laboratories (CRB)
L.G. Hanscom Field
Bedford, Massachusetts 01730

13. ABSTRACT

Results of some fifty different retrieval methods applied in three experimental retrieval systems were subjected to the analysis suggested by statistical decision theory. The analysis validates a previously-proposed measure of effectiveness and demonstrates its several desirable properties. The examination of a wide range of data in relation to this one metric provides a clear and general assessment of the current state of the retrieval art, and shows that the art is still far from what might be considered a desirable state.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Retrieval						
Information Retrieval Tests						
Retrieval Techniques						
System Evaluation						
Measures of Effectiveness						
Performance Criteria						
Applied Decision Theory						
Question-Answering Systems						
On-Line Computer Procedures						