12

Retrieval system tests 1958–1978

Karen Sparck Jones

Previous chapters considered the problems of information retrieval experiment in general, and particular types of experiment. This chapter looks at information retrieval experiment over the last twenty years as a whole, to see what the actual tests which have been carried out show. I shall not attempt an exhaustive review of this work. I shall seek rather to characterize it by referring on the one hand to especially significant tests, and on the other to average or representative ones. My object is to exhibit the development of retrieval experiment in the last two decades in terms of the purpose, quality and influence of the tests which have been carried out. The twenty year period for the survey is a natural one, since it effectively covers the development of modern, especially computer-based, retrieval systems, and equally, most of the significant information retrieval tests.

12.1 Experiment and investigation

The amount of work done under the general heading of information system studies is very large. To see the wood in the trees it is essential to have a clear view of what constitutes an experiment and to restrict the survey, as far as possible, to experiments in the strict sense. Thus for the purposes of this chapter an experiment is distinguished from an investigation in the following ways. An experiment aims at explanation, an investigation only at description: an experiment seeks to answer questions about what happens if such and such is done, by showing why it happens; an investigation indicates only what happens. In the context of information retrieval system testing, an experiment typically focuses on individual variables, where an investigation exhibits system behaviour as a whole. An experiment is in principle hypothesis-guided, while an investigation may be no more than hypothesis-generating. The key requirement of experiment is therefore control over test variables, both primary and secondary. In consequence, experiment is concerned with measurement. Investigation may also produce measurments, and in both experiment and investigation measurement may be merely descriptive. However since an information retrieval system has a function, any measurements must ultimately be related to system performance in terms

of effectiveness, somehow defined. The difference between experiment and investigation is therefore that in experiment explicit comparative measurements are required for different values of the test variables; in investigation, comparison may be only implicit. Further, since retrieval systems have a function, evaluation experiments relating specifically to performance effectiveness, i.e. the ability of the system to retrieve relevant documents and to suppress non-relevant ones, have a special status as the most important kind of experiment.

Unfortunately, the distinction between experiment and investigation just summarized is an ideal which is very difficult to maintain when discussing actual system tests. Much of the work done cannot be described as unequivocally experimental or investigative, expecially where studies of operational systems are concerned. The problem is really that information retrieval systems are so complicated, and so little understood, and there is such a lack of solid theory about them, that really high class experiment can hardly be expected. In a way a review of information retrieval experiment is a review of the inadequacy of information retrieval experiment. The work discussed in this chapter thus ranges from experiments proper to better conducted and relatively systematic investigations. A particular problem is that while both experiment and investigation can in principle refer to operational system studies, in practice there have been few thoroughly controlled operational system tests, and experiment and investigation typically imply laboratory and operational environments respectively. There are indeed, as is noted in other chapters, considerable difficulties about conducting rigorous operational system experiments.

Within the area of information retrieval experiment we can then sort the tests done according to the degree of control they involved, and according to the type of hypothesis they invoked. Control is exhibited by comparison, and the degree of control corresponds largely to the scope or level of the factor being studied as the primary experimental variable, i.e. variable on which the experimenter's interest is focused. Thus at the highest level we may compare whole indexing and searching subsystems within the fixed environment represented by a certain body of users and of literature; at the medium level we may compare different indexing thesauri; and at the lowest level we may vary indexing exhaustivity using a given thesaurus. As long as the environment parameters are held constant, all of these are comparisons implying some degree of control, but in the case of whole indexing or searching subsystems control will be minimal. The consequence is that any observed differences (or similarities) in system performance will not be explicable in any detail since the indexing and searching subsystem as a whole subsumes many lower-level variables. The problem most severely felt by research workers has been that of identifying useful, meaningful unit variables in retrieval systems, i.e. those variables capable of affecting performance for determinable reasons. A closely related problem is that of managing secondary, related variables, since their identification and manipulation are associated with the treatment of the primary variables. The treatment of indexing exhaustivity and specificity in relation to an index language are good examples of this problem.

Similar points can be made about the hypotheses underlying information retrieval experiment. Some hypotheses are rather general, for instance that

in retrieval searching is more important than indexing, and so are difficult to test. In some cases, hypothesis may be of the weakest kind illustrated by any statement that some variable must be important so its behaviour is worth study: the requirement is then to find out why it is important. Statements to the effect that a certain value of some variable is superior to another value, or that one value is precisely twice as good as another, are then progressively stronger hypotheses. Again, a major problem for information retrieval research in the past two decades has been that of formulating testable explanatory hypotheses about information system behaviour, and especially hypotheses given a definite interpretation by a formal model. The distinction between the explanatory hypotheses of experiment and the descriptive hypotheses of investigation is not always easy to maintain. Explanatory hypotheses about the behaviour of a retrieval system ultimately refer to the way it functions in relation to its purpose, i.e. to its performance. Descriptive hypotheses are often assumed to have some connection with system function, but the nature of the connection may be far from clear. Descriptive hypotheses may indeed be tested, but in such cases they are either implying explanatory hypotheses or referring to certain system elements simply as data. Bibliometric and also user studies are examples of descriptive hypothesis tests. Thus bibliometric studies may be concerned to test hypotheses about the distribution of citations in journals, or of citation links between papers, with the test variable the subject area of a literature, for instance. However in interpreting such tests we have either a presumption that describing the structure of a literature has some bearing on retrieval system behaviour, or are in fact concerned with another type of information system as a phenomenon for study. Information retrieval research over the past twenty years could perhaps be described as a long and not altogether successful attempt to convert descriptive hypotheses into explanatory ones.

12.2 Approaches to the historical review

There are thus various ways in which the experimental work of the past two decades can be treated. One possibility is a straightforward historical account; another is a review focusing on the development of methods of experiment (or lack of it); and yet another is a characterization of the research in terms of the attempt to generate theories and models motivating experiments. The last two taken together would indicate the quality of experimental work in information retrieval. There are, however, further possibilities. One is to survey the experiments done by topic, i.e. to consider what particular questions within the whole range of questions that could be asked about document retrieval systems have attracted most attention, or produced the most significant results. The other possibility is to consider the experiments done in terms of their influence, actual or potential, on operational systems. There are in fact no very clear patterns to be seen, since experiments important on one count may not be so on others: for example we can have a methodologically sound experiment concerned with an unimportant question, or a good experiment without influence. There are, however, some major studies of importance for more than one reason, like Cranfield 1 and 2[1–3], or Salton's Medlars test[4]; and though the overall pattern is not very clear, the general colour of the cloth is plain, and there are some differently

coloured threads to be traced, some of them even standing out as brightly coloured against the overall grey brown.

In my view the questions experiments seek to answer should be viewed as bearing on the quality of experiments. Thus we can evaluate experiments in terms of method, hypothesis, and research or application status, and also information retrieval system concern: some aspects of document retrieval systems are more central and important than others, for example searching and matching as opposed to the quality of abstracts used as a basis for indexing, or the convenience of the online searcher's terminal. The core of an information retrieval system is the document access information, i.e. the character of the indexing data and search mechanisms available. The character of the users, of the literature, of the physical and administrative plant, and so on, represent progressively more peripheral environments of the indexing and search functions. We may therefore, other things being equal, rate studies concerned with the core of an information retrieval system as more important than those directed at the periphery.

The influence of the experiments which have been carried out can, on the other hand, be dealt with by an historical account. A chronicle version of retrieval experiment does not match the logical characterization just described particularly well, so an historical account of testing is required to balance an evaluative one. The choice and sequence of experiments has naturally been influenced by the challenges posed by the findings of specific tests, but it has also been affected by developments in operational systems and in broader changes in attitudes to information system provision.

The remainder of the chapter will therefore be organized as follows. I shall first provide a summary view of the history of information retrieval experiment in its wider context, mentioning noteworthy tests in passing. I shall then consider these and other representative experiments from an evaluative point of view, in relation to their **objectives**, i.e. their focus, motivation, and underlying assumptions; in relation to their **forms**, i.e. broadly speaking their data and conduct, which can be itemized utilizing Bourne's useful scheme[5] as covering

(1) corpus size (requests and documents) and subject,
(2) source of the requests,
(3) degree of request negotiation with the user,
(4) number of relevance levels (excluding non-relevance),
(5) status of the relevance judges and basis of their judgements,
(6) performance measures;

and in relation to their **results**, i.e. their findings, the interpretation given to these findings, and their implications. This evaluation will be primarily retrospective, but some reference to what the experiments looked like at the time may be appropriate. Overall this survey will seek to show whether and how experiments have changed in their objective or type of objective, their form, and their results, and more particularly if any changes reflect a growth of experience in the conduct of information retrieval tests and in the understanding of retrieval systems. Following the detailed discussion I shall summarize the main features of the test work done, viewed as a whole. It turns out that the research of the period covered by the chapter can be naturally divided into that of the decade 1958–1968, and that of the decade

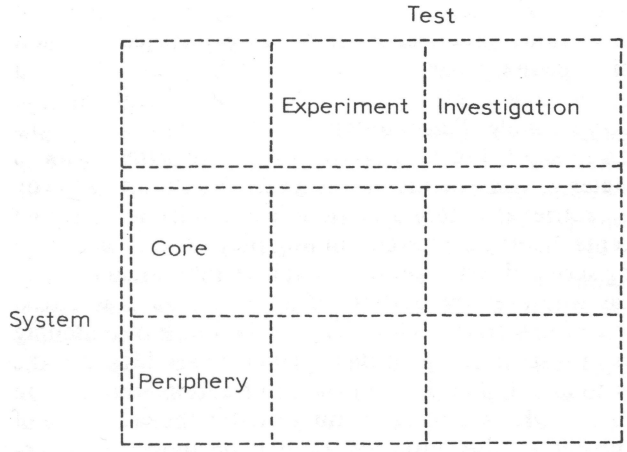

Figure 12.1

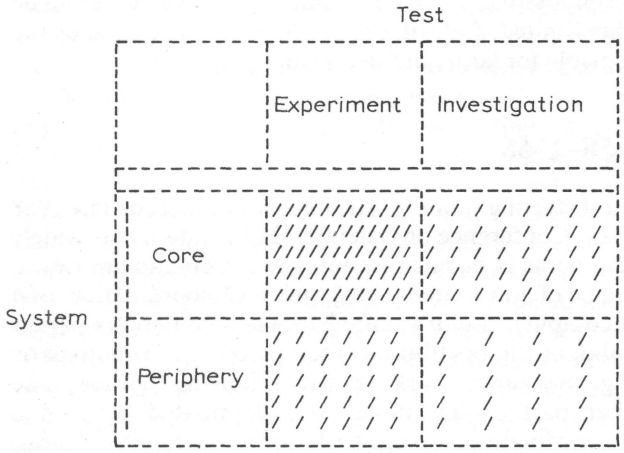

Figure 12.2

1968–1978, and the two will therefore be treated separately. In conclusion, I shall consider the status that experimental work in retrieval as a whole has achieved, and its wider impact, especially on operational systems.

Thus summarizing the discussion so far, we have testing ranging from experiment proper to investigation, and system phenomena ranging from core to periphery. Setting the one against the other gives us *Figure 12.1*. We then superimpose the concern with system function, which we can represent by shading, so we get *Figure 12.2*. The main focus of the chapter is therefore with the most densely shaded area of the diagram: evaluation experiment. However, as the boundaries are not absolute, some reference to the broader picture will be made, but only as far as this is justified by the need for a broad historical account.

In describing and relating the various tests, some simplification is

inevitable. In particular, it is virtually impossible to apply any rigorous definition of a 'unit' test or 'unit' experiment as, say, an explicit comparison between two values of a primary variable, with all other variables held constant, or perhaps a comparison between two values of a primary variable for two of some secondary variable. This is to some extent because definitions would lead to intolerable detail, but also because much reported work is rather difficult to characterize consistently at this level: this in turn is partly because, as noted above, retrieval system behaviour is not well characterized in terms of its components. Some large or continuing projects can indeed be described as conducting series of tests. But in general, an individual test will be taken, informally, as whatever the authors of a paper regard as a test, which is chiefly a matter of objectives. This has the advantage of matching the authors' own views of tests in terms of their primary variables, but the disadvantage of failing to take full account of the information embodied in multi-variable tests. That is, where authors are interested in the behaviour of a primary variable subject to the variation of one or more secondary variables, we may turn the test upside down and view the secondary variables as primary. However attempting to examine the mass of tests done from all points of view would be impossible, so, though some alternative views will be noted, these will be rather limited, and will be mainly those recognized by the research workers responsible for large, multi-variable tests.

12.3 The decade 1958–1968

The year 1958 is a natural starting point for the historical account. The 1958 Washington International Conference on Scientific Information was widely felt to mark new developments in documentation and information retrieval, specifically the appearance of a new intellectual tool, post-coordination, and a new physical tool, the computer. Luhn's auto-abstracts of conference papers may be taken as a symbol of the possibilities then perceived for automatic information processing. Research work in the following decade, and especially in the earlier part of the 1960s, was dominated by studies comparing newer post-coordinate indexing, perhaps involving a thesaurus, with older classificatory approaches. The expansion of computing was associated on the one hand with research on fully automatic indexing and searching systems, and on the other with work on automated searching. As had already been demonstrated by the use of punched card machines, post-coordination was especially suited to automation, and formed the basis of studies of automatic indexing and searching. Research on statistically-based indexing, stimulated by Luhn, was especially prominent in the early 1960s. It was soon recognized that identifying indexing keys by direct automatic content analysis was not a realistic shorter-term aim, and statistical techniques for extracting information about words and word relations were proposed as substitutes. There was considerable enthusiasm for automation, and optimism about its potentialities, reflected in the effort devoted to machine translation. The hardware and software limitations of the machines available nevertheless made research into all kinds of automatic information processing methods very difficult.

Post-coordination and automation were essentially responses to the

increasing specialization and growing volume of the literature. The felt inadequacies of crude natural language indexing of the kind originally represented by Taube's Uniterms, and the difficulties of replacing it by anything more sophisticated done automatically, encouraged the characteristic development of the 1960s: the large thesaurus. This was a human product, used for manual indexing, but to provide index descriptions of documents increasingly exploited in machine searching, or more accurately, in machine scanning for document descriptions matching a human search specification. The old professional dogma about the need for sophisticated indexing, and the new economic fact about the potentialities of automated databases, were combined in the large batch document retrieval systems established during the middle 1960s.

The proposals for novel intellectual and technological approaches needed testing by controlled experiment, or at least investigation. Thus quite apart from automation, it was apparent that the new indexing methods, especially post-coordination, whether applied with natural or a controlled indexing language, should be compared with more established methods. This applied to facetted classification as well, for example. The application of conventional methods within an automated enviroment also called for studies, primarily relating to costs. At the same time, the innovative approaches to automatic indexing as well as searching required extensive testing, both for feasibility and effectiveness. Thus during the first decade after 1958, experimental work was primarily focused on tests comparing forms of manual indexing, primarily in terms of the indexing languages used, on studies of the effects of automation on systems involving manual indexing, and on wholly automatic methods of document and request characterization and searching. However most studies of systems involving automatic searching with manual indexing were less studies of the effects of automation as such than studies of the behaviour of indexing languages. Indeed the salient feature of the testing done between 1958 and 1968 was its concern with indexing languages.

Most of the tests done in the period, and all of the major ones, therefore fall into one or the other of two groups: one concerned with manual indexing using manually constructed indexing languages, and the other with automated indexing. The first group includes the various Cranfield tests[1–3, 6], Schuller's test[7], the Syntol work[8], Altmann's[9], Blagden's[10], and Shaw and Rothman's[11] projects, Lancaster's Medlars investigation[12, 13], and the series of (Case) Western Reserve University (CWRU) studies[14, 15]. The problems encountered with post-coordinate indexing using a thesaurus led to a whole subgroup of tests on roles and links, including those of Sinnett[16], Herner, Lancaster and Johanningsmeier[17], Cohen, Lauer and Schwartz[18], Montague[19], and van Oot *et al.*[20]. The second group of tests, on automatic indexing, includes those conducted by Dale and Dale[21], O'Connor[22], Damerau[23], Borko[24], Tague[25], Melton[26], Dennis[27], and Stone and Rubinoff[28], by the Smart Project[29–32], and at A. D. Little[33, 34], as well as the other research reported in Stevens, Heilprin and Guiliano[35] and Stevens[36]. The tests of indexing languages in the first group focused mainly on controlled languages, with some work on simple natural language indexing. Those of the second were sometimes concerned simply with automatic indexing, e.g. Dale and Dale's and Dennis' tests, sometimes with comparison between automatic and manual indexing, as in the Smart tests and Melton's, Damerau's and Borko's

experiments. An important difference between the tests in the two groups was that on the whole the tests on manual indexing were evaluation tests of system performance, while more of the work on automatic indexing was concerned solely with demonstrating that automatic indexing was feasible and produced plausible output: this applies to the studies done by O'Connor, Borko, Stone and Rubinoff, to A. D. Little's NASA study, and indeed to most of the work reported in the two Stevens volumes. Within the two groups some tests can be described as experiments proper, involving some degree of control of variables and explicit comparison, ordinarily between different indexing languages. The Cranfield and CWRU studies fall into this category. Others, like Lancaster's Medlars test, were investigations. Some of the comparative tests, like those of Schuller or Cohen *et al.*, as well as the investigations of Lancaster, were directly related to operational systems; others, including virtually all of the work on automatic indexing, were laboratory studies.

The fact that in most of the work the emphasis was on the indexing language used gives the research of the decade a distinctive character. Indeed 1968 genuinely marks the end of one phase of research. For manual indexing, it could be called the Cranfield decade. The relatively uncontrolled comparisons of Cranfield 1 were followed by the more detailed tests of Cranfield 2. The experience gained in different Cranfield projects was, moreover, applied in, e.g. the Herner *et al.* study of the Bureau of Ships system, in the Medlars investigation, and the CWRU Comparative Systems Laboratory work. The CWRU Report of 1968[14] essentially constitutes an extended presentation of the testing methods developed in this whole context, and can be said to summarize the experience gained during the period. At the same time it was evident by 1968 that automatic indexing raised more problems than had been expected: the effort and difficulty involved in conducting well-organized and informative tests was clearly shown by Dennis' heroic experiments. Salton's book, *Automatic Information Organisation and Retrieval*[30] nevertheless marked the beginning of a new period since it emphasized the whole range of novel possibilities for information retrieval systems made available by computers, and the importance of viewing an automated information system as an integrated whole. Overall, the conclusion to be drawn from the work of the decade was expressed in the CWRU Report: the indexing language used is much less important in determining system performance than had been supposed.

Given this general characterization of tests between 1958 and 1968, we can now consider the objective, form and result of the different projects in more detail. This will involve both the substantive and the methodological properties of the tests. In the discussion I shall treat the two groups of manual and automatic indexing tests separately since, as already noticed, they were very different in character.

Index language tests

We start with the indexing language tests, and first consider them substantively. As noted, these tests included the classical evaluation studies carried out at Cranfield and Western Reserve. The Cranfield 1[1] and 2[2, 3]

projects are considered in detail in the next chapter, and so will be referred to here only as far as is necessary to give a coherent picture of the work done. The tests carried out by Schuller[7], the Syntol group[8], Altmann[9], Lancaster[12, 13], and Shaw and Rothman[11] are significant because of their character or consequences. Tests like Blagden's[10], Spencer's[37] and Newell and Goffman's[38] can be regarded as representative minor ones of the period.

In terms of the concerns of the period, the most important of these tests were those with the objective of comparing indexing languages, typically some or all of more conventional classifications, or pre-coordinate subject headings, with newer types of classification and post-coordinate indexing, particularly using a thesaurus. Thus the focus of Schuller's test was a comparison between the UDC and Uniterms, of Cranfield 1 a comparison between the UDC, alphabetical subject headings, a facetted classification, and Uniterms; the Cranfield-WRU test[6], which can be labelled Cranfield 1½, compared WRU telegraphic abstracts (essentially representing role-link indexing) with facets, Cranfield 2 a whole range of languages falling into broad groups based on natural language or controlled language terms, Altmann the 'ABC' pre-coordinate system with a simple KWIC index, Shaw and Rothman pre- and post-coordinate natural language and also a simple KWIC index, and CWRU telegraphic abstracts with manual keywords, automatic keywords, a 'meta-language' or controlled language, and subject headings. Spencer compared conventional abstract journal classification schemes with SCI. In the role/link subgroup Montague's test[19] also involved comparisons with other types of indexing. The other tests in this subgroup, by Herner *et al.*[17], Sinnett[16], Cohen *et al.*[18], and van Oot *et al.*[20] involved various combinations of terms, links, and roles. From one point of view these could be described as different indexing languages, but compared with the much larger differences of language studied in the other tests, the link/role tests could be deemed studies of a single type of language. More straightforward evaluations of single languages were those of relational indexing by the Syntol group, simple post-coordinate terms by Blagden, and the Medlars controlled language by Lancaster.

The motivation for the comparative tests tended to be to demonstrate the superiority or at any rate competitiveness of the more novel approaches involved, for example the use of facets in classification, of a post-coordinate thesaurus as opposed to pre-coordinate subject headings, of telegraphic abstracts, or of the SCI. Non-comparative tests like Blagden's were intended to show that a new method could provide satisfactory performance. In most cases the tests were concerned with effectiveness, but in some cases costs were explicitly investigated. Projects interested in the convenience or competitiveness of novel approaches were often implicit cost evaluations.

In general, the assumption made in these tests was the one mentioned earlier: that the index language is a major, or perhaps the crucial, factor influencing performance, so that the choice of language is very important.

A further closely related assumption was that indexing languages need to be sophisticated, though the contrary was occasionally held, for example by Shaw and Rothman. The specific substantive assumption made by the different projects then tended to be that the particular type of sophisticated language advocated was desirable, e.g. one using links and roles, or the more sophisticated relations of Syntol. Only the more extensive comparative tests

like Cranfield 1 or Cranfield 2 could be described as neutral with respect to specific type of language, while assuming that some sophisticated language was required. (In Cranfield 2 simple natural language was taken as a base for improvement in sophistication by the application of various devices.)

In form, these tests had much in common, perhaps not altogether surprisingly in view of the influence of Cranfield both organizationally, as in the link with CWRU, and intellectually, as in the application of Cranfield methods in Herner *et al.*'s Bureau of Ships test. The data for the tests generally consisted of less than 100 requests, and some hundred or a few thousand documents. Lancaster's Medlars test was quite exceptional in scale with over 300 requests and over half a million documents. Otherwise only Cranfield 1, Sinnett, and Cohen *et al.* used more than 5000 documents (though several experimental reports do not indicate how many documents were used). Queries varied, in some cases being genuine user queries, in others pseudo queries, and in yet others, following Cranfield 1, ones specifically based on a source document. On the whole there does not seem to have been much negotiation about the query with the user. Relevance assessments were usually made by requesters, and were normally of search output, perhaps pooled from several alternative searches; one or perhaps two grades of relevance were typical. Evaluation of any particular match output was commonly by precision, and by recall (or sensitivity as CWRU called it), though the CWRU tests substituted specificity for precision, and Sinnett noise. Blagden used noise alone, while Lancaster added novelty to recall and precision. In some cases simple numbers of relevant and non-relevant documents retrieved were used. CWRU combined sensitivity and specificity in a single measure of effectiveness. Recall was normally calculated relative to some subset of the possible relevant documents, say those identified by assessing some or all of the pooled output of alternative searches, or by assessing an independently obtained collection subset. With very few exceptions, like most of the Cranfield 2 performance characterizations, performance was calculated for simple sets of retrieved documents, giving one figure for each measure and so, for example, a pair of precision and recall values, for each particular test option.

With respect to the test results, again looking at the tests substantively rather than methodologically, the most striking feature of the actual findings for the comparative tests was the very wide variation in performance. This is true both of individual studies, and, insofar as such cross comparisons are legitimate, of groups of similar tests. Variations in the findings obtained by different projects have to be treated with reserve, since they may be attributed as much to specific measurement procedures or data statistics as to the system factors, especially languages and their application, being studied, or to their environment. In particular, variations in relative recall for different projects, especially those using pooled output, are only of real significance within the context of individual projects.

The fact that even the more plausibly grouped tests may differ in detail, for example by using average of numbers rather than average of ratios, or in using an external sample rather than pooled output for relative recall, and that in addition I have worked some figures myself, might suggest that there is no point in giving specific findings. But this is worth doing, to give the real flavour of the tests. It should however be noted that as many tests consisted

of a range of subtests, the figures are an illustrative selection; moreover, when ranges of performance figures are given, these may, for projects with rather heterogeneous subtests, be only for the strictly comparable alternatives of a single subtest.

The variations in individual project findings is well illustrated by Montague's different experiments, where precision ranged from 4–9 per cent in one case, with relative recall 83–31 per cent, to 46–74 per cent and 93–31 per cent respectively in another. Considering the tests comparable in type, i.e. in objective and form, the tests conducted by Schuller, Altmann, and Shaw and Rothman, with Cranfield 1½, can be considered as a group, along with those of Sinnett, Cohen *et al.*, Montague, and van Oot *et al.* on links and roles. The individual projects report differences in precision ranging from 12.5–41.7 per cent (Schuller), 51.4–96.5 per cent (Cohen *et al.*, for variable query sets), 46–74 per cent (Montague) or 67.3–88.7 per cent (Altmann), and in relative recall from 31–93 per cent (Montague) or 80.7–100 per cent (Cohen *et al.*) Taking precision and (relative) recall together for comparable data runs in multi-test projects we get such variations as 12.5 per cent precision and 73.1 per cent recall—41.7 per cent precision and 77.4 per cent recall (Schuller), 42 per cent precision and 57 per cent recall—55 and 66 per cent (Shaw and Rothman), 57.4 per cent precision and 100 per cent recall—94.0 and 80.7 per cent (Cohen *et al.*, with variable request sets), and 70 per cent precision and 84 per cent recall—90 and 77 per cent (van Oot *et al.*). The operational non-comparative studies of Herner *et al.*, Melton, and Lancaster were broadly in the 50–60 per cent range for both recall and precision. For recall (sensitivity) alone, CWRU results ranged from 16 to 98 per cent, while the normalized recall results for Cranfield 2 ranged from 44.6 to 65.8 per cent. It may be noted that the specificity results for CWRU ranged from 12 to 98 per cent. Over all these tests taken together, results range from a low of 4 per cent (Montague) to a high of 96.5 per cent (Cohen *et al.*) in precision, and, as far as the comparison is proper, from 31 to 100 per cent in recall (Montague and Cohen *et al.* respectively).

These variations can or could, as noted, be accounted for partly by methodological differences and partly, of course, by the real properties of the languages being investigated, or the values of dependent variables like indexing exhaustivity; they might also be attributable to environment factors like collection subject area. These points are more fully discussed later. It must nevertheless be emphasized that the variations are not wholly explicable: if they were we would know how to design information retrieval systems; and the sheer scale of observed performance variation is worth noticing.

The interpretations of the findings are equally varied, though there is a natural tendency, for the more limited tests, for their authors to conclude that whatever was to be demonstrated has been demonstrated. For example, Shaw and Rothman conclude that roles and links are not needed, while Schuller, testing novel Uniterms against UDC, finds Uniterms superior, though he concedes the complementary utility of UDC. However in some cases the results, like those of Cranfield 1½, were contrary to expectation, and in the more broadly ranging comparative tests, like Cranfield 2 and CWRU, the results were surprising: in the first that natural language is competitive, and in the second that the indexing language is not very important.

Index language test overview

The most fruitful way of looking at the results obtained in the tests of the period is to see how the specific findings and the interpretations given them by those concerned can be seen, when the tests are taken together, to show common trends with implications for our understanding of information retrieval system behaviour in general. What emerges from the tests of 1958–1968 is unlikely to be novel to those familiar with the work, but it is worth emphasizing the fact that these very broad conclusions are supported by the results obtained over a range of projects, and are not simply based on single tests.

Thus the tests comparing different languages, like Spencer's, Shaw and Rothman's, the various Cranfield tests, and the CWRU project, show that, other things being equal, different languages achieve comparable levels of performance though they may retrieve different sets of documents and especially relevant documents. A natural corollary is that fancy indexing has no especial merit, as Cranfield 1½, Montague, and Melton show, or, to put it the other way round, that simple indexing has merit, as Blagden, Cranfield 2 and Shaw and Rothman indicate. A related conclusion is that supported by Cranfield 1 and CWRU, that the indexing subsystem is not the overwhelmingly important factor in determining system performance. In Lancaster's and Saracevic's view, based on the Medlars and CWRU studies respectively, the treatment of the question, and specifically its proper development, emerges as much more important. Saracevic's general conclusion is that human factors are the most important ones. The related system factors most affecting language performance seem to be the exhaustivity or depth of indexing, noted for Cranfield 1 and 2 and CWRU, and also, according to Cranfield 2, the specificity of the indexing language.

The tests taken together indeed support the statement that there is an inverse relationship between recall and precision, which was explicitly studied in the Cranfield 2 experiments, which is influenced both by indexing policies for documents or requests, determining exhaustivity, and by indexing resources in languages, determining specificity. The more detailed studies of the link/role test subgroup provide particular evidence here, showing that both links and roles are precision devices, with roles especially restrictive: and other studies of indexing with relations, like the Syntol test and Cranfield 2, show a similar restrictiveness. Thus the statement that, other things being equal, languages perform the same has to be read as meaning that languages perform the same if document dependent factors are held constant and the languages are not explicitly oriented in opposite directions with respect to recall and precision: if good levels of both recall and precision are required, then when document variables are held constant, languages representing, for example, rather different classificatory philosophies do not differ materially in behaviour.

Of course, these observations can only be taken as very broad generalizations, given the great variations in the details of the tests of the period, and also their many methodological deficiencies. The latter might indeed be regarded as sufficiently gross in many cases to undermine any conclusions to be drawn from the tests, but an alternative view is that the tests, however defective, were sufficiently varied that any common result can be regarded as

likely to represent a persistent underlying reality. The methodological inadequacies of many of the tests, as illustrated by Sinnett, Cohen *et al.*, or Montague, for example, are nevertheless very conspicuous. The CWRU project was indeed specifically intended to constitute a study and development of retrieval system testing methodology. The defects of many of the tests are incidentally compounded for subsequent criticism by inadequate reporting, for example about the number of documents searched.

Many experiments suffered from a general lack of control, both with respect to the values of the variables of interest and those of more obviously or possibly related variables. Thus if the variable of interest is the indexing language, when only one is studied it is not obvious how far the resulting performance should be attributed to the language itself. Conversely, since for example indexing depth may be a dependent variable, depth should at least be held constant, and preferably also systematically changed. A minimal test would therefore compare languages A and B with respect to indexing depths I and II. The Cranfield 2 project was deliberately intended to improve on Cranfield 1 in such respects, and, as just noted, the CWRU project was designed to make such properly controlled comparisons, and indeed included subsidiary tests to validate the effectiveness of the controls. Other tests, notable examples being several of those on links and roles, attempted reasonably careful comparisons. A number of studies, though perhaps not involving a high degree of control, included failure analysis. This was done by van Oot, for instance, and on a large scale by Lancaster. Failure analysis is not part of an experiment proper, but makes a very important contribution to the broader study of retrieval system behaviour.

Some authors indeed comment, like Schuller, on the problem of testing, or at least recognize the limitations of their own tests, for example in sample size. However some particular methodological inadequacies recur in the tests of the period, along with the specific failings of individual tests. These defects can be categorized as first those concerned with the propriety of the way a real system is being modelled, second those concerned with statistical aspects of the tests, and third those of evaluation.

In the first category the most noticeable deficiency is the wide use of 'bogus' queries, i.e. queries not put to the system in the ordinary way by its users. In Cranfield 1 and tests influenced by it like that of Herner *et al.*, source document questions were used, i.e. questions based on and designed to retrieve specific documents; and in other tests, like those of Montague and Cros *et al.*, synthetic, made-up questions, assumed typical of real ones, were used. The results obtained with real and artificial queries may not differ, but where this has not been demonstrated, there must be doubts about the validity of tests with artificial queries. Some tests, like the subset ones with 200 documents in Cranfield 2, or Newell and Goffman's, used specifically constructed document files, i.e. ones with a high density of related papers. Others, like van Oot *et al.*'s, used languages specially constructed for the test document set.

As far as the statistical aspects of testing are concerned, one of the most striking features of the tests of the period, taken as a whole, is the small number of requests used. For example, Sinnett used 22, Shaw and Rothman 9, Cohen *et al.* 14–33, Montague 29, 33 and 10, Melton 12, Spencer 1 (admittedly not a query in the ordinary sense), Shaw and Rothman 9, and

Herner *et al.*, for calculating recall, 10. Document samples were better, but were sometimes small: for example Cranfield 2 used 200 documents in many experiments, Newell and Goffman 210.

In the measurement of performance, the most pervasive methodological inadequacy is the arbitrary treatment of recall, or at any rate the rather particular interpretation given to it, without any awareness of possible bias. Newell and Goffman, and Melton, for instance, define decall as the retrieval of cited documents, Cranfield ½ (Warburton and Cleverdon), 1, and 1½ recall relative to a source document, the Syntol group recall for automatic abstracts relative to manual, and Lancaster recall relative to an independently obtained set of relevant documents. In other cases recall is measured relative to the pooled output of alternative searches, but then performance for an individual language being tested depends on the character of the different pool contributions, which may not be strictly comparable.

Individual tests moreover reveal a variety of other dubious procedures, for instance Cohen *et al.* compare various link/role combinations using different numbers of queries, and even in the in many ways model CWRU tests, results are lumped together oddly, for example those for different indexing languages are combined to provide performance figures for different indexing sources.

Overall, the tests taken together can only support the broadest and most tentative conclusions: the variation in data was vast, and the performance measures used were not only directly incomparable, for instance where one project uses precision another opts for specificity, but incomparable in more subtle ways, for example in averaging technique. Moreover, as relative recall depends on an 'arbitrary' base, it can give very different results according to base: specifically, values will be absolutely higher for languages with a similar performance than for those with a different, but complementary, performance.

Automatic indexing tests

As noted earlier the character of work in automatic indexing was rather different from that done on manual indexing. The many theoretical and computational problems involved meant that more work had to be put into simply establishing the feasibility of procedures and prima facie plausibility of results. There were therefore more studies of a non-evaluative kind, and fewer evaluative ones.

This is not the place to review work on automatic indexing in detail. Briefly, it was primarily concerned, on the one hand, with statistical methods of identifying, by extraction from text, words representing individual documents or sets of documents, and on the other with statistical methods of recognizing relations between words supplying substitute or additional search keys. The work was very much within the framework of post-coordinate indexing, and was chiefly devoted to the examination and programming of statistical methods. Related research was concentrated on simpler methods of keyword selection, for example by text location, as in O'Connor's work[22], or on the application of statistical techniques in assigning items from a manual indexing vocabulary, as in Gotlieb and Kumar's test[39]. Closely related ideas studied were those of term weighting and output ranking.

Actual evaluation tests included Dale and Dales's[21], designed to examine alternative clustering bases for establishing term associations, and similar ones conducted by Smart workers[29–32], especially Lesk's[40], at A. D. Little[33, 34], by Tague[25], and by Dennis[27]. The Smart workers and Williams and Perriens[41] also investigated weighting. The Smart tests and Tague's included comparisons between automatic and manual indexing.

These tests were too few and too heterogeneous for systematic comparison under headings to be worthwhile. It is sufficient to note that, overall, they tended to show rather little difference in performance for the various statistical techniques studied, In comparisons with manual indexing, generally using simple extracted terms but sometimes, as in some Smart tests, thesaurus terms, the general conclusion was that performance is roughly comparable.

Unfortunately these tests were vitiated by even greater methodological defects than those associated with manual indexing studies. Dale and Dale, for example, used only 4 requests, Smart tests very small document sets, sometimes containing less than 100 documents. Dennis' studies are an honourable exception since by the standards of the day they were on an enormous scale, particularly as far as the document sample was concerned. However the request set was typically small, only 6 in one experiment. Dennis' test was in many ways typical of the period in mixing valiant attempts at control in some directions with serious failures in others, to produce a rich but unevenly cooked whole.

The major non-evaluation tests included, for example, Damerau's of text extraction using 7 articles[23], Stone and Rubinoff's of the collection vocabulary[28], Borko's extended study of classes obtained by factor analysis[24], and the A. D. Little study of term association techniques for a very large NASA document collection[34]. Both Damerau and Borko judged their automatic indexing results by comparisons with manually selected or grouped words. The A. D. Little study in fact included a crude performance test for a single search, but evaluation was chiefly simply by inspection of the statistical association products.

A few interesting studies, for example by the Smart Project[30] and by Melton[26], were concerned with non-statistical automatic text analysis specifically designed to identify syntactic relations between words. However these tended to show similar results to those obtained with manual syntax, and the work could in any case not be carried far because of the unresolved general problems of linguistic analysis.

The main emphasis in automatic indexing work was indeed on statistical approaches, but even here the retrieval system testing done was much less substantial than that done on manual indexing. This is not wholly surprising, since the methods had to be worked out before they could be tested. However the many difficulties encountered damped the enthusiasm of the early 1960s, particularly since the problems of devising and validating statistical methods were compounded by the problems of conducting information retrieval tests of any kind which were increasingly recognized by research workers. Dennis' project was at least as discouraging in showing absolutely poor performance for vast work as encouraging for showing that something could be done, and only a few projects like those of Sparck Jones and Salton were involved in serious statistical indexing and evaluation testing by 1968.

Other tests

Outside the two groups of tests discussed were a few others concerned with what we earlier referred to as the retrieval system core: Tague's study of the role of question terms in matching relevant documents is an example[42]. Further, supporting the evaluation experiments involving the retrieval system core were some non-evaluative studies, usually of an investigative rather than experimental character, concerned with such topics as the character of indexing vocabularies or properties of document sets. Their importance in automatic indexing has already been mentioned: in connection with manual indexing such studies as those of Houston and Wall[43] and Heald[44] can be mentioned.

Round these core tests we can then group studies in other more peripheral areas. Among these are two large subgroups, of user studies and bibliometric studies. User studies naturally began to appear accompanying the development of novel, large, or automated systems in the 1960s, and a great many have been carried out. Early studies were mostly based on questionnaires. Unfortunately, as such reviewers as Menzel[45] and Herner and Herner[46] noted, many of these studies suffered from methodological failings like poor sampling or the use of ill-designed questionnaires. Bibliometric studies also became popular in the 1960s, boosted by the Science Citation Index, but these too often exhibited methodological failings, especially in the assumptions made about the propriety of the clustering techniques used.

Finally, it should be noted that alongside the work discussed so far, which was explicitly or implicitly concerned with effectiveness, went studies of system efficiency, i.e. cost. Some of the evaluation tests already mentioned, like van Oot *et al.*'s, included cost analyses, but other studies only of costs were carried out in the period (see King[47]). The development of techniques for conducting cost analyses is of course relevant to that of testing in general.

12.4 Conclusion on 1958–1968

Looking at the decade 1958–1968 as a whole, it is possible to detect some consolidation of actual findings, and some development of testing methods and improvement in experimental standards. The main findings were those mentioned earlier as conclusions to be drawn from the indexing language tests, with the tentative rider from the automatic indexing work that the simple indexing found competitive in the manual tests can be provided automatically.

The main findings of the decade were strikingly exemplified by the Cranfield 2[2, 3] and CWRU[14, 15] results, and are well expressed by Saracevic's comments on the latter. Thus in his conclusion to the CWRU Report[14] Saracevic notes, as overall observations about information retrieval systems, the importance of human factors in maintaining adequate performance (a comment endorsed by Lancaster in calling for quality control for Medlars[12]); the fact that system performance can nevertheless only reach a middling level; and that an inverse relationship holds for getting relevant documents and avoiding non-relevant ones. The inverse relation of recall and precision was emphasized by Cleverdon, and, as Lancaster and Mills noted[48], as there is an inverse relation, one should *design* a system for a particular point along

the recall–precision line. These conclusions were endorsed by the multi-collection tests done in the later part of the decade by the Smart Project.

Methodologically the Cranfield 2 project showed how informative testing required a more systematic breakdown of a retrieval system into its various factors than was common earlier. Gross comparisons between distinct languages were replaced by a much more detailed study of recall and precision devices generating families of languages. The CWRU project was similarly directed toward a much more careful treatment of system factors in a range of comparative experiments than was generally adopted. However it is interesting to note that the attempt to ensure control in CWRU led to new difficulties, in this case to a very artificial and perhaps perverted treatment of queries, i.e. maintaining constant queries for different languages tended to suppress distinctive features of the languages themselves. The same trend is well shown in the Smart Project work which by 1968 was well into a very large range of detailed studies. In this case the emphasis on automatic systems provided not only new opportunities for system design, for example in permitting ranked output, but also ones for system testing in the comparative ease with which grinding tests over ranges of slightly different variable values could be conducted, and in the application of complex measurement and statistical evaluation techniques. However the sheer proliferation of explicit parameter settings served to bring out not only the increasing numbers of runs needed for proper comparative experiments, but the difficulties of ensuring a meaningful experimental design.

The Cranfield 2 and CWRU projects in many ways looked backward, seeking to improve on the initial index language tests of the decade. But they also, in the challenge implied by the comparative flatness of their findings, and in their methodological quality, presented a reference point for the work of the next decade. The Smart Project, while sharing this character to some extent, is a more genuine pointer to future work in more throroughly embracing the possibilities offered by the computer, particularly for sophisticated search strategies and non-conventional methods of massaging term descriptions, for example by numerical weighting and feedback techniques.

But even these projects suffered, as already indicated, from many limitations; and the general character of the testing done in the field during the decade is well described by Saracevic:

> 'At present real and productive testing of total retrieval systems, taking into account and controlling all inside and environmental factors, is not feasible and not possible. At present, it seems that generalisable, formal, quantified results of high validity and reliability on all or even on the majority of factors affecting the performance of retrieval systems cannot be attained.
>
> The reasons are fairly evident. There is an absence of a well-formulated theory taking into account all or a majority of the factors operating on retrieval systems. There is only an intuitive understanding of objectives of retrieval systems—thus, the measures indicative of the achievement of objectives are not totally reflective of the real objectives and not comprehensive. There is an inadequate knowledge of processes involved within or outside the IR systems and, without a thorough understanding of processes, comprehensive testing is unattainable. There is a lack of standardised methodologies for experimentation, which precludes testing.

> Most importantly, there is an inadequate understanding of controls in experimentation with IR systems and the controls are essential in monitoring the factors under consideration and distinctly sorting out the factors contributing to the performance. There is a lack of an effort to cumulate and synthesise knowledge on IR systems as it exists.' (Ref. 14 Part II, pp. 183–4)

In general, therefore, the situation at the end of the first decade of information retrieval system testing was that while the test results tended (broadly) to agree on what happens in retrieval systems, they did not sufficiently explain why it happens. In particular, at the more detailed level, in cases where performance differences were observed, these were not always attributable to specific system factors or, more importantly, to the interplay between system factors. It was thus not at all obvious how systems should be designed to perform well, modulo a preference for recall or precision, in particular environments, especially outside established frameworks like those represented by the Medlars system, or for situations and needs clearly resembling those of existing systems. It was even less evident how 'optimal', i.e. attainably good, performance was to be achieved for a given area of the recall–precision spectrum. For while there is a general inverse relationship, it does not follow that for a specific value of precision (or recall) one cannot establish a better average recall (or precision) than the current one. One needs at any rate to know whether the current performance level is a good one. Greater understanding was thus the prime need in the next decade's testing.

12.5 The decade 1968–1978

The testing work of the decade 1968–1978 differs from that of 1958–1968. It shows both a shift in the main topics of concern and, especially in laboratory work, greater refinement in the attempt to distinguish and control variables. The volume of experimental work seems to have been greatest in the earlier part of the decade, with a number of projects in particular stimulated by the major tests of the previous decade like CWRU's and Cranfield 2. In the latter part of the 1970s there has been a noticeable decline in the number of laboratory experiments, presumably because the rapid extension of online services has been widely, though in some opinions too uncritically, accepted as solving all the information user's problems. This development has been naturally associated with service investigations and management and cost-oriented studies.

Overall, the evaluation tests of the decade fall into five major groups, compared with the two of the previous decade, though these five groups do perhaps, as we shall see, fall into two very broad classes roughly corresponding to the two groups of the previous decade.

In the early 1970s there were a number of reports on manual index language tests of the kind conspicuous in the previous decade; and indeed these projects had typically been started in the late 1960s: examples are the tests done by Aitchison *et al.*[49] and Olive, Terry and Datta[50], and Keen's ISILT experiment[51, 52]. Some of the tests involved retrieval using different

bibliographic record fields, like titles or abstracts, which were regarded as representing different indexing languages, rather than a single language, namely natural language, used for indexing from different sources. The UKCIS investigation[53, 54] illustrates this approach. However somewhat greater care was taken in this group of tests in the treatment of such dependent variables as indexing exhaustivity than was usually the case in the previous decade's tests.

The second group of tests was indeed concerned with indexing rather than with the indexing language used, and particularly with exhaustivity and specificity. Thus Schumacher, March and Scheffler's test[55], for instance, was concerned with the effects of exhaustivity on performance, as was ISILT. Tests on indexing language specificity, like Svenonius'[56], also fall into this category of more detailed studies of single variables.

The conclusions about the importance of searching reached in some of the earlier tests were followed up in a number of studies of searching, which has also been a topic of interest to those responsible for online services. Some of these studies, like that of Katzer[57], were concerned with the form of the query in a narrow sense, others like those of the UKCIS group or Leggate *et al.*[58], with the properties of user queries, and yet others like those of Barraclough *et al.*[59] or Keen's EPSILON test[60], with the behaviour of users in searching. Tests with broad or narrow question formulations like Aitchison *et al.*'s also fall into this group.

A particular trend of the 1970s has been an interest in weighting and its natural corollary, output ranking. In some cases weighting has been determined by the properties of individual documents, or of the collection as a whole, so the tests really fall under the heading of index language or indexing studies; but in other cases weights are associated specifically with query terms, representing *a posteriori* rather than *a priori* document indexing, and weighting here is more properly subsumed under searching and the organization of search output, especially by non-boolean matching functions. Different tests have to be examined very carefully here to determine their true rather than apparent concern: for example document set weights calculated at search time for the query terms only are nevertheless logically distinct from individual query weights. In fact, though tests with manually assigned weights have been carried out, for example by Evans for query terms[61, 62], most of the work done on weighting has been done in the context of automatic indexing. The development of research on weighting in this context has, however, paralleled that of work on manual indexing, in that the emphasis has increasingly been on the role of weights in searching. Thus the most noticeable feature of retrieval research in the 1970s has been the experimental work on the general idea of relevance feedback, and on relevance weighting in particular, within automatic systems. Research in this area was begun by the Smart Project in the 1960s, and is represented by a long series of experiments through the decade[4, 63–67]. Other tests in the area have been conducted by Miller[68–70], UKCIS—Barker, Veal and Wyatt[54, 71], and subsequently Robson and Longman[72, 73]—Cameron[74], and Sparck Jones[75–77]. This approach to searching is particularly interesting in being that most conspicuous in the whole area of information retrieval in having some solid theoretical underpinning.

These relevance feedback and weighting techniques are largely statistically

based, and so are connected with other statistical approaches to retrieval. The statistical work of 1968–1978 is in turn linked with that of the earlier decade. As noted, much of the automatic indexing research done between 1958 and 1968 did not progress as far as evaluative performance testing. The early 1970s saw reports on statistical term cluster evaluation by Vaswani and Cameron[78] and Sparck Jones[79, 80], and recent experiments by Harper and van Rijsbergen[81] have specifically combined the use of term associations with that of relevance weights. Other rather crude methods of non-statistical automatic (or in principle automatic) indexing are represented by O'Connor's work on passage retrieval[82, 83] and by Atherton's BOOKS project[84]. Klingbiel and Rinker[85] and Evans[86] report tests of semi-automatic indexing involving some reference to a dictionary or thesaurus. In general statistical clustering has proved very disappointing, and the main thrust of statistical work has been on the more promising weighting. For the purposes of discussion we can therefore consider two groups of tests: those on automatic and especially statistical indexing not involving relevance information, and those on relevance feedback and weighting.

Some of the language, indexing, and searching tests were carried out in the context of operation services, for example by Aitchison *et al.*, Barker *et al.*, Olive *et al.* and, recently, Cleverdon[87]. There have also been more restricted investigations, rather than experiments proper, relating to services, such as those carried out by Rowlands[88], Lancaster, Rapport and Kiffin Penry[89], Leggate *et al.*, Hansen[90], Simkins[91], and Pollitt[92]. Such operational tests were often concerned, and perhaps more than those of the previous decade, with cost efficiency as well as performance effectiveness, and some studies, like that of Katzer[57], have been wholly devoted to costs. The increasing volume of information and development of information services have also been matched by a corresponding growth of user studies, data base coverage investigations, and so on. There have also been many bibliometric studies, some of a very academic character.

Overall during this period we can detect two major strands in testing, reflecting an increasing divergence between the concerns of operational system managers and those of research workers. Projects under the first head concentrated initially on indexing languages and then on the related topics of indexing and searching. Research workers have also concentrated increasingly on searching, as in the relevance weighting experiments, but within the framework of theoretical approaches implying sophisticated procedures like output ranking. It is therefore paradoxical that some research findings which appear particularly suited to modern computer systems should have made no impact on the operational scene.

From a more intellectual point of view it will be evident that we can combine the five topic groups listed to form two broader groups of test which in fact continue the previous decade's interests in manual and automatic systems respectively. Thus the work on index languages, on indexing, and on user searching strategies is all oriented towards manual systems or the human elements of automatic systems. The work on statistical or other 'mechanical' forms of indexing, and on statistical and 'mechanical' techniques for query modification, on the other hand, is a continuation of the automatic indexing research of the 1960s. In what follows, these two broad groupings should be borne in mind, though the detailed discussion is more conveniently and,

from the point of view of test evaluation usefully, organized by the five specific topics.

The work of 1968–1978 is more variegated than that of the previous decade, and it is less easy to describe in a tidy way. Concrete comparisons between tests are more difficult to make, and comprehensive generalizations about groups of tests cannot always be provided. In some groups no tests stand out as especially important; however for the decade as a whole we can single out the tests or sets of tests done by Aitchison *et al.*[49], Keen[51, 52], Vaswani and Cameron[78], Miller[68–70], UKCIS[53, 54, 71–73], and perhaps Sparck Jones[75–77, 79, 80, 93–95] as significant in terms of scope, conduct or result. Some tests, like Aitchison *et al.*'s and Keen's, resembled Cranfield 2 in touching on a wide range of questions. The Smart Project work as a whole is very important[4, 63–66, 96–98].

Turning now to individual tests, or more particularly experiments, both important and representative, the question is what changes and developments are detectable in their objectives, forms, and results. As in the discussion of the work of the decade 1958–1968, the tests will be considered first from the substantive, and then from the methodological points of view; but in this case all five groups will be treated substantively before methodological questions are considered.

Index language tests

The tests in the first group were focused on comparisons between different indexing languages. This group is exemplified by Jahoda and Stursa's test[99], Cleverdon's three tests[87, 100, 101], and those of Aitchison *et al.*[49], Barker *et al.*[53, 54], Olive *et al.*[50], and Keen[51, 52]. Jahoda and Stursa compared single subject access with a KWIC index, Cleverdon controlled thesaurus-type languages with natural language, Keen's ISILT several controlled languages and natural language, Aitchison *et al.* and Barker *et al.* chiefly different natural language texts like titles and abstracts. Smart Project experiments included manual controlled versus automatic natural language comparisons in the Medlars tests[4], and Miller, in working on searching, tested controlled MeSH versus natural language[68, 69]; Evans compared manually and automatically assigned thesaurus terms[86], and Klingbiel and Rinker manual and semi-automatic natural language indexing[85]. Keen's printed subject investigation, EPSILON, can also be regarded, though the emphasis is on searching, as partly a language test[60].

The most conspicuous feature of these tests is the inclusion of natural language; index language tests in the previous decade were typically confined to different forms of controlled language. The inclusion of natural language, represented either by manually-selected keywords or by automatically-searchable titles or abstracts, must be seen as responding in part to the findings of earlier projects like Cranfield 2 (this was indeed explicitly the case in, for example, Aitchison *et al.*'s test), and in part to the increasing use of machine files for which title searching in particular is especially appropriate. The cost of using a controlled language with very large files, whether for indexing or searching, must be a contributing factor too. Some of the tests, like Cleverdon's DOAE test[100], and Keen's, explicitly covered dependent variables like indexing exhaustivity, and Aitchison *et al.* included question

formulation breadth as a secondary variable. In general the language tests were concerned with post-coordinate systems, though it has to be recognized that many thesauri include compound terms and so allow a kind of hybrid pre- and post-coordinate indexing. Keen's ISILT test included pre-coordinate languages, and the printed subject indexes of EPSILON embody pre-coordination. The most ambitious test of a thorough pre-coordinate system is represented by Yates-Mercer's non-comparative evaluation of Farradane's relational indexing[102]. It is noteworthy that classifications figured much less largely in the tests of this decade than in those of the previous one.

The motivation for these tests was generally the straightforward one of simply comparing the languages concerned, or perhaps of evaluating natural language compared with a given controlled language. The underlying assumption tended to be either that the different languages behave much the same, or more particularly, that natural language is competitive, as in Salton's Medlars test, for example. Yates-Mercer's investigation is thus noteworthy in that it was explicitly aimed, in contrast, at justifying a very sophisticated relational approach. Of course in all these tests, as in those of the previous decade, the implicit assumption is that the indexing language used in a retrieval system is important.

In form these tests tended to follow by now standard practice, though with rather more emphasis on real user requests, with evaluation ordinarily by precision and recall or, for the larger document sets, relative recall. Aitchison *et al.*'s and Keen's tests used exhaustive recall, but the majority of the tests were restricted to recall relative to collection subsets. The collections used tended, as before, to be rather small: only Barker *et al.* used more than 100 requests. However they, Miller, Olive *et al.* and Cleverdon[87] all used large document sets represented by regular service files. The numbers are not always given, but Miller, for example, searched some 210 000 documents.

With respect to the test results, again considering broadly comparable tests in terms of both objective and conduct, the specific findings as before show very different values for precision and recall, again not surprisingly for relative recall. Thus individual project ranges for precision were from 12 to 15 per cent for Miller (my calculation from his thesis[68]), from 27 to 51 per cent for Cleverdon (calculated by extrapolation[101]), from 46.3 to 89.9 per cent for Klingbiel and Rinker, and from 38.8 to 66.0 per cent for Barker *et al.*; for relative recall from 30–52 per cent for Cleverdon, 51–64 per cent for Miller, 49.2–73.3 per cent for Klingbiel and Rinker, to 56.4–95.7 per cent for Barker *et al.*, with absolute recall ranging from approximately 4–28 per cent for Aitchison *et al.* to 57.9–92.3 per cent for Keen. For this group as a whole precision ranged from 12 per cent (Miller) to 66.0 per cent (Barker *et al.*), and absolute recall from about 3 per cent (Aitchison) to 100 per cent (Keen). Other measures, like the numbers of non-relevant documents retrieved used by Keen, ranged from medians of 8.6 to 24.4. These are figures for simple sets of retrieved documents. A significant feature of the work of this period was the use of ranked output, for which performance representations may be obtained, with the document cutoff methods used by Aitchison *et al.*, for example, or the recall cutoff techniques used by the Smart Project. Graph comparison presents many problems; it may therefore simply be noted that, for the same calculation methods, large and presumably significant differences between graphs appear in Aitchison *et al.*, for instance. Thus in

their set of experiments, in one specific comparison on co-ordination matching (Ref. 49, *Figure 1.39*), one performance graph ranges from recall 27 and precision 6 to recall 1 and precision 33, while another, far away, ranges from recall 81 and precision 4 to recall 2 and precision 100, both having the characteristic sagging shape. This is an enclosed area difference of several hundred per cent. There are also considerable differences in the relative locations of graphs for different projects. Thus by comparison with the Aitchison *et al.* graphs just mentioned, Cleverdon[101] gives co-ordination matching figures generating graphs for recall 99 with precision 34 to recall 10 and precision 97, without sag.

In interpreting their findings the authors of the various projects tended to conclude that, other things being equal, different languages perform much the same. Controlled languages are perhaps slightly superior, but natural language is very competitive; Klingbiel and Rinker specifically found that machine-aided indexing could be very successful. The more specific conclusion drawn was that, especially where costs are concerned, natural language, and particularly automatically scanned text, is a good bargain, though absolute performance is not striking. However the results obtained by Yates-Mercer for a non-trivial document set, namely recall of 76 per cent with precision of 77 per cent, apparently show that a much higher level of performance is attainable than that generally achieved in the comparative tests, or in service investigations like Lancaster *et al.*'s (relative recall 48.0 per cent, precision 59.3 per cent)[89]. Collectively, the implications of these tests are those of the comparable tests of 1958–1968, namely that, where dependent variables like exhaustivity are controlled, languages behave similarly, and it is the other factors like exhaustivity and searching which are much more important; languages matter only in relation to the system's specific recall or precision performance objectives. The inverse relation between recall and precision is again quite clear.

Indexing tests

The studies of indexing were, as noted, especially concerned with exhaustivity: see, for example, Cleverdon[100], Keen[51, 52], and Schumacher *et al.*[55], and also Sparck Jones[94]. The general style of these tests was very like that of the language studies just described, and indeed the two were often closely connected as, for example, in Cleverdon and Keen. Schumacher *et al.*'s experiment tested description exhaustivity over an exceptionally wide range, his specific aim being to investigate the use of progressively longer sources for controlled indexing, from titles, through abstracts, to the body of the text, the sources being associated with increasing exhaustivity of index description. The topic well illustrates the difficulties of testing since the use of different sources to provide descriptions of differing exhaustivity may also introduce quality variations. Schumacher *et al.*'s test is open to this criticism, as are Aitchison *et al.*'s[49] and Barker *et. al.*'s[53, 54] studies of the use of different text descriptions in machine searching, which may be viewed as exhaustivity tests. Cleverdon, Keen, and Sparck Jones, for example, were more careful to test for indexing from the same source. In form the tests were like those of the main group, most using rather small data: Schumacher used 99 requests and 984 documents, for instance; but they differ too much in their detailed

conduct for systematic comparisons. However taken together the results show that large differences of exhaustivity do affect performance, typically trading recall for precision. Schumacher *et al.*'s findings (assuming constant indexing quality) show this very clearly: with increasing exhaustivity he obtained a substantial gain in recall, with a gradual, though not enormous, decline in precision. Thus recall relative to the full text relevant retrieved progressed from 25 per cent for titles to 72 per cent for titles plus abstracts, contents lists and author keys, while precision dropped from 65 to 56 per cent. Keen found recall rose from 74.7 to 85.8 per cent, but for an increase in median non-relevant retrieved from 18.9 to 24.4, for controlled language document indexing on two levels of exhaustivity. Cleverdon found that for varying natural language exhaustivity for both requests and documents, performance ranged from 70.5 per cent recall (relative to an independent sample) and 32.2 per cent precision to 80.6 per cent recall and 18.1 per cent precision. However, as Sparck Jones suggests, small differences are not important and exhaustivity in document indexing can be consciously counterbalanced by the treatment of requests. This is indeed implicit in the use of extended profiles for title searching in operational services. Cleverdon's results also suggest the possibility of trade-offs, as do Aitchison *et al.*'s tests of different query formulations, broad, medium or narrow.

Searching tests

The evaluation tests on searching include some of the most interesting of the decade. It is, however, difficult to give a coherent account of them, since the whole searching subcomponent of a retrieval system is an extremely complicated one, and one which is not well understood, and the different tests done have been scattered over the large area of searching as a whole.

Searching refers both to the entire interaction between a user seeking documents relevant to a need from a document file, and to any particular expression of this need used to scan some or all of the file. The latter includes the treatment of individual terms and that of the logical structure of the query, and the complex relationship between the two. This is not the place for a detailed discussion of searching, and in the summary account which follows its different aspects will be referred to very crudely. For this purpose we will therefore simply use the term 'strategy' for the searching process for a query as a whole, 'specification' for any individual matching prescription, 'logic' for the formal structure of such a prescription, and 'formulation' for the broad or narrow scope of a specification. With respect to logic, the great majority of experiments and investigations have, following operational practice, been concerned with boolean queries, and hence with the measurement of performance for simple sets of retrieved documents. However the idea of subsearches (especially broadening a search) naturally allows for an ordering of output, and some approaches to indexing, notably those involving weighting, can only be properly, or at any rate sensibly, interpreted as generating a ranked, i.e. ordered output. (It should be emphasized that this has nothing to do with the representation of Boolean structure by weights, which is merely a matter of notation.) The Cranfield 2 experiments provided an ordered output, and as noted earlier, it became

common in the 1970s, for unweighted as well as weighted searching. It is customary in Smart, for example.

Unfortunately, proper comparative experiments in searching present many problems. As Keen notes in describing EPSILON[103], and discusses more fully in Chapter 8 in this volume, the difficulties of designing and conducting satisfactory experiments in manual searching are very great; and Barraclough *et al.*'s study of users' search behaviour[59], for example, was a very simple observational one. It is also not clear how radically different logics, like boolean and ranking ones, should be compared: indeed Evans[61, 62] asks whether such comparisons are meaningful (see also Chapter 14 in this volume). It should be noted too, that evaluating recall for very different strategies using pooled output may introduce dangerous biases.

The searching tests done between 1968 and 1978 were chiefly concerned on the one hand with search logics, and on the other with searcher behaviour, in both cases in operational environments. The many experiments on query modification by relevance feedback, like those carried out by the Smart Project are, as mentioned earlier, more naturally considered under the heading of automatic indexing, since the role of the user in detailed decision making in query modification is very limited, and numerical calculations of a kind justifying a computer are ordinarily involved in the searching.

At the detailed level of objective, form and result, the tests on manual searching have rather little in common, not surprisingly considering the complexity of the topic. Both systematic comparisons and generalizations about them can therefore only be limited. The tests on search logics included Evans[61, 62] and Miller's[68–70] experiments, and also Katzer's cost-oriented investigation[57]. Some of Aitchison *et al.*'s work[49] is also relevant, as is that of Cleverdon[101]. Studies of searcher behaviour were carried out during the period by Barber, Barraclough and Gray[104] and Barraclough *et al.*[59], Olive *et al.*'s test comparing manual scanning for SDI with automated searching was a 'semi' search test[50]; and service studies like Lancaster, *et al.*[89], Leggate *et al.*[58], and the UKCIS investigation[53, 54] involved examination of search specifications and searching.

The focus of Evans' experiment (see Chapter 14 in this volume) was to compare a range of different term or term group weighting schemes involving ranked output, and also boolean specifications, of Miller's to compare the use of weights (in fact relevance weights) with boolean searches, of Katzer's to compare 'grades' of boolean logic, specifically for cost. Cleverdon compared boolean searching with co-ordination ordering, and Aitchison *et al.* indirectly compared boolean and co-ordination-ordered searching, and also different query formulations (which they called strategies), broad, medium and narrow. Olive *et al.* compared automatic scanning with human for current awareness, Barber *et al.* users and experts as searchers. Barraclough *et al.* simply observed user behaviour in searching online, and the other investigative projects like the UKCIS one just noted features of searches.

The motivation for the comparative tests in the first group was to evaulate the less restrictive, especially weighting, schemes, that of the second group the simpler, less expensive approaches, like automatic searching in Olive *et al.*'s study, having the user rather than an expert search in Barber *et al.*'s test. The common assumption was that the simpler approaches were adequate.

In form the tests followed the general pattern, the main feature of interest

being the large document sets used, for example Olive *et al.*'s of 12 765. Recall and precision evaluation with retrieved sets was usual, even for Miller, who applied a threshold to the ordered weighted output. However Evans and Aitchison *et al.* evaluated over rankings.

The findings for the search logic tests were competitive and even superior performance for the ordering methods. For example, Evans found relative recall for ordering methods at a cutoff of 25 ranged from 45.2 to 49.6 per cent, compared with that for weighted boolean matching of 40.8 per cent, while an alternative method of comparison found the ordering strategies and simple boolean searching similar. Miller had precision of 17 per cent and relative recall of 46 per cent for boolean compared with 15 and 64 per cent for cutoff ordering. Cleverdon had boolean performance for different languages ranging from 27 to 51 per cent precision for 52–30 per cent recall (by external sample), with non-boolean at a middling co-ordination level ranging from 69 to 68 per cent precision for 25–39 per cent recall. It is interesting to note the very low recall levels of Aitchison *et al.*'s boolean searches. It is also of interest that Barraclough *et al.*'s Medusa observations show users most often starting with narrow searches and broadening them, effectively following the co-ordination strategy of Cranfield 2. Aitchison *et al.*'s comparisons of formulations show large performance differences, but these are in part due to variations in document indexing exhaustivity. For example Aitchison *et al.*'s searches ranged from 28–67 per cent precision with 50–11 per cent recall for broad formulations to 50–75 per cent precision with 9–4 per cent recall for narrow, though graph comparisons show rather smaller differences. The findings for the user studies show the competing alternatives very similar, for example performance for users and experts respectively in Barber *et al.*'s test. The authors generally interpret the findings as showing that the various simpler approaches advocated are justified. The more general implication of the tests, taken together, is that different strategies of the same general type produce very similar results, and even that strategies of quite different types may do so. However the tests all support the proposition that it is worth taking some trouble about the search specification: the 'simpler' approaches tested were by no means crude.

The tests so far mentioned dealt with systems which were wholly or essentially manual, i.e. the document indexing might be manual and the request indexing certainly was, even if the searching process was executed mechanically. Thus in systems with automatic scanning of titles or abstracts, like those discussed by the UKCIS workers, the real work was done by the manually constructed profiles. Automatic systems at their most exigent, like those studied by Smart, involve automatic indexing of documents and of requests, represented by initial user need statement texts, or at least substantial automatic modification of given manually-constructed document and request descriptions. The phrase 'automatic indexing', while loosely applicable to automatic scanning, is more properly applied to more extensive automatic processing, which in the 1970s was focused largely on the treatment of requests, compared with the earlier concern with document indexing.

Automatic indexing tests

As noted, the work on automatic indexing and searching of the 1970s was

devoted to much more thorough performance evaluation than that of the previous decade. A significant feature of the generally statistical approaches adopted has been the idea of relative rather than absolute merit, whether in the characterization of individual documents, of a collection, of requests, or of document-query matches. Manual indexing tends to involve an all or nothing approach to indexing and retrieval. Numerical measures of merit can of course be used with a threshold to select items in indexing or searching, but more power, because more discrimination, is involved in the general idea of weighting; and as indicated earlier, a good deal of the theoretical work in information retrieval in this decade has been concerned with the notion of ranking determined by probability.

Evaluation tests on automatic indexing and searching were chiefly devoted to statistical methods, not simply in the absence of non-statistical techniques, but with the support of the theories justifying statistical approaches to indexing and matching. The tests have included ones on individual document index term weighting, though not selection, on vocabulary selection and weighting, on term clustering and document clustering, and on query term selection and weighting. A number of projects have carried out experiments on more than one of these: the Smart Project work in this decade in particular has included tests in all of these areas[4, 63–66, 96–98]. Sparck Jones has been concerned with vocabulary selection and weighting, term clustering[75–77, 79, 80, 93–95], and query weighting, and van Rijsbergen with term clustering, document clustering, and query weighting[81, 105–107].

Automatic methods not using relevance information

To consider work on automatic methods other than those involving relevance information first. There has been no evaluation testing of methods for the direct selection of terms for documents along the lines of Damerau's earlier investigation, though Evans tested indexing by automatic assignment of manual thesaurus terms[86]. Simple weighting by within-document term frequencies has been studied by the Smart Project[96, 97]. More attention has been devoted to the treatment of the collection vocabulary, as in Salton's use of discrimination functions to select and weight vocabulary terms[96, 97], or the use by Salton[96, 97] and Sparck Jones[93, 95] of inverse document frequency weights. A whole range of tests with term clusters, used either to define classes of substitute terms or sets of additional terms was carried out by Vaswani and Cameron[78] and by Sparck Jones[80, 95], and a more restricted test by Cagan[108]. Smart Project tests on term clustering during the decade have been rather restricted ones with modified manual thesauri and 'statistical phrases'[65, 97, 98]. Document clustering has been studied by the Smart workers[63] and by van Rijsbergen[105–107].

The focus, motivation and assumptions of these tests were very much those of the previous decade. The general aim has been to demonstrate the value of statistical selection, weighting and classification techniques for retrieval, mostly by comparison with their absence, but sometimes, as in some Smart tests, by comparison with manual alternatives. More specific concerns have been to evaluate competing statistical methods for providing a given device, for example approaches to term classification in Vaswani and Cameron's and Sparck Jones' experiments, and to term weighting in many

Smart tests. It is worth noticing that the statistical techniques have been increasingly seen as means for improving any natural language input, rather than as tools for totally automatic as opposed to manual indexing and searching. Thus the motivation for weighting experiments has often been to show that simple natural language keyword indexing, regarded as itself having been shown to be competitive with controlled language indexing, can be improved by the application of devices using statistical information. Throughout the assumption has been that statistical techniques pick up or effectively exploit information neglected or inadequately handled by the human indexer or searcher.

The form of these experiments is well illustrated by Smart ones. They were generally characterized by small request and document sets (very often the test collections of earlier projects like Cranfield 2), and in evaluation by the use of recall and precision graphs for their ordered search output. The fact that recall/precision graphs may be obtained by different techniques must be emphasized, as this may explain large apparent differences between projects. Van Rijsbergen has utilized a measure of effectiveness combining recall and precision. Cagan used logical rather than real recall and precision. Evans' test was of a conventional kind using boolean profiles, and rather larger document sets than the studies of fully-automatic methods.

Overall, the results of the different tests have been very similar. The detailed findings show considerable variation in performance for the different options tested in the more extensive comparisons like those made by Vaswani and Cameron, Sparck Jones, and the Smart Project. Unfortunately, the fact that the recall/precision graphs produced were obtained by different techniques means that specific comparisons between projects are impossible, and even comparisons between the relative ranges of performance have to be made with caution. Moreover Sparck Jones and Salton each conducted such large series of tests that it is very difficult to describe them briefly. We may therefore simply note that for vocabulary selection and weighting both Salton and Sparck Jones found performance differences with recall/precision graphs ranging from 5 to 20 per cent, which were usually also improvements over simple term matching graphs. For term classification Vaswani and Cameron, using cutoff ranked output, found classification methods ranging from 16.4 to 21.3 per cent precision with 37.3–49.1 per cent relative recall, compared with 21.4 and 49.1 per cent for keywords alone; the Smart Project's test with classes and phrases show small performance graph differences and improvements; Sparck Jones found variations of as much as several hundred per cent between best and worst classification graphs. Evans found that automatic assignment with a well-organized thesaurus could provide quite competitive performance. Van Rijsbergen for document clustering found small improvements for clusters at the best end of the performance range, representing gains in precision but loss of recall; these were rather better results than those obtained by Smart. A noticeable feature of the tests is that among the better performance options, similar performance may be obtained by a variety of approaches. Collectively the authors concerned interpreted their findings as showing that vocabulary weighting is more effective than selection or clustering. Selection may impact recall (a point concealed in the Smart type of performance representation), and is no better than weighting, and both Smart workers and Sparck Jones have found inverse document frequency

(i.e. collection frequency) weighting of some utility. However neither Vaswani and Cameron nor Sparck Jones, in substantial series of experiments, could obtain real performance improvements with clustered rather than unclustered terms. Cagan found clustering useful, but in a highly eccentric test. In document clustering, precision can be maintained, but recall suffers. Taken together, the tests would imply that simple statistical techniques are as good as more elaborate ones, but even then yield only modest performance improvements. The main inference to be drawn from the tests was that vocabulary distribution properties may be important for retrieval: for example Sparck Jones found clustering rare terms far more useful than clustering common ones. This observation contributed to the work on weighting.

Automatic methods using relevance information

The final group of tests to be considered are those concerned with relevance feedback and weighting. The automatic indexing methods discussed so far are based on information about the occurrences and co-occurrences of terms in *any* documents. The use of more specific information about term occurrences and co-occurrences in *relevant* documents leads to the relevance feedback and relevance weighting schemes which have been especially important in the research work of the decade. The Smart Project's early tests[4, 63, 64], especially those of Ide[109], concentrated on feedback methods for adding terms to, or removing them from, queries; later experiments[65, 66] were concerned with relevance or 'precision' weighting. Sparck Jones has carried out a series of experiments with relevance weights[75–77, 95], as have Harper and van Rijsbergen[81]. In an operational context, Miller[68–70], the UKCIS staff (Barker *et al.*[54, 71] and subsequently Robson and Longman[72, 73]) have studied relevance-controlled query expansion or weighting schemes. Cameron's approach was rather different, clustering documents using relevance information[74].

The character of these experiments has been very like that of other automatic indexing tests. Thus the focus of the tests has been a comparative evaluation of searching with and without relevance information. The context has nearly always been that of natural language indexing, though Miller applied relevance weights to MeSH terms. The motivation has been to demonstrate the value of statistical methods of indexing utilizing relevance information. An additional motivation in laboratory tests like those of Robertson and Sparck Jones[75] or Harper and van Rijsbergen has been to validate a formal theory; while in the service studies like the UKCIS ones, the statistical feedback techniques have been seen as devices assisting the user in reducing the effort of profile preparation.

In form the experiments follow the mainstream pattern with recall and precision evaluation, but with output ordering in the laboratory tests where the operational tests have concentrated on comparisons with boolean performance. It is of interest that in this group of experiments quite large test collections were used, not only in operational tests like Miller's, but in some laboratory tests: thus Cameron used some 12 000 documents (though few requests), and Sparck Jones over 27 000 documents.

The results of these tests are some of the most striking of the decade. There are again large variations in the different tests, for example for different weighting formulae; and there are also, making due allowance for different performance representation methods, quite wide variations between tests. But the findings for relevance weighting in particular show large improvements in performance in the recall/precision graphs for weighted compared with unweighted searching. Considering the specific findings, Smart findings for relevance feedback adding or deleting terms show comparative performance graph differences of 5–15 per cent, Barker *et al.*'s user-modified profiles exploiting relevance information show gains in relative recall of 10–30 per cent for no loss of precision. Robson and Longman's complicated tests for producing profiles via relevance weighting showed automatic profile performance ranging from 25.6 per cent precision and 61.3 per cent relative recall or 36.5 per cent precision and 76.3 per cent relative recall compared with manual results of 31.0 per cent precision and 82.1 per cent recall or 39.9 per cent precision and 87.9 per cent recall, for different categories of profile. The experiments weighting given query term lists show a wide range of differences: Miller's cutoff weighting output gave 15 per cent precision and 64 per cent relative recall compared with unweighted boolean 17 per cent precisions and 46 per cent recall (by my calculation). The Smart Project and Sparck Jones show graph differences ranging from about 5 per cent to several hundred per cent: Sparck Jones' weighting differences ranged from 50 per cent to some hundred per cent, compared with unweighted performance, though these particular findings, like any others, may be partly attributable to the performance representation methods used. Cameron's experiment in this area showed a gain in recall from 44.3 per cent to 60.6 per cent, for a decline in precision of 60.7 per cent to 41.0 per cent.

The predominantly favourable findings have naturally been interpreted as demonstrating the value of the various statistically-based techniques for utilizing relevance information, particularly as in some cases weighting improves on already competitive or good performance. Thus Miller's test shows statistical weighting competitive with standard Medlars boolean searching, and the Smart workers and Sparck Jones claim that their substantial series of tests show that relevance feedback and weighting are useful. In the operational context the UKCIS workers note that the automatic weighting methods do effectively reduce user effort.

Certainly the implication of these tests would appear to be that using relevance information, possibly in the particular, theory-motivated way advocated by Robertson and Sparck Jones and van Rijsbergen[110], is a helpful approach to retrieval.

A small amount of work has been done on non-statistical automatic indexing, for example by O'Connor[82, 83] and (pseudo-automatically) by Atherton[84]. Both of these studies are of interest in concentrating on neglected areas of retrieval, namely of passages and monographs respectively. O'Connor shows simple proximity or rudimentary syntactic strategies effective, Atherton crude significant location utilization procedures. Klingbiel and Rinker's interesting test[85], mentioned earlier, was of machine-aided indexing utilizing elementary parsing and a dictionary; the findings showed performance could compete very successfully with manual indexing. The project is an exception to the general trends of the period, linking recent

document retrieval research with much earlier work and with current research on non-bibliographic databases.

12.6 Conclusion on 1968–1978

Overall, when we look at the evaluation tests of the decade from a substantive point of view, we can see on the one hand a rounding out of the work done in the previous decade, and on the other one possible line of non-conventional performance improvement. The experiments as a whole show simple indexing as good as sophisticated as far as the language is concerned, with the actual treatment of the query more important as a determiner of performance than anything else, and performance in any case very difficult to raise above a broad 50 per cent precision–50 per cent recall level. It appears, more specifically, that the requirement to be met to reach this level, is that of adequate exhaustivity of indexing, primarily of the query (unless the user's only concern is with precision). The relevance weighting techniques studied during the decade may not raise performance above the middling level apparently at best attainable in practice, but they may provide a very helpful and cheap way of raising performance from the lower levels likely to be actually attained in practice. It is however the case that the substantive remarks that can be made about the results obtained between 1968 and 1978 are, like those of the previous decade, very general, and the more novel ones are rather tentative.

Methodologically, the tests taken together show some improvement over those of the previous decade's, but regrettably not enough. A particular contribution has been made by the use of individual test collections by more than one project, and of several collections by individual projects. In the first case the Cranfield 2 data especially have been widely used, but other test collections like Keen's ISILT one and a UKCIS one have also been utilized by more than one project. This does not of course mean that any defects of the data as created are removed, but at least the results of one project can be related to others, and equally, particular results supported by related tests, including those involving different methods of performance representation. The Smart Project was a pioneer of multi-collection tests, and the importance alike of those tests showing different results for different collections and those showing the same result for different collections, cannot be overestimated. Sparck Jones has also used a range of increasingly large test collections in laboratory experiments.

As noted earlier, a good many of the tests of the decade exhibited more careful control of both primary and secondary variables, and a concern with the validity of findings which has led to a wider application of statistical significance tests, as for example by the Smart Project and by Keen. (It must nevertheless be admitted that the basis for applying significance tests to retrieval results is not well established, and it should also be noted that statistically significant performance differences may be too small to be of much operational interest.)

Unfortunately, the tests of 1968–1978 still show many methodological deficiencies. Thus it is to be regretted that in Yates-Mercer's test of relational indexing indexers and searchers were not sufficiently independent; tests like

Schumacher *et al.*'s did not, as noted, distinguish closely related factors like indexing source and quality and index description source and quality, and Svenonius' experiment suffered from similar defects. In evaluation, Cleverdon indulged in some not very well justified performance extrapolation. In a good many tests it is difficult to attribute any significance to the absolute values of relative recall obtained, and indeed, as noted in connection with searching, perhaps to comparative values.

The most welcome feature of the decade has been the increase in test collection size, though reports of operational tests in particular tend not to indicate precise size. There are nevertheless far too many tests using quite small document sets and, much more importantly, small numbers of requests. For example Miller used only 25 requests, Cleverdon has often used less than 20, Katzer used 18, and Cameron 12. It is very doubtful whether the smaller sets produce results which can be regarded as more than suggestive for other contexts. The largest request sets used were UKCIS' 193 and Leggate's 160. It is particularly regrettable that the many Smart Project tests have typically used small collections, in recent years three consisting of some 25 requests and 450 documents each. Several studies, like UKCIS', have unfortunately also used different sizes of request set in individual, closely related experiments.

To complete the detailed discussion, we may note that as in the previous decade, the main body of evaluation tests on retrieval system core factors discussed so far was surrounded by other studies of different types. These have also followed the changing trends of the decade. Non-evaluation studies in the core area include a number, especially in the earlier part of the decade, in the area of automatic indexing, like those of Artandi and Wolf[111], Carroll and Roelloffs[112], Williams[113], Harter[114, 115], and Field[116], all concerned with terms, and of Litofsky[117] and Schiminovich[118] on document clustering. They all claimed some degree of plausibility or merit in the devices studied. Outside the five groups discussed, or at least on a higher and more comprehensive level, have been service oriented tests and studies like those of Rowlands on SDI[88], Hansen on *Chemical Abstracts* costs[90], and Simkins[91] and Pollitt[92] comparing services for particular user communities. Investigations like those of Lancaster, *et al.*[89] and Leggate *et al.*[58], briefly mentioned earlier in connection with searching, really fall into this category. These studies naturally reflect consumer interest in the increasing range of competing services but incidentally, as is shown by Pollitt's study, provide valuable raw data. In the more peripheral areas there have again been many studies of users, and a whole range of bibliometric investigations, for example of citation patterns. There has also been an increasing interest in data base coverage and overlap.

12.7 The outcome of 20 years' testing

What conclusions can be drawn about the state of information retrieval research from such a survey? More specifically, what progress has been made over the last 20 years in obtaining substantively valuable results from methodologically sound experiments?

Overall, the impression must be of how comparatively little the non-negligible amount of work done has told us about the real nature of retrieval systems. Of course, compared with areas like biological research, the number of tests has been extremely small; and a point brought out by the survey is how few really serious tests there have been. In his 1970 review of evaluation tests Cleverdon[119] includes perhaps a couple of dozen tests; and as the present chapter suggests, the number would not be more than doubled 10 years later. One might nevertheless suppose that enough experimental and investigative work had been done to provide some concrete information about retrieval systems. Yet the most striking feature of the test history of the past two decades is its lack of consolidation. It is true that some very broad generalizations have been endorsed by successive tests: for example that performance is pretty middling, or that different languages perform the same; but there has been a real failure at the detailed level to build one test on another. As a result there are no explanations for these generalizations, and hence no means of knowing whether improved systems could be designed.

It is of course unreasonable to expect a high degree of consistency in the conduct of experiments: this would presuppose a framework for system characterization and evaluation which does not exist. Conducting large test programmes in document retrieval is also extremely laborious; it requires resources which are not available to many individual projects. It is nevertheless the case that the lack of solid results must be attributed primarily to poor methodological standards. As the test details presented in this chapter show, there is so little control in individual tests and so much variation in method between tests that interpretations of the results of any one test or of their relationships with those of others must be uncertain.

The general inadequacy of information retrieval tests, but also the practical reasons for it, are best exhibited by looking again at the conditions for a retrieval test. There are the data on one hand, the mechanism on the other. The data consists of the actual documents and the actual user queries and assessments. The mechanism consists of the indexing and searching apparatus. We initially think of the data (D) as given, the mechanism (M) as chosen, i.e. we exploit specific techniques in a specific environment, to obtain a total retrieval system. The minimal system study then consists simply of noting the performance of this system: call this D:M. We then recognize that different mechanism options are available and, selecting some part of the mechanism, say the indexing language, for study as the primary experimental variable, we compare two of its values, say M11 and M12, in a test D:M11/M12. We then consider connections between different parts of the mechanism and proceed to relate the behaviour of variable M1 to that of some other variable M2, say indexing exhaustivity, for a test with the structure D:M11/M12:M21/M22. In this we perhaps regard M2 as the secondary variable. We can naturally extend the test series for any set M1, M2 . . . Mn of mechanism variables we choose to examine.

But of course, from the point of view of understanding retrieval systems in general, D is as important as M. The behaviour of retrieval systems is a function of both D and M. We should therefore consider the constituent variables of D, say types of user, giving us D11/D12: M11/M12: M21/M22, and, further, say different document types, giving us

D11/D12:D21/D22:M11/M12:M21/M22; and we can clearly continue, for as many variables and values of each as we can identify.

In general retrieval system tests have exhibited biases in the way they have approached this set of study possibilities. Much more attention has been paid to the mechanism variables M than the data variables D. The mechanism variables have been made explicit, the data one left implicit: in other words, though test authors have often paid lip service to the possible influence of their data variable values on their results, they have nevertheless tended to characterize the entire system performance in terms of the mechanism variables studied; variable D has been left undifferentiated, while perhaps several values of a single M variable, or a few values of several M variables, have been examined. It has not, moreover, been open to third parties to put different tests together on the grounds that while their data variable values have differed their mechanism variable values have been the same, so amalgamating the tests would permit the effects of data variation to be examined: the mechanism variables have generally not been identically or sufficiently similarly treated.

Some projects, like those of Salton and Sparck Jones, have begun to tackle this problem by working with more than one data set; but it has to be recognized (as the data details of Sparck Jones and Bates[95] make plain) that the characterization and control of data variables in these test series is much less systematic even than that of the mechanism variables. It is moreover generally the case that where the same data have been used by different projects, the treatment of the mechanism variables has been too heterogeneous for it to be possible to combine the test results to obtain information about an extended set of mechanism variable values.

12.8 Methodological and substantive achievements

Thus if we accept that a proper understanding of retrieval systems can be achieved only with the aid of both a well-organized descriptive framework and extensive series of experiments, each bearing on the other, and look now at the evidence of the chapter survey, what methodological and substantive progress has been made in achieving this understanding?

If we compare, say, Montague's test of 1965[19] with Evans' of 1975[61, 62], we can detect some methodological improvements and a substantive development: Montague's test was vitiated by the use of incomparable query sets and incomparable document sets, i.e. the individual experiments in the group could not be compared usefully with one another because the data sets used differed. In many of them the query set used was also very small. Evans used a constant set of queries and documents for a range of comparisons, and a somewhat larger query set than any of Montague's. The substantive development is represented by the shift from document indexing, studied in Montague's test, to query formulation, the focus of Evan's test. At the same time, the difference between the tests is not as large as might be hoped for, in methodological solidity or depth of understanding. While Montague's test is open to criticism in mixing real and synthetic queries, in Evans' the amount of output assessed for relevance per query was somewhat arbitrarily varied. Again, while Montague's test explored a variety of rather arbitrarily related

document indexing options, Evans' investigated a fairly heterogeneous selection of search strategy options.

More generally, while it might be claimed that multi-collection experiments comparing search procedures like those of Sparck Jones[95] represent some methodological and substantive advance compared with Dale and Dale's test of some ten years earlier[21], Atherton's BOOKS test[84] does not represent much advance methodologically over Cranfield 1 (Ref. 1), and is indeed substantively focused on the same system component, document indexing.

Is it nevertheless possible to point to any general methodological and substantive advances?

To take methodology first. It is possible to point to a general advance in the quality of retrieval experiments. Specifically, we can say that experiments now are more likely than they were twenty years ago:

(1) to use real data, for example not requests based on source documents;
(2) to have enough data, for example fifty requests rather than ten, and five thousand documents rather than five hundred;
(3) to introduce more control in relation to data variables, for example by using more than one collection;
(4) to discriminate better among mechanism variables, for example by distinguishing language specificity from indexing specificity;
(5) to utilize more appropriate performance measures, for example by interpreting recall in relation to sets of documents rather than single sought documents;
(6) to conduct tests more carefully, for example by utilizing Latin square designs for assigning tasks to people;
(7) to evaluate findings properly, for example by applying significance tests.

But, as the survey has shown, not all experiments meet these conditions. The effort of conducting proper experiments, which proposals like those for the 'ideal test collection' were intended to reduce, remains very great, so many tests are limited in scope. Again, though tests appear to reflect a growing consensus, for example in the use of recall and precision, many test reports suggest that little attempt has been made to learn from the experience of previous workers. Where such complicated matters as techniques for deriving recall/precision graphs are concerned the lack of rigour is not surprising, but it is still unfortunate.

These defects of current test methodology are nevertheless really only the manifestation of deeper problems about the substantive aspects or retrieval systems. Differences of test design in part reflect genuine differences of test purpose and emphasis. Thus there is no very good reason why an efficiency oriented test to determine indexing speed under different working conditions should have much in common with an effectiveness oriented test to evaluate the utility of not-logic in boolean searching. However there are many features of information retrieval systems which are not sufficiently understood, even at the level of reliable description, let alone analytical modelling: appropriate choices of performance measure are an example.

Looking at the substantive side of retrieval systems, what contributions have twenty years of testing made to system understanding and hence system design? As pointed out earlier, statements here can only be very general ones. Thus it appears that the tests which have been carried out show:

(1) that artificial indexing languages do not perform strikingly better than natural language;
(2) that complex structured descriptions do not perform strikingly better than simple ones;
(3) that the number of searching keys is more important than their individual quality;
(4) that the characterization of queries is more important than that of documents;
(5) that formal properties of the data may be turned to advantage, as in weighting schemes.

But of course, as these statements all refer only to mechanism variables, they can have real meaning only by being related to their environment of data parameters; and the main failure of information retrieval research has been in determining those environment properties significant for system operation and in establishing the relationship between data and mechanism variables. Cleverdon[120] in 1971 maintained that 'it is, in theory, possible to design and operate a system that will achieve a given satisfactory performance, at the least possible cost, in a particular environment'. But he also observes that while it is possible, in any given situation, to design an effective system, 'a problem that is still unsolved is how it is possible to predicate exactly what a situation will be. . . . Designing for the hypothesised, but probably non-existent, "average" user, we may produce systems that satisfy no-one'. (pp. 67–8)

Some advance in this area since 1971 can in fact be detected: a good deal of rather crude evidence about systems has been gathered; and some system models have been proposed which have stood up to initial testing, for example the Robertson[121] and van Rijsbergen[110] probabilistic theories. But it remains the case that our ignorance is large: to take a conspicuous instance, we have virtually no information about the real recall levels of large online search systems, or about real recall for many retrieval schemes investigated by research workers.

12.9 The current state of retrieval system understanding

After an evaluative survey of the retrieval test literature, van de Water *et al.*[122] concluded that the standards and content of tests were slightly higher than those found in a survey carried out five years earlier, but that information science was nowhere near established as a science. This is certainly true; but perhaps this is aiming too high too soon. A more reasonable question is whether retrieval research has any more modest, but nonetheless material, achievements to its credit.

The best way of answering this question is to ask whether there have been any research results which have been applied to operational systems. Even allowing for some delay, one would hope that after five or ten years good research results could have had operational outcomes.

Cleverdon[123] considered this question in 1976. Looking at the historical development of retrieval systems, he asked whether some more conspicuous research projects had contributed, either positively or negatively, to the

operational systems of the mid-1970s. Assuming four groups of system components, relating to input, store, search, and overall organization, he notes that, with respect to the store, the important breakthrough was the National Library of Medicine's use of computers for the preparation of printed indexes; with respect to input, the key factor was the boost given to post-coordination by Taube's company, Documentation Inc.; with respect to searching, the vital development was the growth of online computing; and with respect to overall system organization, the significant contribution was made by computer network technology. So, Cleverdon concludes,

> 'We now have mechanised systems which not only allow the user to do everything which was possible with a card catalogue or printed index, but also give him many additional facilities. We can search in natural language—or a controlled vocabulary if we still cling to the old beliefs. There is the power and flexibility of postcoordinate searching, output can be automatically printed in a number of different forms, and the citations can be in a ranked order of probable interest. There are, of course, some corresponding disadvantages. Many people would consider online searches to be expensive while others find the systems awkward and complex to use. Both these are aspects that can only change for the better.'

According to Cleverdon, the contributions to retrieval system development made by testing have been very limited. He singles out as critical the 1953 Documentation Inc. comparison between uniterm and alphabetical indexes[124], and Swanson's 1962 comparison between simple automatic text searching and conventional manual indexing[125]. Though the results obtained were not properly understood at the time, Cleverdon argues that the common factor explaining the comparative success of the Uniterms in Taube's test and of the text indexing in Swanson's was the use of natural language. Subsequent tests have like Cranfield 2 (Refs. 2, 3) confirmed the value of natural language indexing. For Cleverdon other important tests, in terms both of their individual results and the natural way in which these results could be combined for whole system characterization, were Cranfield 1½ (Ref. 6), which demonstrated the inverse relation between recall and precision, and the sequence of Smart tests exhibiting the value of direct text utilization as a mode of natural language indexing, of matching producing ranked output, and of iterative searching. Cleverdon finds that by the late 1960s, 'it was obvious that we had acquired the knowledge that would enable mechanised systems to be designed that were both effective and economic'. Subsequent developments in computing technology permitted us to take advantage of this knowledge. In Cleverdon's opinion, 'it is clear that no single research investigation made a major contribution to the present position, and that most of the significant advances have come as a result of setting up operational systems, from which developments flowed'. In fact, even the widespread, though by no means exclusive, use of natural language in operational systems may be as attributable to practical factors as to the application of research findings; and the systematic exploitation of the recall/precision relationship, of ranking, and of coherent interactive procedures, do not in fact figure in operational systems. The natural language text available for searching also tends to be limited.

Cleverdon's position, however, is that while he is enthusiastic about the

available online systems, this is not to imply that they are perfect. We are, he suggests, 'at much the same stage with mechanised information retrieval systems as was the automobile at the beginning of this century. All the essential ingredients are now there and working, but much effort is still required before we can have economical, reliable, and widely used models'. My view is that this is rather optimistic: Cleverdon's position can only be justified by a rather restricted engineering attitude to retrieval system research. Current systems can perhaps be better tooled, but they rely so heavily on the human user that the opportunities for a radical redesign of the rest of the system are limited. Changing the relative balance between the user and the rest might allow more interesting design possibilities; and to pursue these effectively we need more information than we have, which can only come from further theory development and testing. After all, while economic arguments may have recommended natural language to operational system managers, it is quite possible they would not have accepted these so readily without the intellectual confirmation provided by the results of experiments like Cranfield 2. It has taken quite a long time for the test results of the 1960s to filter through into practice, and we must therefore expect as slow responses to the experiments which have been carried out in the 1970s, particularly since Cleverdon's review, or which should be carried out in the 1980s.

References

1. CLEVERDON, C. W. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Cranfield (1962)
2. CLEVERDON, C. W., MILLS, J. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems*, 2 Vols, College of Aeronautics, Cranfield (1966)
3. CLEVERDON, C. W. The Cranfield tests on index language devices, *Aslib Proceedings* **19**, 173–194 (1967)
4. SALTON, G. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART), *Journal of the American Society for Information Science* **23**, 75–84 (1972)
5. BOURNE, C. P. Evaluation of indexing systems. In: *Annual Review of Information Science and Technology*, Vol. 1 (Ed. C. A. Cuadra), Interscience, New York (1966)
6. AITCHISON, J. and CLEVERDON, C. W. *Report on a Test of the Index of Metallurgical Literature of Western Reserve University*, College of Aeronautics, Cranfield (1963)
7. SCHULLER, J. A. Experience with indexing and retrieving by UDC and Uniterms, *Journal of Documentation* **12**, 372–389 (1960)
8. CROS, R. C., GARDIN, J. C. and LEVY, F. *L'Automatisation des Recherches Documentaires. Un Modèle Général: Le SYNTOL*, Gauthier Villars, Paris (1964); 2nd edn with new preface (1968)
9. ALTMANN, B. A multiple testing of the natural language storage and retrieval ABC method: preliminary analysis of test results, *American Documentation* **18**, 33–45 (1967)
10. BLAGDEN, J. B. How much noise in a role-free and link-free coordinate indexing system?, *Journal of Documentation* **22**, 203–209 (1966)
11. SHAW, T. N. and ROTHMAN, H. An experiment in indexing by word choosing, *Journal of Documentation* **24**, 159–172 (1968)
12. LANCASTER, F. W. *Evaluation of the MEDLARS Demand Search Service*, National Library of Medicine, Bethesda, Md (1968)
13. LANCASTER, F. W. MEDLARS: a report on the evaluation of its operating efficiency, *American Documentation* **20**, 119–142 (1969)
14. CASE WESTERN RESERVE UNIVERSITY. *An Inquiry into Testing of Information Retrieval Systems*, 3 Vols, Comparative Systems Laboratory, Centre for Documentation and Communication Research, Case Western Reserve University (1968)

15. SARACEVIC, T. Selected results from an inquiry into testing of information retrieval systems, *Journal of the American Society for Information Science* **22**, 126–139 (1971)
16. SINNETT, J. D. *An Evaluation of Links and Roles Used in Information Retrieval*, Air Force Materials Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio (1964)
17. HERNER, S., LANCASTER, F. W. and JOHANNINGSMEIER, W. F. A case study in the application of Cranfield system evaluation techniques, *Journal of Chemical Documentation* **5**, 92–95 (1965)
18. COHEN, S. M., LAUER, C. M. and SCHWARTZ, B. An evaluation of links and roles as retrieval tools, *Journal of Chemical Documentation* **5**, 118–121 (1965)
19. MONTAGUE, B. A. Testing, comparison and evaluation of recall, relevance and cost of coordinate indexing with links and roles, *American Documentation* **6**, 201–208 (1965)
20. VAN OOT, J. G. *et al.* Links and roles in coordinate indexing and searching: an economic study of their use and an evaluation of their effect on relevance and recall, *Journal of Chemical Documentation* **6**, 95–101 (1966)
21. DALE, A. G. and DALE, N. Some clumping experiments for associative document retrieval, *American Documentation* **16**, 5–9 (1965)
22. O'CONNOR, J. Automatic subject recognition in scientific papers: an empirical study, *Journal of the ACM* **12**, 490–515 (1965)
23. DAMERAU, F. J. An experiment in automatic indexing, *American Documentation* **16**, 283–289 (1965)
24. BORKO, H. Research in computer based classification systems. In: *Classification Research: Proceedings of the Second International Study Conference* (Ed. P. Atherton), Munksgaard, Copenhagen (1965)
25. TAGUE, J. M. An evaluation of statistical association measures, *Proceedings of the American Documentation Institute* **3**, 391–397 (1966)
26. MELTON, J. S. Automatic language processing for information retrieval: some questions, *Proceedings of the American Documentation Institute* **3**, 255–263 (1966)
27. DENNIS, S. F. The design and testing of a fully-automated indexing-searching system for documents consisting of expository text. In: *Information Retrieval—A Critical View* (Ed. G. Schechter), Thompson, Washington D.C. (1967)
28. STONE, D. C. and RUBINOFF, M. Statistical generation of a technical vocabulary, *American Documentation* **19**, 411–412 (1968)
29. SALTON, G. The evaluation of automatic retrieval procedures—selected test results using the SMART System, *American Documentation* **16**, 209–222 (1965)
30. SALTON, G. *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York (1968)
31. SALTON, G. and LESK, M. E. Computer evaluation of indexing and text processing, *Journal of the ACM* **15**, 8–36 (1968)
32. LESK, M. E. Performance of automatic information systems, *Information Storage and Retrieval* **4**, 201–218 (1968)
33. GIULIANO, V. E. and JONES, P. E. *Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems*, Arthur D. Little Inc., Cambridge, Mass. (1966)
34. JONES, P. E. *et al. Application of Statistical Association Techniques to the NASA Document Collection*, Arthur D. Little Inc., Cambridge, Mass. (1968)
35. STEVENS, M. E., HEILPRIN, L. and GIULIANO, V. E. (Eds.) *Statistical Association Methods for Mechanised Documentation*, National Bureau of Standards, Washington D.C. (1965)
36. STEVENS, M. E. *Automatic Indexing: A State of the Art Report*, Monograph 91, National Bureau of Standards, Washington D.C. (1965)
37. SPENCER, C. C. Subject searching with *Science Citation Index:* preparation of a drug bibliography using *Chemical Abstracts, Index Medicus*, and *Science Citation Index* 1961 and 1964, *American Documentation* **18**, 87–96 (1967)
38. NEWILL, V. A. and GOFFMAN, W. 'Searching titles by man, machine and chance, *Proceedings of the American Documentation Institute* **1**, 421–423 (1964)
39. GOTLIEB, C. C. and KUMAR, S. Semantic clustering of index terms, *Journal of the ACM* **15**, 493–513 (1968)
40. LESK, M. E. Word-word associations in document retrieval systems, *American Documentation* **20**, 27–38 (1969)
41. WILLIAMS, J. H. and PERRIENS, M. P. *Automatic Full Text Indexing and Searching System*, IBM Federal Systems Division, Gaithersburg, Md (1968)
42. TAGUE, J. M. Matching of question and answer terminology in an education research file, *American Documentation* **16**, 26–32 (1965)

43. HOUSTON, N. and WALL, E. The distribution of term usage in manipulative indexes, *American Documentation* **15**, 105–114 (1964)
44. HEALD, J. H. *The Making of TEST: Thesaurus of Scientific and Engineering Terms*, Department of Defence, Washington D.C. (1967)
45. MENZEL, H. Information needs and uses in science and technology. In: *Annual Review of Information Science and Technology*, Vol. 1 (Ed. C. A. Cuadra), Interscience, New York (1966)
46. HERNER, S. and HERNER, M. Information needs and uses in science and technology. In: *Annual Review of Information Science and Technology*, Vol. 2 (Ed. C. A. Cuadra), Interscience, New York (1967)
47. KING, D. W. Design and evaluation of information systems. In: *Annual Review of Information Science and Technology*, Vol. 3 (Ed. C. A. Cuadra), Encyclopedia Britannica, Chicago (1968)
48. LANCASTER, F. W. and MILLS, J. Testing indexes and index language devices, *American Documentation* **15**, 4–13 (1964)
49. AITCHISON, T. M. *et al. Comparative Evaluation of Indexing Languages, Part II: Results*, Report R70/2, INSPEC, Institution of Electrical Engineers, London (1970)
50. OLIVE, G., TERRY, J. E. and DATTA, S. Studies to compare retrieval using titles with that using index terms, *Journal of Documentation* **29**, 169–191 (1973)
51. KEEN, E. M. and DIGGER, J. A. *Report of an Information Science Index Languages Test*, 2 Vols, College of Librarianship Wales, Aberystwyth (1972)
52. KEEN, E. M. The Aberystwyth index languages test, *Journal of Documentation* **29**, 1–35 (1973)
53. BARKER, F. H., VEAL, D. C. and WYATT, B. K. Comparative efficiency of searching titles, abstracts and index terms in a free-text data base, *Journal of Documentation* **28**, 22–36 (1972)
54. BARKER, F. H., VEAL, D. C. and WYATT, B. K. *Retrieval Experiments Based on Chemical Abstracts Condensates*, Research Report No. 2, UKCIS, The University, Nottingham (1974)
55. SCHUMACHER, H. H., MARCH, J. F. and SCHEFFLER, F. L. *The Use of Selected Portions of Technical Documents as Sources of Index Terms and Effect on Input Costs and Retrieval Effectiveness*, University of Dayton Research Institute, Dayton, Ohio (1973)
56. SVENONIUS, E. An experiment in index term frequency, *Journal of the American Society for Information Science* **23**, 109–121 (1972)
57. KATZER, J. The cost performance of an on-line, free-text bibliographic retrieval system, *Information Storage and Retrieval* **9**, 321–329 (1973)
58. LEGGATE, P. *et al. The BA Previews Project: The Development and Evaluation of a Mechanised SDI Service for Biologists*, Experimental Information Unit, University of Oxford (1973)
59. BARRACLOUGH, E. D. *et al. The Medusa Current Awareness Experiment*, Computing Laboratory, University of Newcastle upon Tyne (1975)
60. KEEN, E. M. *On the Performance of Nine Printed Subject Index Entry Types. A Selective Report of EPSILON*, College of Librarianship Wales, Aberystwyth (1978)
61. EVANS, L. *Search Strategy Variations in SDI Profiles*, Report R75/21, INSPEC, Institution of Electrical Engineers, London (1975)
62. EVANS, L. *Methods of Ranking SDI and IR Outputs*, Report R75/23, INSPEC, Institution of Electrical Engineers, London (1975)
63. SALTON, G. (Ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, N.J. (1971)
64. SALTON, G. Recent studies in automatic text analysis and document retrieval, *Journal of the ACM* **20**, 258–278 (1973)
65. SALTON, G., WONG, A. and YU, C. T. Automatic indexing using term discrimination and term precision measurements, *Information Processing and Management* **12**, 43–51 (1976)
66. YU, C. T. and SALTON, G. Precision weighting—an effective automatic indexing method, *Journal of the ACM* **23**, 76–88 (1976)
67. SALTON, G. and WALDSTEIN, R. K. Term relevance weights in on-line information retrieval, *Information Processing and Management* **14**, 29–35 (1978)
68. MILLER, W. L. *The Evaluation of Large Information Retrieval Systems with Application to MEDLARS*, Ph.D. Thesis, University of Newcastle (1970)
69. MILLER, W. L. The efficiency of MEDLARS titles for retrieval, *Journal of the American Society for Information Science* **22**, 318–321 (1971)

70. MILLER, W. L. A probabilistic search strategy for MEDLARS, *Journal of Documentation* **27**, 254–266 (1971)
71. BARKER, F. H., VEAL, D. C. and WYATT, B. K. Towards automatic profile construction, *Journal of Documentation* **28**, 44–55 (1972)
72. ROBSON, A. and LONGMAN, J. S. *Automatic Aids to Profile Construction*, 2 Vols. UKCIS, The University, Nottingham (1975)
73. ROBSON, A. and LONGMAN, J. S. Automatic aids to profile construction, *Journal of the American Society for Information Science* **27**, 213–223 (1976)
74. CAMERON, J. S. *Automatic Document Pseudo-Classification and Retrieval by Word Frequency Techniques*, Computer and Information Science Research Center, Ohio State University (1972)
75. ROBERTSON, S. E. and SPARCK JONES, K. Relevance weighting of search terms, *Journal of the American Society for Information Science* **27**, 129–146 (1976)
76. SPARCK JONES, K. Experiments in relevance weighting of search terms, *Information Processing and Management* **15**, 133–144 (1979)
77. SPARCK JONES, K. Search term relevance weighting given little relevance information, *Journal of Documentation* **35**, 30–48 (1979)
78. VASWANI, P. K. T. and CAMERON, J. B. *The National Physical Laboratory Experiments in Statistical Word Associations and Their Use in Document Indexing and Retrieval*, Publication 42, Division of Computer Science, National Physical Laboratory, Teddington (1970)
79. SPARCK JONES, K. and JACKSON, D. M. The use of automatically-obtained keyword classifications for information retrieval, *Information Storage and Retrieval* **5**, 175–201 (1970)
80. SPARCK JONES, K. *Automatic Keyword Classification for Information Retrieval*, Butterworths, London (1971)
81. HARPER, D. J. and VAN RIJSBERGEN, C. J. An evaluation of feedback in document retrieval using cooccurrence data, *Journal of Documentation* **34**, 189–216 (1978)
82. O'CONNOR, J. Text searching retrieval of answer sentences and other answer passages, *Journal of the American Society for Information Science* **24**, 445–460 (1973)
83. O'CONNOR, J. Retrieval of answer-sentences and answer-figures from papers by text searching, *Information Processing and Management* **11**, 155–164 (1975)
84. ATHERTON, P. *Books are for Use: Final Report of the Subject Access Project*, School of Information Studies, Syracuse University (1978)
85. KLINGBIEL, P. H. and RINKER, C. C. Evaluation of machine-aided indexing, *Information Processing and Management* **12**, 351–366 (1976)
86. EVANS, L. *Evaluation of the ISPRA Automatic Indexing Programs SLC-II*, 2 Vols, INSPEC, Institution of Electrical Engineers, London (1978)
87. CLEVERDON, C. W. *A Comparative Evaluation of Searching by Controlled Language and Natural Language in an Experimental NASA Data Base*, Cranfield Institute of Technology (1977)
88. ROWLANDS, D. G. The Unilever Research SDI system, *Information Storage and Retrieval* **6**, 53–71 (1970)
89. LANCASTER, F. W., RAPPORT, R. L. and KIFFIN PENRY, J. Evaluating the effectiveness of an on-line natural language retrieval system, *Information Storage and Retrieval* **8**, 223–245 (1972)
90. HANSEN, I. B. CA Condensates as a retrospective search tool, *Information Storage and Retrieval* **9**, 201–205 (1973)
91. SIMKINS, M. A. A comparison of data bases for retrieving references to the literature on drugs, *Information Processing and Management* **13**, 141–153 (1977)
92. POLLITT, A. S. *The Cancerline Evaluation Project*, British Library Research and Development Report 5377, School of Medicine, University of Leeds (1977)
93. SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation* **28**, 11–21 (1972)
94. SPARCK JONES, K. Does indexing exhaustivity matter?, *Journal of the American Society for Information Science* **24**, 313–316 (1973)
95. SPARCK JONES, K. and BATES, R. G. *Research on Automatic Indexing 1974–1976*, 2 Vols, British Library Research and Development Report 5464, Computer Laboratory, University of Cambridge (1977)
96. SALTON, G. and YANG, C. S. On the specification of term values in automatic indexing, *Journal of Documentation* **29**, 351–372 (1973)
97. SALTON, G. *A Theory of Indexing*, Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia (1975)

98. SALTON, G., YANG, C. S. and YU, C. T. A theory of term importance in automatic text analysis, *Journal of the American Society for Information Science* **26**, 33–44 (1975)
99. JAHODA, G. and STURSA, M. L. A comparison of a keyword from title index with a single access point per document alphabetic subject index, *American Documentation* **20**, 377–380 (1969)
100. CLEVERDON, C. W. *An Investigation into a Suitable Mechanised Information Retrieval System at the Defence Operational Analysis Establishment*, Cranfield Institute of Technology (1970)
101. CLEVERDON, C. W. and HARDING, P. *Report of an Investigation into a Mechanised Information Retrieval Service in a Specialised Subject Area*, Cranfield Institute of Technology (1970)
102. YATES-MERCER, P. A. Relational indexing applied to selective dissemination of information, *Journal of Documentation* **32**, 182–197 (1976)
103. KEEN, E. M. 'On the processing of printed subject index entries during searching, *Journal of Documentation* **33**, 266–276 (1977)
104. BARBER, A. S., BARRACLOUGH, E. D. and GRAY, W. A. On-line information retrieval as a scientists tool, *Information Storage and Retrieval* **9**, 429–440 (1973)
105. JARDINE, N. and VAN RIJSBERGEN, C. J. The use of hierarchic clustering in information retrieval, *Information Storage and Retrieval* **7**, 217–240 (1971)
106. VAN RIJSBERGEN, C. J. Further experiments with hierarchic clustering in information retrieval, *Information Storage and Retrieval* **10**, 251–257 (1974)
107. VAN RIJSBERGEN, C. J. and CROFT, W. B. Document clustering: an evaluation of some experiments with the Cranfield 1400 collection, *Information Processing and Management* **11**, 171–182 (1975)
108. CAGAN, C. A highly associative document retrieval system, *Journal of the American Society for Information Science* **21**, 330–337 (1970)
109. IDE, E. New experiments in relevance feedback. In: *The SMART Retrieval System: Experiments in Automatic Document Processing* (Ed. G. Salton), Prentice-Hall, Englewood Cliffs, N.J. (1971)
110. VAN RIJSBERGEN, C. J. A theoretical basis for the use of cooccurrence data in information retrieval, *Journal of Documentation* **33**, 106–119 (1977)
111. ARTANDI, S. and WOLF, E. H. The effectiveness of automatically-generated weights and links in mechanical indexing, *American Documentation* **20**, 198–202 (1969)
112. CARROLL, J. M. and ROELOFFS, R. Computer selection of keywords using word frequency analysis, *American Documentation* **20**, 227–233 (1969)
113. WILLIAMS, J. H. Functions of a man-machine interactive information retrieval system, *Journal of the American Society for Information Science* **22**, 311–317 (1971)
114. HARTER, S. P. A probabilistic approach to automatic keyword indexing. Part 1: On the distribution of specialty words in a technical literature, *Journal of the American Society for Information Science* **26**, 197–206 (1975)
115. HARTER, S. P. A probabilistic approach to automatic keyword indexing. Part 2: an algorithm for probabilistic indexing, *Journal of the American Society for Information Science* **26**, 280–289 (1975)
116. FIELD, B. J. Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing, *Journal of Documentation* **31**, 246–265 (1975)
117. LITOFSKY, B. *Utility of Automatic Classification Systems for Information Storage and Retrieval*, Ph.D Thesis, University of Pennsylvania (1969)
118. SCHIMINOVICH, S. Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm, *Information Storage and Retrieval* **6**, 417–435 (1971)
119. CLEVERDON, C. W. Evaluation tests of information retrieval systems, *Journal of Documentation* **26**, 55–67 (1970)
120. CLEVERDON, C. W. Design and evaluation of information systems. In: *Annual Review of Information Science and Technology*, Vol. 6 (Ed. C. A. Cuadra), Encyclopedia Britannica, Chicago (1971)
121. ROBERTSON, S. E. *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems*, Ph.D. Thesis, University of London (1976)
122. VAN DE WATER, N. *et al.* Research in information science: an assessment, *Information Processing and Management* **12**, 117–123 (1976)
123. CLEVERDON, C. W. A survey of the development of information retrieval systems. In: *EURIM 2: A European Conference on the Application of Research in Information Services and Libraries (1976)*, Proceedings (Ed. W. E. Batten), Aslib, London (1977)

124. GULL, C. D. Seven years of work on the organisation of materials in the special library, *American Documentation* **7**, 320–329 (1956)
125. SWANSON, D. R. Searching natural language text by computer, *Science, N.Y.* **132**, 1099–1104 (1960)