13

The Cranfield tests

Karen Sparck Jones

As the broad survey of Chapter 12 suggests, Cleverdon's Cranfield 1 and 2 projects have played an extremely important part in the development of retrieval system experiment. The purpose of this chapter is to discuss their contribution in more detail. I shall consider the objectives, form and results of the two projects in turn, and the reactions to them, concluding with a discussion of their influence on retrieval testing as a whole. Cranfield 1 and 2 were major tests from every point of view—in their scope, their conduct, and their impact—and as such must be examined. I shall not, however, attempt an exhaustive review of all the Cranfield-related literature, but rather focus on the two projects in their contexts, referring as appropriate only to the more important comments on them. I shall not attempt either to discuss all of Cleverdon's own papers, but will refer only to those bearing on the Cranfield 1 and 2 tests as experiments. In Cleverdon's own work Cranfield 1 was preceded by the collaboration with Thorne, and was paralleled by the English Electric test and the joint Cranfield–Western Reserve University experiment, referred to in the previous chapter as Cranfield 1½. Cranfield 2 was followed by the DOAE and Precision Engineering tests, and more recently by the NASA study. Together these tests constitute a major contribution to information retrieval experiment, showing those features which may be said to mark the Cranfield tradition in the methods and concerns of retrieval experiment, but also reflecting the major change in retrieval systems brought about by automation.

In this chapter, the account of Cranfield 1 is followed by a discussion of related work and of the reactions to the test, and by an assessment of these reactions; Cranfield 2 is then treated in the same way. This approach has been adopted to make the position of the tests in their historical contexts clear, but it must be emphasized that at Cranfield itself the second test immediately followed the first, and that the intellectual continuity between the two was strong: Cranfield 2 was very much addressed to questions which arose during Cranfield 1 as well as to dealing with early criticisms of Cranfield 1. It should also be noted that there have been changes in standard terminology since the early Cranfield work: in particular in relation to performance measurement relevance is now generally called precision, and what was originally referred to as efficiency would now, given the way the tests in practice limited performance measurement, be interpreted as referring primarily to effectiveness.

13.1 Cranfield 1

The ancestor of Cranfield 1 was a pilot experiment in the use of Uniterms for indexing aeronautical documents which was carried out at Cranfield in 1953. The data for the test consisted of 40 questions based on source documents which were searched on 200 documents; performance was measured as the success rate in retrieving the query sources, which was 83 per cent (strictly 82.5 per cent). The test formed part of a group described by Thorne[1] in 1955 involving a comparison between Uniterms and the UDC as used at the Royal Aircraft Establishment, and a subsequent comparison between the NLL specialized aeronautical indexing language developed at National Aeronautical Research Institute, Amsterdam and the UDC. Performance for the 40 questions searched with UDC and RAE was 50 per cent, compared with 85 per cent for Uniterms, while performance for the NLL language using another set of questions was 88 per cent as opposed to 80 per cent for the UDC on a subset of these.

This group of tests, though defective for example in using different question and document sets, already exhibits some key features of the Cranfield tradition: methodologically the use of source document queries, substantively an interest in subject-oriented indexing. Thorne's account of the set of tests also emphasizes costs, which have been one of Cleverdon's persistent concerns.

The part played by Cleverdon's experience in these tests in promoting Cranfield 1 is clearly indicated in the Preface to the first volume of the Cranfield 1 Report[2]. Thus Cleverdon notes that while the specialized NLL scheme developed by 1953 would apparently complement the more general UDC satisfactorily, something cheaper was desirable. Uniterms were a possibility, but the results obtained from a second test following the one just described were not too promising. However the work on the NLL scheme itself meant that test procedures had to be devised and, as Cleverdon says,

> 'by this time [1954] I had become convinced that the only way to obtain a valid comparison between systems would be to control conditions in such a way that there was an economic basis for the comparison. At the Conference of the Aslib Aeronautical Group in 1955 I read a paper in which, for the first time, the necessity for controlled experiments was put forward.' (p. ii)

This conference was, moreover, partly sponsored by the Classification Research Group, which was actively discussing novel classification techniques; and new approaches to indexing and retrieval were also being put forward in the United States. 'It was clear', Cleverdon continues,

> 'that claims were being made by proponents which, while possibly correct, could not be considered proven by results; just as clearly many of the arguments being used by opponents of the systems were equally unproven or trivial. It seemed desirable that a serious investigation should be made

> so that opposing claims could be evaluated, and by this time we had definite views as to how such an investigation could be carried out.' (p. ii)

In a paper to a Special Libraries Association meeting in 1955 Cleverdon argued that independent evaluation of rival claims was needed, and this led to National Science Foundation funding of the Aslib Cranfield Project in 1957.

The project was designed to investigate 'the comparative efficiency of four indexing systems', involving the indexing of:

> '18,000 research reports and periodical articles in the general field of aeronautical engineering, with half of the documents dealing with the specialized subject of high speed aerodynamics.' (p. 1)

The general objectives of the test were determined by the problems presented by the growth of the scientific literature, the increasing complexity of research work, and corresponding proposals for new retrieval systems and implementations. As the project proposal stated,

> 'in all the controversies that have raged during the past fifty years on the basic points of a book catalogue or card catalogue, with an alphabetical subject arrangement or a classified arrangement, it is interesting to note that no attempt has been made to carry out any controlled tests that would enable one to make statements based on fact rather than voice theoretical opinions.' (p. 4)

The proposal quotes a remark by the Editor of *American Documentation* in 1955 to the effect that we must regard

> '*documentation systems as useful devices, the benefits of which must be determined, not by polemics, but by the intelligent measurement of such benefits in relation to needs and costs.*' *(p. 5)*

As the proposal noted,

> 'the complication in attempting to evaluate the comparative efficiency of any two retrieval systems is due to the number of various factors which have to be considered. These can be summarised as follows:
> (1) The documents which are to be indexed.
> (2) The system of indexing.
> (3) The indexer's subject knowledge of the documents being indexed.
> (4) The indexer's familiarity with the indexing system.
> (5) The size of the index.
> (6) The type of question which is to be put to the index.
> (7) The equipment to be used in recording or retrieving data.
> (8) The overall efficiency, which is made up of:
> (a) The time cost in preparing the index.
> (b) The time cost in locating required information.
> (c) The cost of equipment used.
> (d) The probability of producing the required answer.
> (e) The absence of irrelevant answers ('noise').
> (f) The number of searches made.' (p. 5)

The essential features of the project were thus that it was a comparative one, focusing on indexing languages, and seeking to identify and control other system factors bearing on the study variable.

The indexing systems chosen for test were

'(a) The Universal Decimal Classification.
(b) Alphabetical subject catalogue.
(c) A facetted classification scheme.
(d) The Uniterm system of coordinate indexing.' (p. 8)

According to the Report,

> 'the basis for this selection was that the schemes differed as fundamentally as is possible and represented the principal types of retrieval systems which have any significance in the present state of the art.' (p. 8)

The UDC was chosen as the most widely used system of the 'enumerative' type, illustrating tree-of-knowledge classification and the use of a decimal notation. Facet classification specifically lacks these three features. The alphabetical index is deliberately anti-classificatory and relatively uncontrolled in language, while Uniterms are equally uncontrolled, but allow all permutations and combinations of individual terms to define subjects. The facet and Uniterm languages both represented novel approaches to indexing, the former being particularly carefully prepared.

With respect to the other main test variables, the document subject area, aerodynamics, was determined by convenience, while the set of documents used was chosen ad hoc, to allow variability in detailed document topics, types, and sources. The indexers were chosen to have various types of experience and familiarity with the systems being tested. (It may be noted that 'the project imposed a severe mental strain on the indexers' (p. 2).) Equipment, i.e. the physical form of indexes, was not studied as a test variable. The size of the document set for the indexes was primarily based on the wish to ensure that retrieval was not too obvious, without being wasteful; more specifically, with 60 permutations of the three major variables, obtained as explained below, and taking 100 documents as a convenient number for each permutation, there were 3 'subprogrammes' of 6000 documents for the total of 18 000. In fact, since half the documents were in a specialized subject area, the collection was deemed representative of a much larger set.

Overall,

> 'the assumption on which the investigation [was] based [was] that the only valid way to measure the efficiency of any system of indexing is by basing measurements on economic costs and in this there are always three matters to be considered, these being:
>
> (a) The cost of indexing.
> (b) The cost of preparing the physical index.
> (c) The cost of searching.' (p. 18)

So on the indexing side, since equipment questions were deliberately excluded, the major variables being investigated were the system, the indexer, and the indexing time, i.e. 4 systems × 3 indexers × 5 times = 60 permutations.

The procedures adopted in the indexing are described in detail, in relation to such questions as determining the time spent on identifying subject content versus that spent on assigning notations or headings. There were many problems to be dealt with, and

> 'we came up against many difficulties that had not been envisaged' (p. 27),

but as the Report notes,

> 'by the time that the first 6000 documents had been dealt with, the indexers had established a tempo which was maintained to the end of the whole of the indexing.' (p. 25)

Even so, it is clear that indexing to time was a major effort for the indexers. A point of particular interest is that experience with the indexing led to developments of the schemes used. Indeed at one stage

> 'it was clear that with the facet classification we were getting into a complete mess and a technical revision of the schedules was necessary, together with some procedural alterations.' (p. 28)

Again,

> 'as indexing continued, the original rules for the alphabetical subject indexing appeared to be too restrictive and we had to make some slight modifications.' (p. 29)

The UDC presented some, though fewer, problems, while Uniterms presented least difficulties. The main feature of the whole indexing operation was the idea of main and subsidiary assignments, i.e. each indexer would process a batch of documents using scheme A, say, as base, and then supply appropriate descriptions for B, C and D; he would then process another batch using B as base, and so on. In fact this approach to avoiding global biasses towards one language had to be simplified, though it was maintained in essentials. As Cleverdon says,

> 'it is difficult to know exactly how to assess our experience in setting up these systems, because few people appear to have attempted to review objectively their own experience.' (p. 29)

These general statements about the four indexing languages are amplified in detailed discussions of the particular problems encountered with the individual languages. For example, for the UDC these included the treatment of synthesis, the interpretation of ambiguous headings for specific technical concepts, the allocation of concepts where several separate placements offered, and also the provision of an alphabetical index. For the alphabetical index an initial problem was that there was no existing index which could simply be utilized, and the index was built up using rules developed during the project defining the character of main headings and subheadings; the particular problems encountered were those of preferred order for different forms of simple or complex concept modification, of direct versus inverted entry, and of multiple entries and cross references. The facetted scheme, which was based on thoroughly argued principles, was specifically designed for the test. The problems encountered were those of adhering to the preferred citation order, or of maintaining constant word forms for terms in the chain index, and of entry order in the catalogue. With the Uniterm scheme, neatly described as based on 'literary warrant', the main problem was in splitting terms into elementary units. Taken together, these details illustrate both how familiar problems of indexing recur in any environment, and how they can perhaps be handled for medium-sized systems.

The Report also gives statistical details about the indexing which are both of interest in themselves and relevant to the indexing performance observed. For example for the indexing subprogramme for the last 6000 documents there were 2350 different notational elements for UDC, 2684 main alphabetical headings, 1686 facet notational elements, and 3174 Uniterm terms. For individual documents, variations in the numbers of terms for the different indexers were noted, and the average numbers declined with indexing time. For example, average postings for journal articles were 7.6 for UDC for 16 min as opposed to 2.3 for 2 min, 4.7 and 2.2 respectively for alphabetical subject headings, 1.7 and 1.2 for facetted, and 11.0 and 6.0 for Uniterms (Table 6).

To check the main indexing, alternative, independent ('supplementary') index descriptions were supplied by people outside the project. These suggested that the project indexing was well enough organized, but also tended to show a lack of agreement in the indexing done by different people for the same document.

In the course of the indexing a good many low-level administrative decisions had to be taken, for example about index formats; and while the project was, as is noted in the Report, primarily concerned with intellectual questions, the account of the various low-level devices and procedures is very useful in illustrating the large amount of nitty gritty involved in any substantial test, and the need for care in these aspects of testing. These processes incidentally provided useful information about the time taken in clerical operations, and also emphasize the scale of this early project: 100 000 cards were punched for the Uniterm indexing, for example.

In introducing the account of the searching representing the test proper of the four indexing systems, in Volume 2 of the Report[3], Cleverdon comments on the need for

> 'a method which would enable an assessment to be made of the effects of the variables which had been built into the indexing.' (p. 7)

It was decided that though each group of 100 documents had distinctive characteristics, and so should be studied individually, testing over a 6000 document subprogramme set, and more specifically that of the final subprogramme reflecting the indexer's established indexing experience, would be sufficient. As Cleverdon notes, there was very little guidance available on the conduct of tests, the ASTIA Uniterm test of 1953 being both inadequately reported and evidently unsound methodologically in lacking controls, especially with respect to relevance decisions. The organization of the search testing was thus strongly motivated by a desire to avoid getting

> 'bogged down in the quagmire of arguments concerning relevancy' (p. 8),

and also by the need for statistically valid results. The first requirement was met by using questions based on source documents and searching for these, the latter by complicated (and, it must be said, very opaquely described) sampling procedures. A set of 1200 questions was therefore obtained by soliciting questions based on sample document lists from those involved in the independent indexing. The critical problem was then that of devising appropriate approaches to searching. The searching itself was carried on until either the source document was retrieved, or no further plausible search strategies (called 'programmes' in the Report) could be devised; and controls were placed on the distribution of the searches among the staff. It was regarded as important that the searching should simulate real life searches. The searching was therefore done by devising specific search strategies, searching for these and noting the numbers of references retrieved, and devising new strategies as these seemed to be required. (The wanted documents were identified for the searches only by their numbers.) However as Cleverdon notes, some artificiality was introduced by the fact that the documents retrieved for each strategy were not read, since given numbers were being looked for, so successive strategies were not produced by reference to document content as well as question content. Further, it turned out that it was very difficult to define single searches, i.e. strategies, and to determine stopping points. These problems were dealt with by defining a strategy as any permutation of a set of elements, and a new strategy as dropping one element; a search was terminated when more than one element had to be dropped. But of course in practice such rules could not be applied rigorously, and the discussion of examples in the Report is very instructive.

The searches were in fact done in three rounds, the first, on 400 questions, as a pretest, leading to the definitions of strategy and stopping point applied with a second set of 400. Failure analysis of the results for this second round suggested that sufficient information about searcher variation had been collected, and the third round of 400 was therefore done in a way designed to eliminate searcher and hence associated strategy variation. The searching was mainly done by project staff, but other categories of searcher were also involved in the second round. (These searching procedures are discussed in more detail in Keen's chapter in this volume, which also comments on the confidence which can be placed in the results obtained with them.)

The actual results reported for all the searches show a success rate in retrieving the wanted documents of 75.6 per cent for UDC, 81.5 per cent for alphabetical, 73.8 per cent for facet, and 82.0 per cent for Uniterm (Table 3.1). These relationships are broadly paralleled in more detailed breakdowns for searcher category, indexing time, indexer, and subject. The results were subjected to a variety of statistical tests, e.g. to determine the standard error rate and the effects of various factors like language, tested by correlation of question and wanted document title words. Further, an exhaustive analysis of failures in the second and third rounds was also made, covering defects in the questions, in the indexing, in searching, and in the languages: for example whether questions were too detailed or not easily understood, the indexing descriptions omitted important concepts or were plain careless, searches were misconceived or unsystematic, and indexes defective in placing ambiguities and gaps in the schedules. Overall (Table 5.4), question failures accounted for 7.7 per cent of the failures, indexing for 60 per cent, searching for 17 per cent and the systems for 6 per cent, with UDC especially defective in indexing, alphabetical and facet in searching, and Uniterm in questions.

The main test was supplemented by special studies, one being concerned with relevance. As Cleverdon says,

> 'a first task was to find out exactly what was being measured, exactly what was implied when it was said that Uniterm, for instance, had an efficiency of 85% [sic].' (p. 51)

It was argued that this meant that searches were retrieving X per cent of all documents at least as relevant as the source. But against this it could be maintained that the relation of question and source was unnaturally close. To test the interpretation of efficiency searches were made for documents independently supplied as a bibliography for some 41 questions. Source documents were excluded and the bibliography items were assigned to three grades of relevance: as useful as the source, somewhat useful, and not in fact useful. Searches for the new relevant documents showed success rates for the highly relevant of 74 per cent, 75 per cent, 60 per cent and 75 per cent respectively for UDC, alphabetical, facet and Uniterm. Thus efficiency is reduced compared with the main test. This suggested that the operating conditions for searching were important, and specifically that the success rate in the main test would have been lower if less strategy relaxation had been permitted: as suggested for Uniterms, an inverse relationship of recall and precision ('relevance') ratios applies. Thus, the Report claims,

> 'there is the possibility of quoting three different performance figures, those with Uniterm as an example being:
> 65% when all concepts are required,
> 85% when one less concept than the required is accepted,
> 97% when a single Uniterm is accepted.' (p. 55)

Further,

> 'the only practical method of showing these various points is by plotting them against relevance [i.e. precision] ratio, that is the percentage of retrieved documents which have an agreed relevance.' (p. 55)
>
> Then
>
> 'as the recall figure (i.e. the percentage of potentially relevant documents in the collection) rises, the relevance ratio (i.e. the percentage of relevant documents amongst the total of those retrieved) must fall and conversely as the recall figure drops, so the relevance ratio will improve.' (p. 55)

A study of precision for 79 questions, assessing a sample of retrieved documents and extrapolating, showed precision ranging from 7 per cent for UDC via 7.5 per cent for facet and 12 per cent for Uniterms to 12.5 per cent for alphabetical, for highly relevant documents. However checks suggested that quite different figures could be derived, and, more importantly, that searching beyond the point of retrieving the source document might well retrieve more relevant documents, and so improve precision. But, as Cleverdon points out,

> 'this somewhat tortuous analysis serves to emphasise nothing more than the extreme danger of placing too much credence on any of the figures which are not otherwise corroborated.' (p. 58)

He nevertheless concludes that the claim made for efficiency levels being the same for all relevant documents as for source documents is probably true with the proviso that extra searching may be needed, though he acknowledges that the main test results could be taken as indicating that the test questions were slanted towards the source documents.

In commenting on the test results as a whole Cleverdon notes that, when allowance is made for standard error, the systems could differ by as little as 3.2 per cent, or as much as 13.7 per cent. However, taking the results at their face value, and considering those for the final subprogramme, Uniterm appears 3.8 per cent (actually 3.9 per cent) superior to alphabetical, the latter 5.3 per cent superior to the UDC, and UDC 3.5 per cent superior to facet. The four languages also had some individual advantages and disadvantages, particular points of interest being

> 'the great value and importance of the alphabetical index for the [UDC] schedules' (p. 90),

the fact that the alphabetical system

> 'was far more effective than had been expected' (p. 91),

and that

> 'Uniterm, as a descriptor language, can be given a high rating on many counts. It achieved the best overall figures in the test, . . . it appears to have as good a relevance figure as any other system, and . . . it did not compare unfavourably in the recall of non-source relevant documents.' (p. 92)

With respect to times, longer times raised the success rate from 72.9 to 84.3 per cent over all systems. The results did not show marked differences between indexers or subjects, nor improvements due to learning in indexing, except for Uniterm, or to learning in searching, except for UDC. In discussing the design of search programmes Cleverdon says that

> 'we were not prepared to spend the length of time in physical searching which some organisations appear willing to do' (p. 87);

further, if results are to be produced quickly,

> 'the formulation of [search] programmes must be a reasonably straightforward matter, and this was the position with the project searches.' (p. 88)

This is emphasized by the fact that most failures were due to indexing. Finally, Cleverdon maintains that the main and other tests together

> 'have shown that the general working level of I.R. systems appears to be in the general area of 60%–90% recall and 10%–25% of relevance, . . . This is a considerable distance away from the oft-made assertion that systems *are operating in the general area of [both high recall and high precision].*' (p. 89)

Further,

> 'it can now be said that the inverse relationship between recall and relevance has been conclusively shown, and it should now be possible to design and operate systems that will satisfy, in the most economic way, stated requirements. There will be situations where the emphasis must be on the highest possible recall level, and the resulting penalty of the low

> relevance figure will be accepted. In other cases recall is less important, and greater emphasis will be placed on improved relevance.' (p. 90)

Finally, in considering recall and precision, it is

> 'necessary to consider the environment in which [a] system is operated and here the most important factor is the type of question which will be put to it.' (p. 100)

Working back to indexing this bears on questions of the exhaustivity of indexing, its specificity, the provision for syntax, and weighting. Discussing these, Cleverdon maintains that

> 'it remains true that given the same concept indexing, any two descriptor languages will have the same information content, and therefore the same potentiality for retrieval.' (p. 104)

In other words,

> 'it is not the alternatives of classified or alphabetical arrangement, of post-co-ordinate or pre-co-ordinate indexing (much less the alternatives of manual or mechanical searching) which make any real difference in performance but the power of the descriptor language, allied to the standard of the indexing. The "power" of a descriptor language is in its ability to eliminate irrelevant references, and in addition to a hospitality for specific indexing, there are at least two other devices which can be used, namely "syntactic indexing" and "weighted indexing".' (p. 105)

Parallel tests

Cranfield 1 was paralleled by two investigations of existing systems designed to throw light on the extent to which its results were influenced by the artificialities of the test design.

One study was of the facetted system set up for the English Electric library at Whetstone (see Cleverdon[3], Chapter 7). This test again involved searching for source documents on which questions were based, by both Cranfield project staff and, for a subset of the queries, by English Electric staff; the success rates for each, 77.4 per cent and 73.5 per cent, and the reasons for failures, paralleled those of Cranfield 1, and the same major problems of preferred order for the chain index were encountered. The conclusion was that the test methods developed at Cranfield were applicable in other environments, and perhaps that the results obtained in the tests represent the level of performance to be expected.

The second test was the joint Cranfield–WRU (Cranfield 1½) test of the WRU metallurgical index[4], intended primarily as a study of testing techniques rather than as an evaluation of the index, at that time incomplete. In the test the WRU system, regarded at the time of the test as one of the most sophisticated novel approaches to indexing, was compared with facet indexing for 114 questions, again based on source documents, searched over 950 documents. Since evaluation simply by searching for source documents (though in this case without the wanted document numbers being known to the searchers) had been criticized, the test included an exhaustive assessment of other documents for relevance and the calculation of recall and precision for these. The assessments allowed two grades of relevance, for documents as relevant as the source, and for less relevant documents. The results, for both grades and including the source documents, were 75.8 per cent recall for the WRU index, with 17.7 per cent precision, compared with recall of 69.5 per cent and precision of 33.7 per cent for the Cranfield facet index. (Removing the source documents from the calculations reduces recall to 70.6 per cent and 59.1 per cent respectively, with 13.0 per cent and 24.0 per cent precision (my figures using Ref. 4 pp. 12–15 and Appendix 3C; there appear to be some discrepancies in the various figures in Ref. 4).) A detailed analysis of the failures showed the searching was responsible for most, namely 67.1 per cent, with indexing 18.4 per cent. In considering the results. Aitchison and Cleverdon say that

> 'before the test started, we were convinced that W.R.U. would be able to achieve one of three results.
> (a) obtain a high recall figure;
> (b) obtain a high relevance [i.e. precision] figure;
> (c) obtain a recall figure and a relevance figure which would both be somewhat higher than that achieved with the Cranfield facet index.
> The high level of exhaustivity of the indexing and the complex semantic factoring in the index language gave them the ability to achieve (a); the specific index language, with the added controls would allow them to achieve (b); the combination of these factors could bring about (c).' (p. 47)

The poor WRU precision figure was then explained by two factors:

> 'the main factor in W.R.U. failures to retrieve relevant documents was the relatively poor standard of many of their search programmes,' (p. 48)

while an investigation of non-relevant documents showed that

> 'the high level of exhaustive indexing was partly to blame' (p. 48)

They comment on various features of the test emphasizing the inverse recall/precision relationship, and note that the test influenced the work on Cranfield 2, then under way, in sharpening the idea of index language device.

Cranfield 1 summarized

The main summary account of Cranfield 1 was the Lancaster and Mills[5] account of 1964, which also discussed the English Electric and Cranfield 1½ tests. The account emphasizes the need for the study of indexing itself rather than the manipulation of its products in searches, and comments on the critical role of the Aslib Cranfield project in this. In the Lancaster and Mills' view the most significant results were related to recall, showing its inverse relation to precision and comparable performance for the languages studied (including facets when implemented rather differently from the main test); for indexing times, showing a short time is good enough; for indexers, showing technical knowledge of the subject is not necessary; and for failures, showing inevitable human error to be important. Lancaster and Mills accept Cleverdon's conclusion that the 'artificial' questions did not invalidate the results, and themselves conclude that they were not affected by other factors like the stopwatch indexing. In Lancaster and Mills' view the English

Electric test supported the Cranfield results in relation to success rates and reasons for failures, while the WRU test was notable first for searching without reference to wanted document numbers, and second, for the use of both recall and precision; the results for recall for the facet system were like those of the main test, while the WRU failures were attributable chiefly to poor searching.

According to Lancaster and Mills, the Aslib Cranfield Project as a whole served to establish

> 'reasonably reliable figures . . . for the degree of recall and relevance [i.e. precision] likely to be achieved by a good index employing fairly exhaustive indexing' (p. 8),

i.e. performance in terms of operating efficiency. Further, the broad implication is that

> 'as one factor amongst several determining the operating efficiency of indexes, the indexing system, per se, is of less importance than has been assumed generally.' (p. 9)

However in their view

> 'the first Cranfield investigation was perhaps less important for the light it shed on four actual indexes than for its contribution to the development of techniques for the testing of information retrieval systems and for exposing the basic parameters in the operation of an indexing system. . . . What was, to begin with, a rather blunt tool is being developed into a sharp analytical instrument.' (p. 9)

Lancaster and Mills conclude that the inverse relationship between recall and precision is inescapable, and that the recall/precision character of a system is determined by indexing exhaustivity and specificity, determining recall and precision respectively. Moreover, since search relaxation can improve recall,

> 'we are left with the specificity of the index as its key characteristic. Hospitality to specific indexing, by which we mean the ability to describe precisely the concepts chosen as retrieval tags, is the most important single criterion of an index language' (p. 11);

or

> 'to summarise, for every index language, depending upon its hospitality for specific indexing, there is a maximum relevance ratio which cannot be exceeded. However it is always possible to improve the recall ratio along the fixed performance curve to its maximum by means of variations in search programmes. But the ability to do this is dependent upon the exhaustivity of the indexing.' (p. 11)

13.2 Criticisms of Cranfield 1

The aims and character of the Aslib–Cranfield Project meant that the Cranfield 1 results attracted much attention. Those with indexing language positions to defend tended to attack the results without offering any concrete evidence to support their own views. The reviews of the final Report focusing on Cranfield 1 as an experiment and raising methodological questions are thus of more interest. Some of the points made were sound, but others were not. The more conspicuous critiques are therefore presented here as they were originally published, for the historical record, and are then briefly assessed.

The main critique of Cranfield 1 came from Swanson[6], who wrote:

> 'the significance of these data, in my opinion, should be fully understood by every student of indexing and classification. Such understanding can be reached only after discriminating study, however, for the experimental design of the project, particularly in its early stages, provides a rather unsteady foundation for the superstructure of conclusions subsequently erected. . . . The early Cranfield results have by now been extensively cited, widely quoted out of context, and usually misinterpreted.' (p. 1)

Swanson's particular concern, given

> 'the snowballing tendency to cite [the] results out of context of the experimental conditions' (p. 2),

is with the implications of the project's experimental design for its 'findings'. Thus he says that

> 'the *design itself* seems to have guaranteed many of the results that were found, so that the evidence which supports such results is questionable. The experimental design in fact was such that certain phenomena no doubt would have been detected whether they existed in real systems or not. The "phenomena" to which I refer are the following widely claimed, widely quoted (and, I think, widely accepted) "findings" of the Cranfield project' (p. 2),

namely that:

(1) indexing times over 4 minutes give no real improvement in performance;
(2) a high quality of indexing is obtainable from non-technical indexers;
(3) reliable figures for recall and precision have been obtained;
(4) systems operate at recall 70–90 per cent and precision 8–20 per cent;
(5) maximum recall has effectively been reached;
(6) a 1 per cent improvement in precision costs a 3 per cent loss in recall;
(7) no indexing finesses can substantially improve recall without a loss of precision;
(8) low precision in the WRU index is due to bad searching, though WRU perfectionism pays no dividends;
(9) there is an inverse relationship between recall and precision;
(10) all four indexing methods give similar performance.

Swanson argues that these statements, whether or not they are true, are not established by the test, but are products of its design. He singles out in particular the use of source documents as probably accounting for all of 1–8, the lack of control on relevance, and the influence of human memory on indexing and searching. In Swanson's view, the artificiality of the questions derived from specific documents is less important as bearing on the questions themselves than for its implications for evaluation due to the close relationship between the question searched and the document sought:

> 'in a real situation, a source document generally does not exist. In information-retrieval experiments, *if specific documents are used as sources of questions, any meaningful tests of retrieval-system effectiveness must be made on new documents which can be presumed free of any unusual or direct influence on the wording or nature of the question.*' (p. 6)

Swanson maintains that the source documents should have been excluded since there might be a difference in retrievability between source and non-source documents. He argues that the 'bias' of the test is likely to have been exacerbated by the very close verbal relation between source document titles and questions, and argues from a sample he investigated that simply using a machine-based match between questions and titles would have given recall of 85 per cent. Thus if the different index language descriptions covered the titles well, a high and similar level of recall would be reached whatever the properties of the languages. Swanson moreover argues that a similar bias existed in the Cranfield 1½ test, which also used source documents. Thus the conditions of the test in both Cranfield 1 and Cranfield 1½ could not be expected to distinguish the languages tested adequately, even if they are genuinely distinguishable in performance.

As a corollary, Swanson argues, the results for indexing times are hardly surprising, and also the fact that maximum recall was approximated is not surprising either, i.e. if the real match is in fact focused on the document title, more sophisticated or extensive indexing is unlikely to be useful.

With respect to indexer/searcher memory, Swanson contends that while indexing memory did not influence searching, the fact that the same people were involved means that possible memory effects cannot be excluded as influencing the results. The lack of technical knowledge of the indexers might well not have been important given the question/document link. Finally, Swanson's view is that the heavily stressed result that the four languages performed the same is of little value when recall is considered in relation to precision; and he comments on the fact that included in the test data were figures about the average number of documents retrieved which show that, assuming only one of the retrieved is relevant, as is required, performance is very different for the four languages, Uniterms being much superior to UDC, with alphabetical and facet in between. Swanson maintains that the figures for recall cannot be taken at their face value.

Swanson further points out that the supplementary Cranfield experiment designed to test recall of non-source documents was defective in using relevance assessments made by referring other documents to the source document, and he notes that, given the different search procedure also used, in the absence of precision figures, comparison for recall with the main test results is very dubious.

Swanson also argues that the Cranfield 1½ results do not support the Cranfield 1 findings in any clear way, and in particular show different performance for source and non-source documents. He nevertheless agrees that

'so far as *average behaviour* of information-retrieval systems is concerned

> there is no doubt a very definite *tendency*, and an obvious one, for recall to decline as relevance [i.e. precision] improves and vice versa' (p. 17),

at least under the conditions that

> 'as the average number of index terms per document is increased the recall ability of the system will also increase (inevitably), but relevance, averaged over many search questions, will *tend* to decline',

and

> 'as a *search* is *broadened* (e.g. by moving upward in a classification hierarchy), then, of course, recall will improve (inevitably); and there is a *tendency* for relevance to decline' (p. 17);

and he concludes by saying that his criticisms have been mainly

> 'directed at the inaccurate interpretations and generalisations of the Cranfield data. The value of the project as a whole has been unquestionably great'. (p. 18)

Other reviewers agreed for example in condemning the use of source documents, and made additional criticisms. For instance Mote[7] commented on design failures like the fact that there was interference between the four systems in indexing, on the lack of realism represented by the absence of user/searcher feedback, and on defects in the presentation of the results such as multiple entries for individual queries under different sources of failure. Some of these points were also made by Richmond[8], who comments on the consequences of the primary/subsidiary indexing strategy for comparability of the systems, namely that strictly only the 4000–5000 primarily indexed documents for each system matter. She argues, too, that as indexing times were averaged they are relatively useless. However Richmond's main attack is on the extremely poor presentation of the detailed figures; she points out that

> 'so few of the tables are comparable' (p. 308)

and that

> 'so many of the factors were not equalised . . . that one wonders how valid the results are' (p. 209),

and provides many examples of the consequent difficulties of interpretation. She also notes that the main test and subsidiary relevance test results do not match up as the Report text suggests, as the supplementary results in fact show differential loss of performance. Richmond notes that the general conclusions are not dogmatic, the four systems performing somewhat similarly and better than expected, with Uniterm most efficient and facet least: this difference is attributable according to Cleverdon to the depth of indexing allowed by the different languages, or, according to Richmond, to the effects of timing. Richmond concludes that the house is built on rock, rather than sand, but it is not well built: the test is important, but future tests need to be more careful and much better reported.

In reviewing Cranfield 1½ Sharp[9] roundly condemns the source document approach and comments that as the environment becomes more natural, i.e.

in involving real relevance assessment, the initial poor WRU performance improves, indeed until it becomes superior to that for facet: performance figures are reversed as the test becomes more realistic.

There is no doubt that such reviewers point to failures in the related Cranfield 1 and 1½ tests. However it is clear from the comments that it was recognized that it was only through the experience of major tests that significant progress could be made in experimental design and system understanding. The reviewers all emphasize the importance of Cranfield 1 in particular: for example Mote says

> 'this project represents the most serious attempt yet made to derive a basis for the comparison of indexing systems.' (p. 81)

At the same time, the way forward was indicated by Richmond's call for more care and Sharp's condemnation of source documents: as Sharp said,

> 'the source-document principle should be dropped and future tests carried out taking into account *all* relevant documents retrieved.' (p. 174)

As indicated at the beginning of the chapter, these criticisms are reproduced here to illustrate contemporary methodologically-motivated reactions to Cranfield 1. However some of the attacks on the test were fundamentally mistaken, the most obvious example being Taube's on the 'pseudo-mathematics' of its treatment of relevance[10], which confounded relevance assessment and precision measure. Another example is Mote's misconceived comment on the lack of control of indexing depth in questions.

More importantly, points which were not obviously wrong varied in status as criticisms of the test. Some criticisms disregarded the stated test objectives. Thus Mote's view that the system operations were not realistic is hardly a criticism when it is directed at the constraints required by experimental control. Other critics of the test suggest that particular factors could have influenced the test without demonstrating that they did: an example is Richmond's remark about primary and subsidiary indexing. Such criticisms though suggestive must be regarded as speculative. Yet other criticisms have substance, but not in a narrow sense. These mostly concern the use of source document questions. It has never been shown that source document questions do not either look or behave like 'regular' questions. Thus while Swanson suggests that it is possible that the lack of any real difference between the indexing languages can be attributed to the use of source documents, it does not follow that it must be so attributed. On the contrary, the fact that in many subsequent tests of different kinds indexing languages have tended to perform the same suggests that the Cranfield results were correctly attributed to the language variable. But though the use of source documents may not be grounds for straightforwardly criticizing the test, whether the source documents could have affected the results is a serious question about the test. This was clearly accepted at Cranfield, and a different procedure was adopted for Cranfield 2.

The real limitation of Cranfield 1 was its failure to measure precision, though this was recognized in time for the supplementary test, and Cranfield 1½ and later Cranfield 2 tests were designed to measure precision along with recall.

Reviewing the criticisms of Cranfield 1 now it is evident that though some of the points made stemmed from a failure to appreciate the character of the test as a controlled experiment, others were sound and were explicitly or implicitly seen to be so and hence were catered for in the planning and conduct of Cranfield 2. Even so, the critics agreed on the significance of Cranfield 1 as a retrieval experiment: though its results were not readily accepted, its status as, in Michael Keen's words, a 'pioneering and relevant' test was recognized not only subsequently, but at the time.

The influence of the Aslib Cranfield Project work in the first half of the 1960s can be seen both in specific tests like Herner and Co.'s Bureau of Ships investigation[11], and, more broadly, in the application of particular lessons to be learnt from it in retrieval system testing in general. From the system point of view it suggested that the indexing language might be less important, and other factors more important, than had been supposed, while from the methodological point of view it stimulated more careful design, in terms both of control and realism. For measuring system performance it did much to promote the use of recall and precision. That such lessons were learnt from the Cranfield research is clear from discussions of system testing like Kyle's[12]: she explicitly asks 'What have we learnt?' and 'Where do we go from here?', and seeks to provide some answers. Some of the Case Western Reserve University research[13] was also a direct response to the Cranfield work, as Rees indicates[14, 15]. More generally, Cranfield 1 and 1½ led to a great deal of discussion of retrieval systems and their testing, illustrated by Cleverdon's argument with Swanson about the Cranfield hypotheses[16], and by the debate at the FID/CR Conference[17] in 1964. The wider influence of the Cranfield experiments on system evaluation at a time when this was developing, especially in the context of system automation, was therefore considerable.

13.3 Cranfield 2

However the major impact of Cranfield 1 and its associated experiments was on Cranfield 2, which was specifically designed as a development of Cranfield 1: this is clear from Cleverdon and Mills' account[18] of the philosophy underlying Cranfield 2. Cleverdon, Mills and Keen's view in the first volume of the Cranfield 2 Report[19] was that while Cranfield 1 and 1½ were of general value in demolishing preconceptions about indexing languages, in showing that operational systems could be readily evaluated, in providing considerable data, and in encouraging discussions of systems and their evaluation, they also led to specific hypotheses which were taken as the basis for the new study. These were seven of Swanson's, namely that 4 min for indexing is enough, that technical knowledge is not required, that systems operate at 70–90 per cent recall and 8–20 per cent precision, that there is an optimum level of exhaustivity, that there is an inverse relation between recall and precision, that raising precision 1 per cent lowers recall 3 per cent, and that the most significant Cranfield 1 result was that the four languages performed the same, plux six more: these were that the most important factors to be measured in system evaluation are recall and precision; that the physical form of the store has no effect on performance so measured; that for the same concept indexing different languages will perform much the same;

that the more complex a language is in terms of recall and precision devices the greater its range of performance; and that maximum recall depends on indexing exhaustivity, precision on language specificity. The Cranfield 2 test was based on the more general propositions from this set not limited to specific documents and questions, namely Swanson's fourth, fifth and seventh, and the first one and last three of those just given. Cranfield 2 thus focused on an analysis of the behaviour of recall and precision devices, and further, to ensure control, on an analysis of these devices in laboratory experiments. The Report authors argue robustly for the emphasis on recall and precision as important to users and difficult to measure; for the resolution, following Vickery, of indexing languages into their component devices so the contribution of these to language performance can be assessed; and

> 'to make advances in knowledge regarding index languages',

for

> 'a laboratory-type situation, where, freed from the contamination of operational variables, the performance of index languages could be studied in isolation.' (p. 8)

Thus putting these points more fully, the authors summarise the actual test objective as follows:

> 'we started from the belief that all index languages are amalgams of different kinds of *devices*. Such devices fall into the two groups of those which are intended to improve the recall ratio and those which are intended to improve the precision ratio. . . . The purpose of the test was to investigate the effect which each of these devices, alone or in any possible combination, would have on recall and precision.' (p. 17)

Further,

> 'to enable this to be done, it was essential that it should be possible to hold everything constant except the one variable being investigated.' (p. 17)

The critical factors in the test design were therefore the method of providing questions, the method of providing relevance judgements, and the method of providing index descriptions of documents; and what is most significant about the test design is that the methods of obtaining relevance information designed to provide a firm foundation for recall and precision performance figures effectively determined other properties of the test data. Thus as the authors note, while relevance assessments of output for precision calculation can be both reliably and readily obtained, adequate recall figures require exhaustive document assessment, leading to the use of a relatively small collection. Their view however is that the WRU test had shown that a small document set could provide sufficient data for analysis. The test therefore used 1400 documents, along with 279 requests, providing a larger question sample than previous tests. It should be noted that the composition of the document set was determined by the method of obtaining the questions.

The project aim in obtaining the questions and assessments was that these should be as realistic as possible, though, as the Report authors point out, no actual set of operational questions was available. The approach adopted was therefore to ask the authors of research papers to characterize the problem to which the paper was addressed as a question, adding supplementary questions to the base one if appropriate, and to indicate which citations were relevant. The document collection was then made up of the pooled references of all the question source papers, and further assessments of non-cited papers were made by the question providers, using the output of a crude screening, done by students on titles, of the whole collection, and also some of the output of a bibliographical coupling run based on the source papers.

The authors emphasize that the test design was not perfect, and allow that while the test was concerned with the ability of different indexing systems to retrieve judged relevant documents, the vagueness of the notion of relevance itself could have some hidden influence on the test results. The Report describes the procedures for obtaining questions and assessments in detail: one main and one secondary subject area were chosen for the collection; the questions (and documents) represented a wide range of author types, etc., and the question texts were annotated for more and less important and additional terms; assessments were made using four grades of relevance and one non-relevant. The account of this stage of the test in Volume 1 of the Report is a salutary reminder of the great difficulty and labour of collecting raw test data, and great difficulty of obtaining good data.

Altogether the question and assessment sets were plausibly heterogeneous, though it is difficult to know if there were specific biasses in them; the Report describes the various aspects of the sets in considerable detail, with particular emphasis on the status of the questions. In specific response to the criticisms of the Cranfield 1 procedures it is argued that the connection of source document and question is much less narrow than in Cranfield 1, and that the searching results are in any case not affected by the source documents as these were removed from the collection before searching for each query.

The treatment of indexing in the test reflected the desire to study devices in a controlled way: thus documents were initially indexed 'conceptually', and the common conceptual description was then taken as input to indexing by different languages. In this test, unlike the previous ones, these languages were constructed to embody combinations of devices, and were not simply off-the-shelf. The language devices were characterized as different ways of modifying a simple list of single terms to promote precision or recall. Precision-promoting devices include co-ordination, weighting, links and roles, recall devices, synonym confounding, word form variant confounding, generic hierarchical linkage, and non-generic hierarchic linkage. Bibliographic coupling, statistical associations, and superimposed coding are also regarded as precision devices though only the first of these, bibliographic coupling, was tested, in fact outside the main project. As the Report authors say,

> 'we have tried to distinguish the basic device itself, as a method of class definition, from the different ways in which it might be implemented in different index languages. The latter may be regarded as different amalgams of the various devices, with further differences resulting from the various methods of file organisation.' (p. 47)

The strategy adopted was therefore to take as the base indexing simple natural language, with the document description carefully controlled for exhaustivity and specificity. The controls in fact implied 'maximum' exhaustivity and specificity, reflecting the language of the document. Variations in exhaustivity could be obtained by utilizing importance weights attached to the keywords of the base description, while sufficient specificity was achieved by allowing multi-keyword strings and phrases. Thus the basic description provided both 'concepts' representing interfixed (linked) keywords and 'themes' representing higher-level linked concepts, and also weights for the individual keywords. In other words the description embodied several precision devices, though not roles, which were found to be inapplicable (pp. 56–7), while recall devices were left for application at search time; the precision devices had clearly to be derived from the document itself, but they could be abandoned for study purposes. The initial indexing therefore supplied four languages, single terms, concepts, themes, and weighted keywords, which could of course be combined; recall devices could be naturally applied either to keywords or concepts; for the first these were represented by synonym confounding, word-form combining, both of these together, and three grades of hierarchical reduction, giving a total of eight different languages. The provision of the various types of word classification embodying the recall devices is described in considerable detail. In addition, since these languages were all based on the initial natural language, a conventional controlled language index based on the EJC Thesaurus was used. The title and abstract tests, concurrently being studied by Salton, were regarded as representing other languages. Altogether, as Chapter 1 of Volume 2 of the Report shows, the various types and combinations of device applied to the three starting points of single terms, concepts, and controlled terms gave eight languages for the first, 15 for the second, and six for the third, i.e. 29 languages in total, to be tested.

An interesting point about the test, made by Michael Keen (personal communication), is that the original test design described in Cleverdon and Mills assumed that the initial concept indexing would be so done as to allow 100 per cent recall; however the procedure for checking on this was not followed, suggesting that the idea that system imperfections could be ruled out by experimental procedures was tacitly accepted as unrealistic and was replaced by a principle that care should be taken in indexing, though perfection could not be attained.

The actual conduct of the searching presented many problems, given the many options to be tested, the absence of convenient computer facilities, and the requirement that the physical form of the index should not interfere with its use. The description of the heroic clerical procedures involved, centring on the delightfully-named 'beehive' cabinet, makes interesting reading, and it is worth noticing that even with a computer, the organization of the range of searches involved in Cranfield 2 would be non-trivial.

The first volume of the Report shows the basis for the Cranfield 2 test, i.e. the type of indexing and index language hypotheses involved, and the approaches adopted to providing the test vehicle. The relationship of the primary test variables to others is summarized in Volume 2 (Ref. 19), as a preliminary to the discussion of the results. Thus the Report authors distinguish four classes of retrieval system factor—environmental, including subject field and collection size; software, namely indexing, with respect to exhaustivity, language, with respect to specificity, and searching; operational, including time, personnel, etc., and also performance; and hardware, for example output form. The laboratory mode of testing imposed strict controls on the environmental and operational factors, and for the stated object of the test, hardware aspects could be ignored. However the controls imposed on the environmental and operational factors would naturally play their part in determining performance and hence must be taken into account in assessing the test.

> 'In the artificial environment created for the test it was found that a limited set of changes [to secondary variables] could be investigated; these included several sets of questions picked by different criteria, relevance judgements made in four different grades, collections of three different sizes and tests in two related but different subject fields.' (p. 6)

The search strategies tested were all levels of 'blind' co-ordination, and 6 types of selection/combination of the query terms representing more 'intelligent' search rules appropriate to the basic type of language.

Overall, the range of variable value combinations tested was very large, and the Report authors rightly comment on the

> 'volume, variety and complexity of the tests (p. 16).'

Thus for a particular query and document set there are search results for the different co-ordination levels and other search rules applied to a range of language descriptions representing particular combinations of recall and precision devices, for different initial indexing exhaustivity levels, and taking account of several relevance grades.

Unfortunately, the considerable clerical and intellectual effort involved in doing the tests, combined with a methodological interest in question sets with specific properties, meant that most tests were not carried out with the full 279 questions and 1400 documents, or even with the largest subset of 221 questions and the 1400 documents. In fact, once the investigators had convinced themselves that the smaller sets gave results comparable with the large ones, many of the tests were done with 42 questions and 200 documents. A consequence of the various selections was that tests were done not only with collections having different numbers of requests or documents, but also with collections of different generality, i.e. relevance density.

The scope of the experiments made the details of performance measurement very tricky, and these are discussed at length in the Report. The problems involved are both the higher-level ones of the choice of measure, and the lower-level ones of the application of individual measures to particular data, with special reference to averaging. At the higher level the project was naturally inclined to use recall and precision; at the lower, the particular problem tackled by the project was that presented by averaging searches conducted at different co-ordination levels, representing one case of the general problem of dealing with search output not supplied as simple retrieved document sets. An additional problem for the manually-conducted Cranfield 2 was that the sheer effort of calculation implied by some performance representation methods could not be undertaken. However after trials comparing performance for a pair of languages given by different methods, it was concluded that the simplest direct averaging, totalling documents retrieved across co-ordination levels and then deriving recall and precision, was adequate, and this method was used for the great mass of detailed figures and graphs presented in Volume 2. However since relying only on one form of measure might be dangerous, fallout figures were worked for the main runs; and an alternative representation of recall and precision using ranking rather than levels, the so-called 'document output cutoff' method, was also supplied for the main run outputs.

The discussion of these extremely difficult issues in the Report is important both in showing the attention paid to the question by the project, and in emphasizing their intractibility for any project.

It is impossible here to do more than refer briefly to the great mass of individual results presented in Volume 2: in providing this detail for reader study the Cranfield 2 Report is much superior to that for Cranfield 1. It is sufficient to note that the main results fall into 9 groups: the first group (4.1) gives performance for the 221 × 1400 and 42 × 200 collections for several single term languages, for the authors supporting their view that the smaller collection could justifiably be used for most of the experiments; the second (4.2) compares all the recall devices for single terms for the 42 questions and 200 documents, showing some loss of performance with the most gross term reduction; group 4.3 tests concepts and themes, for small query sets but 1400 documents, showing not much difference in performance; group 4.4 examines exhaustivity levels with single terms, for both large and small collections, again showing not much variation in performance; group 4.5 studies search rules for the single term languages, for small query sets but 1400 documents, suggesting some superiority in the more stringent strategies; the concept languages compared for the 42 × 200 collection in 4.7 show large performance variations, and this is also true of the controlled languages compared in 4.8 for this collection; abstracts and titles regarded as indexing languages are compared in 4.9, again for the 42 × 200 collection, showing abstracts inferior. Section 4.6 covers a secondary variable comparison on the different relevance grades, for the single term languages.

To obtain an overview of the co-ordination level results some comparisons are made of performance at specific co-ordination levels: for example for the 42 × 200 collection at co-ordination level 3 (Figure 6.10T), the various single term languages range from recall 66.7 per cent with precision 14.8 per cent for the simplest language to 82.3 per cent and 7.4 per cent for the most 'condensed' hierarchical one. For co-ordination 2 for the concept languages (Figure 6.12T), deemed comparable with level 3 for the simple terms, performance ranges from recall 84.8 per cent with precision 6.1 per cent for the most condensed to recall 14.1 per cent with precision 50.9 per cent for the given basic indexing language, while for controlled indexing at level 2 (Figure 6.14T) performance ranges from 68.7 per cent with 12.6 per cent precision for the basic to 94.4 per cent and 5.1 per cent recall and precision for the most condensed descriptions. The picture is of low recall and high precision for concepts, higher recall and lower precision for single terms, and highest recall and lowest precision for controlled. Comparing the graphs for the most basic members of the three classes shows single terms and controlled very similar, with concepts with very much lower recall (Figure 6.1P); however when the best members of each class are taken performance is very similar, with single terms probably superior to controlled and definitely superior to concepts (Figure 6.2P).

The main aim of the alternative document output cutoff representation based on simulated ranking was to provide single normalized recall figures for each language, following the Smart model. this supplies the overall merit ordering of the languages given in the key synoptic table, Figure 8.1T. This shows a best normalized value of 65.82 for the single term language with word forms conflated, and a worst of 44.64 for the simple concepts. More globally, all but one of the single term languages are placed first in the list, with normalized recall ranging from 65.82 down to 63.05, followed by two concept languages and then all the controlled languages, these with normalized recall from 61.76 to 59.17, followed by all the remaining concept languages. Abstracts and titles are comparable with controlled terms. A variety of subsidiary analyses show, for example, that absolute performance for different relevance grades varies, but that the inverse recall/precision relationship is maintained. Furthermore, it appeared that better performance was obtained for lower generality questions; that the basic questions performed better than the supplementary; and that different subject areas probably affect absolute if not relative performance.

From this mass of detailed results the Report authors draw two main conclusions. First, that

> 'every set of figures supports the original hypothesis of an inverse relationship between recall and precision. It is immaterial which variable is changed to give a new system; it may be the coordination level . . ., the exhaustivity of indexing . . ., the recall devices . . ., the precision devices . . ., the search programmes . . ., or the relevance decisions . . .; it has been impossible to find any exception to what can be claimed as a basic rule.' (p. 252)

Second, that

> 'quite the most astonishing and seemingly inexplicable conclusion that arises from the project is that the single term index languages are superior to any other type.' (p. 252)

With respect to the different language groups the authors conclude that there was an optimum level of specificity: the initial simple concepts were over-specific, so performance improved as the terms were broadened; the single terms were about right, so broadening degraded performance; and the controlled language came between the two, so broadening depressed performance, but only moderately. Thus more specifically, the authors concluded that:

> '(1) In the environment of this test, it was shown that the best performance was obtained by the use of Single Term index languages.
> (2) With these Single Term index languages, the formation of groups of terms or classes beyond the stage of true synonyms or word forms resulted in a drop of performance.
> (3) The use of precision devices such as partitioning and interfixing was not as effective as the basic precision device of coordination.' (p. 255)

The authors then consider whether the test environment was responsible in some specific way for the results. For as they say, their conclusion that the single term languages are superior

> 'is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used to obtain [the] results, and our own first reaction was to doubt the evidence. A complete recheck has failed to reveal any discrepancies, and unless one is prepared to say that the whole test conception is so much at fault that the results are completely distorted, then there is no other course except to attempt to explain the results which seem to offend against every canon on which we were trained as librarians.' (p. 252)

They conclude that it cannot be said that the subject area could not have distorted the results, but that collection size did not, that the relation between question and cited relevant documents is unlikely to have affected the results, that indexing failures or omissions were highly unlikely to have occurred sufficiently to have influenced the results, and that the classifications used were well prepared and so unlikely to have had any untoward effect.

They conclude that,

> 'with the possible doubtful exception of the subject field, there appears to be nothing in the test environment which could be held responsible for serious distortion of the results as between one system and another' (p. 262),

and continue,

> 'this test has shown that natural language, with the slight modifications of confounding synonyms and word forms, combined with simple coordination, can give a reasonable performance. This means that, based on such practice, a norm could be established for operational performance in any subject field, and it would then be for those who proposed new thesauri, new relational groups, links, or roles, to show how the use of their techniques would improve on the norm.' (p. 263b)

13.4 Criticisms of Cranfield 2

Unfortunately, though it is evident that Cranfield 2 was much more carefully designed than Cranfield 1, it was still open to methodological criticisms. Some of these were made by Vickery[20], who points out that the unexpected conclusions make it especially necessary to examine how the results were obtained. Thus he notes, for example, that the indexes were made for the document set vocabulary, and so certainly do not reflect an ordinary operational situation; that though the vocabulary distributions may have the same shape as those for larger document sets, absolute numbers of postings are low, again not reflecting an operational situation; that the search terms are less likely to be of varying subject generality than those of 'real' vocabularies; that the search broadening was very artificial; and finally, representing a methodological as much as substantive problem, that there could well be unusually close verbal links between relevant documents and queries.

Vickery also comments on the lack of statistical significance tests, and notes that normalized recall, used for the overall language ranking, is not very realistic as it does not depend on the cutoff point. At the same time he remarks with respect to the overall observations based on the ranking, namely that normalized recall is low for concepts as opposed to single terms, that it is low for generic single terms, and that it rises for generic concepts, that the third observation is widely accepted, that the second is not unexpected, but that the first is unexpected and so requires specific refutation by the advocates of controlled languages. In Vickery's view,

> 'the volumes of this report are an impressive account of a complex piece of research, undertaken with care and diligence. They give no final answers, and their conclusions must be treated with caution, but they are a valuable exploration of the retrieval process.' (p. 340)

Following the line of his attack on Cranfield 1, Swanson[21] suggests that the relation between questions and relevant documents is much too close, due to some features of the assessment procedure, especially the second step: thus the students screening for additional documents relevant to a question, used documents already known to be relevant to the question, while the user was allowed to modify his initial query after assessing the extra documents. Swanson indeed maintains that in the initial question provision the documents cited by the source paper, having been read, could have influenced the verbalization of the question. However a more serious criticism, in Swanson's view, is implied by some facts about the additional relevant documents: namely that bibliographic coupling gave a good many relevant documents not identified by the student screeners. This suggests that a large number of relevant documents were in fact missed altogether, the explanation being that the students were poor screeners as they only used titles, while the bibliographic coupling was only done at a high level. The implication is that the results for the whole set of experiments may be unreliable.

This point is considered in detail by Harter[22], who seeks to show, by formal arguments applied to real data, first that a good many relevant documents were missed; second, that changes in the relative proportions of missed to non-missed can affect recall/precision point values (as well as values over a cutoff range), with the important consequence that relative performance ratings can change; and third, using additional relevance data for a sample of languages, that the picture of relative language merit given by Cranfield 2 for these languages is changed when the additional relevance information is utilized. Unfortunately, while Harter's general point is sound, he indulges in some wild statistical extrapolation and very speculative global statements about the Cranfield results. He makes a good case for the principle that there may well be missed relevant documents unless evaluation is truly exhaustive and that omissions can affect performance, but his actual investigation of the data suggests that it is mostly the less important result for the title language, for which test biasses are evident, that was really affected in practice.

Other more general comments were made by, for example, Sharp and Rees. Sharp[23] notes that

> 'Cleverdon *et al.* . . . have qualified the basic recall/relevance thesis so that its application now seems so limited as to be confined to those conditions where . . . its truth is obvious.' (p. 92)

while Rees[24] remarks that

> ‘the problem of a criterion measure remains in that Cleverdon’s measure reflects the overall or ultimate performance of the system or subsystem tested. The sources of variation affecting performance are not adequately pinpointed, and small indication is given as to how to optimize performance.’ (p. 68)

In Rees’ view the basic assumption about relevance underlying Cranfield 2 had not been seriously questioned by 1967, while the methodology of the test was not blatantly defective; he notes that the project was not, unlike that at Case Western Reserve, regarded as having the explicit aim of developing test methodologies. He implies that the results are not seriously suspect, but at the same time argues that

> ‘the generalisability of these findings, and the problem of optimising system performance, remain.’ (p. 68)

He also comments on the difficulty of replicating the results.

Assessing these criticisms of Cranfield 2, it is apparent both that as Cranfield 2 was methodologically superior to Cranfield 1 the scope for criticism was reduced and that greater familiarity with the requirements and constraints of testing meant that some criticisms were more usefully pointed.

As before, some criticsms seem to have been fundamentally mistaken, like Sharp’s condemnation of the Report’s careful statement of the recall/precision relationship. The more plausible criticisms again fall into three groups. Vickery’s remark that the test did not reflect an ordinary operating system situation, like Mote’s earlier, is inappropriate to an explicitly laboratory test. Swanson’s and Harter’s claims about the existence of many more relevant documents than were used are themselves open to a good deal of doubt; they fall into the class of speculative criticisms. On the other hand, their point about the assessment procedure is more substantial, though there is no evidence that, while the procedure could have affected the test results, it actually did so. Both Cranfield 1 and Cranfield 2 were comparative tests and it is therefore necessary, in reviewing criticisms of the two experiments, to distinguish features of the design and conduct of the tests which could conceivably have affected comparative performance from those which were most unlikely in fact to have done so. Many of the criticisms of both tests failed to take this distinction into account. At the same time, the possibility that hidden factors may affect performance has to be raised in relation to every test.

The real defects of Cranfield 2 were the lack of statistical tests, noted by Vickery, and the failure to develop criterion measures pointed out by Rees. Overall, among the comments on Cranfield 2, Rees’ display the most insight, and correctly point the way forward for future tests building on both Cranfield 1 and Cranfield 2.

The relation to Cranfield 2 seems to have been rather less hostile than that to Cranfield 1. There were probably several reasons for this. First, the test was not manifestly open to major methodological criticisms like Cranfield 1 (Swanson’s and Harter’s papers were not published till five years later). In this connection it is worth noting that a subsequent test by Cleverdon with alternative relevance assessments[25], and Lesk and Salton's study[26], showed that comparative performance was not materially affected by different relevance assessments. Second, the test was not concerned with specific, established indexing languages like the UDC, or language types like facetted classifications, so the results could not be regarded as threatening by language proprietors and advocates. At the same time, insofar as the devices studied might be regarded as associated with types of language, it is likely that Cranfield 1 had softened up the potential opposition. Third, other work, notably by the Smart Project, suggested that the results were not eccentric, but could be paralleled for other data, with alternative test designs. Fourth, an increasing interest in the use of natural rather than artificial, controlled indexing languages, both for intellectual reasons among research workers and practical ones among the new computer-based system operators, meshed in with the Cranfield 2 results. An additional reason was perhaps the realization that anyone wishing to subvert the Cranfield findings, either in terms of a laboratory test or operational system investigation, would have to do a great deal of work.

For Cleverdon personally, Cranfield 2 was followed on the one hand by controversy about the recall/precision relationship[27], and on the other by involvment in tests like the DOAE[28] and Precision Engineering[29] studies of 1970, and the NASA experiment[30] of 1977, described in Chapter 12. These tests clearly show their descent from Cranfield 1 and 2, in being concerned with comparing index languages in relation to language control or indexing descriptions in relation to description exhaustivity; but they also represent studies of indexes and indexing in new environments: the NASA test for instance involved online searching. The underlying continuity of Cleverdon's work is also apparent in his interest in costs, which repeatedly figures in his discussions of system evaluation. In the research community there were two direct responses to Cranfield 2. One was the application of Cranfield 2 principles and practices in other tests, usually selectively, or with modifications, but perceptibly: the Inspec tests[31] and Keen's ISILT project[32] are good examples of this trend. Other projects, like that of Sparck Jones[33], utilized Cranfield 2 performance representation methods, for example. The second response was to exploit the Cranfield 2 collection, once it had been made available in machine-readable form. A great many Smart Project experiments were carried out with the Cranfield data[34], and the collection was also used by, for instance, Svenonius[35] in the US and by Sparck Jones and Bates[36] and van Rijsbergen and Croft[37] in the UK.

It is more difficult to assess the indirect influence of Cranfield 2 than of Cranfield 1. Substantively, Cranfield 2 suggested that retrieval systems may not work very well and are difficult to upgrade significantly. But this was being suggested by other tests in the later 1960s as well. Methodologically, it is very probable that Cranfield 2 encouraged care in the analysis of systems and the conduct of experiments, but by the later 1960s the same point was being made by other projects too. Cranfield 2 did not, moreover, lead directly into work on those topics which were regarded as most important in the 1970s, namely the effects of searching on performance and the design of mechanized systems. But that its indirect influence was great can be inferred from the fact that so many research workers in the late 1960s and 1970s referred to Cranfield.

13.5 Conclusion

What, then, is the Cranfield legacy? First, and most specifically, it has proved very difficult to undermine the major result of Cleverdon's work, namely that indexing languages, including natural language, tend to perform much the same: the gross substantive result of the research remains true. Second, methodologically, Cranfield 2, whatever its particular defects, clearly indicated what experimental standards ought to be sought. Third, our whole view of information retrieval systems and how we should study them has been manifestly influenced, almost entirely for the good, by Cranfield.

But none of this means that retrieval system testing is wholly well organized. Cranfield 1 and 2 raised questions about the replication of tests results, and more significantly, their extrapolation to a large scale, which have not been answered. Moreover, as Rees said of Cranfield 2, the work does not provide system design instructions. The major gap in the Cranfield work was indeed the absence of any models which could underpin design recommendations: there was certainly some '*ur-theorie*' underlying the Cranfield tests; but it's a long way from ur-theory to theory proper, and we have so far only taken a few steps along the road.

References

1. THORNE, R. G. The efficiency of subject catalogues and the cost of information searches, *Journal of Documentation* **11**, 130–148 (1955)
2. CLEVERDON, C. W. *Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Cranfield (1960)
3. CLEVERDON, C. W. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Cranfield (1962)
4. 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)
5. LANCASTER, F. W. and MILLS, J. Testing indexes and index language devices, *American Documentation* **15**, 4–13 (1964)
6. SWANSON, D. R. The evidence underlying the Cranfield results, *Library Quarterly* **35**, 1–20 (1965); this paper includes a comprehensive bibliography on earlier Cranfield-related literature
7. MOTE, L. J. B. Review of CLEVERDON, C. W. The Cranfield 1 1962 Report, *Journal of Documentation* **19**, 80–81 (1963)
8. RICHMOND, P. A. A review of the Cranfield Project, *American Documentation* **14**, 307–311 (1963)
9. SHARP, J. Review of the Cranfield-WRU Test Literature, *Journal of Documentation* **20**, 170–174 (1964)
10. TAUBE, M. A note on the pseudo-mathematics of relevance, *American Documentation* **16**, 69–72 (1965)
11. 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)
12. KYLE, B. R. F. Information retrieval and subject indexing: Cranfield and after, *Journal of Documentation* **20**, 55–69 (1964)
13. 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)
14. REES, A. M. *Review of a Report of the Aslib-Cranfield Test of the Index of Metallurgical Literature of Western Reserve University*, Centre for Documentation and Communication Research, Western Reserve University (1963)

15. REES, A. M. The Aslib-Cranfield test of the Western Reserve University indexing system for metallurgical literature: a review of the final report, *American Documentation* **16**, 73–76 (1965)
16. CLEVERDON, C. W. The Cranfield hypotheses, *Library Quarterly* **35**, 121–124 (1965); with rejoinder by Swanson, p. 125
17. CLEVERDON, C. W. The testing and evaluation of the operating efficiency of the intellectual stages of information retrieval systems, with following discussion. In: *Classification Research: Proceedings of the Second International Study Conference* (Ed. P. Atherton), Munksgaard, Copenhagen (1965)
18. CLEVERDON, C. W. and MILLS, J. The testing of index language devices, *Aslib Proceedings* **15**, 106–130 (1963)
19. CLEVERDON, C. W., MILLS, J. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems*, 2 Vols, College of Aeronautics, Cranfield (1966). (The Report material is summarized in CLEVERDON, C. W. The Cranfield tests on index language devices, *Aslib Proceedings* **19**, 173–194 (1967))
20. VICKERY, B. C. Reviews of CLEVERDON, C. W., MILLS, J. and KEEN E. M. The Cranfield 2 Report, *Journal of Documentation* **22**, 347–349 (1966) and **23**, 338–340 (1967)
21. SWANSON, D. R. Some unexplained aspects of the Cranfield tests of indexing language performance, *Library Quarterly* **41**, 223–228 (1971)
22. HARTER, S. P. The Cranfield II relevance assessments: a critical evaluation, *Library Quarterly* **41**, 229–243 (1971)
23. SHARP, J. Content analysis, specification and control. In: *Annual Review of Information Science and Technology*, Vol. 2 (Ed. C. A. Cuadra), Interscience, New York (1967)
24. REES, A. M. Evaluation of information systems and services. In: *Annual Review of Information Science and Technology*, Vol. 2 (Ed. C. A. Caudra), Interscience, New York (1967)
25. CLEVERDON, C. W. *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages*, OSTI Report 5075, Cranfield Institute of Technology (1970)
26. LESK, M. E. and SALTON, G. Relevance assessments and retrieval system evaluation, *Information Storage and Retrieval* **4**, 343–359 (1968)
27. CLEVERDON, C. W. On the inverse relationship of recall and precision, *Journal of Documentation* **28**, 195–201 (1972)
28. CLEVERDON, C. W. *An Investigation into a Suitable Mechanised Information Retrieval System at the Defence Operational Analysis Establishment*, Cranfield Institute of Technology (1970)
29. 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)
30. 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)
31. AITCHISON, T. M. *et al. Comparative Evaluation of Indexing Languages, Part II: Results*, Report R70/2, INSPEC, Institution of Electrical Engineers, London (1970)
32. KEEN, E. M. and DIGGER, J. A. *Report of an Information Science Index Languages Test*, 2 Vols, College of Librarianship Wales, Aberystwyth (1972)
33. SPARCK JONES, K. *Automatic Keyword Classification for Information Retrieval*, Butterworths, London (1971)
34. SALTON, G. (Ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, N.J. (1971)
35. SVENONIUS, E. An experiment in index term frequency, *Journal of the American Society for Information Science* **23**, 109–121 (1972)
36. 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)
37. 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)