6

Evaluation within the environment of an operating information service

F. Wilfrid Lancaster

This chapter deals with the problems of evaluating operating information services, the objective of such evaluation being to determine how successful the service is in satisfying the needs of its users. The major emphasis of the chapter is the 'human aspects' of information services, machine aspects being dealt with in the chapter by Barraclough. The chapter first introduces some basic concepts and definitions related to evaluation. The special problems involved in applying evaluation methods to operating services are then discussed.

There exist a number of possible reasons why the managers of an information centre may wish to conduct an evaluation of the services provided. One is simply to establish a type of 'benchmark' to show at what level of performance the service is now operating. If changes are subsequently made to the services, the effects can then be measured against the benchmark previously established. A second, and probably less common, reason is to compare the performance of several information centres or services. Since a valid comparison of this type implies the use of an identical evaluation standard, the number of possible applications of this kind of study tend to be quite limited. Examples include studies of the coverage of different data bases (e.g. Ashmole *et al.*[1], Davison and Matthews[2]), the comparative evaluation of the document delivery capabilities of several libraries (Orr *et al.*[3]), and the use of a standard set of questions to compare the performance of question-answering services (Crowley and Childers[4]). A third reason for evaluation of an operating information service is simply to justify its existence. A justification study is really an analysis of the benefits of the service or an analysis of the relationship between its benefits and its cost. The fourth reason for evaluation is to identify possible sources of failure or inefficiency in the service, with a view to raising the level of performance at some future date. Using an analogy with the field of medicine, this type of evaluation can be regarded as diagnostic and therapeutic. In some ways it is the most important type. Evaluation of an information service is a sterile exercise unless conducted with the specific objective of identifying means of improving its performance.

Evaluation of an information service may be **subjective**, based solely on the opinions of its users, or it may be **objective**. Subjective, opinion-based studies have their value since it is important for the managers of information services to know how users feel about these services. On the other hand, purely subjective studies have obvious limitations. Generally speaking, the satisfaction of a user with a service is a relative thing, based on what the user knows and what he does not know. To take an obvious example, the recipient of the results of a literature search may express satisfaction with these results based upon what is known to him. Thus, if most of the references retrieved are relevant to his interests, he may be quite happy with the results. But the unknowns of the situation have not entered at all into this evaluation. The requester might be much less satisfied with the results if he knew that a substantial number of relevant references were missed by the search, especially if some of these were, in some sense, 'more relevant' to his interest than those retrieved. An objective evaluation would try to quantify the results of the search: to determine how many relevant items were missed as well as to determine how many of the items retrieved are considered relevant. An objective, quantitative approach is usually needed for diagnostic evaluation purposes.

6.1 Levels of evaluation

An information service can be studied at any of the following levels:

(1) cost;
(2) effectiveness;
(3) benefit;
(4) cost-effectiveness;
(5) cost-benefit;
(6) cost-performance-benefit;

which approximates to a sequence of increasing complexity.

The **cost** of an information service, obviously, refers to the resources expended on that service. While, in theory, the cost analysis of a service may seem quite straightforward, the danger exists of overlooking some tangible or intangible costs. The most common pitfall is that of overlooking costs to the user of the service. The fact that a user may not be required to pay out money for the service does not mean that there is no user cost involved. Clearly, the time and effort of the user is a cost that must be charged against the service in any realistic cost analysis. The user of a literature searching service may spend one hour of his time in making his information needs known to the service and another hour in examining and evaluating the results of the search. Allowing for overheads, this time could be worth, say, $80 in an industrial organization. To ignore this user's time would lead to a completely distorted picture of the cost or the cost-effectiveness of the literature search within the complete institutional environment.

The **effectiveness** of an information service is the extent to which the needs of the users are satisfied by the service. An evaluation of effectiveness attempts to determine **satisfaction level**. It should preferably be objective and quantitative, expressed in such terms as '80 per cent of the document delivery needs of users are satisfied' or '60 per cent of factual questions are answered completely and correctly'.

Measures of **benefit** present great problems in the information environment since it is difficult, if not impossible, to assess the benefit of 'information' or 'information service' in any tangible (e.g. financial) terms. Studies of benefit tend to be more subjective than objective. They deal with 'perceived value' of the service (in the eyes of the user) or attempt to look at some facet of the impact of the service on the user (e.g. its effect on his information seeking behaviour).

A **cost-effectivenss** study, quite obviously, is one that attempts to relate the cost of a service to its level of effectiveness. It looks at both sides of the cost-effectiveness equation. The cost-effectiveness of the service is improved if either:

(a) the level of effectiveness is increased while costs are held constant, or
(b) the costs are reduced while the level of effectiveness remains constant.

Frequently a cost-effectiveness analysis is conducted in order to choose among several competing strategies for implementing some service.

A **cost-benefit** study, similarly, is one that relates the costs of providing the service to the benefits derived from it. Cost-benefit analyses are really attempts to justify the existence of the service; they ask the question 'Do the benefits derived from the service exceed, in some sense, the costs associated with providing it?'. Finally, a **cost-performance-benefits** analysis is one that investigates the entire set of interrelationships among system costs, performance (level of effectiveness) and benefits (Lancaster[5]).

Because benefits are so difficult to pin down, it is difficult to achieve true credibility in a cost-benefit study of an information service. Several attempts have been made, however, with varying degrees of success. Approaches used include those that try to justify the cost of an information service by:

(1) Showing that, if an in-house information service did not exist, it would cost the organization more to buy the same level of service from outside sources (e.g. Mason[6], Magson[7]).
(2) Proving that costly research may be duplicated if adequate information services are not available or are not used effectively (Martyn[8]).
(3) Showing that information services can improve the quality of decision making or can reduce the professional level of the staff needed to make various types of decisions (McDonough[9]).
(4) Showing that, if an in-house information service were not available, costs to the organization would increase as engineers, scientists or other professionals are forced to spend more of their own time in information-seeking activities with resulting loss in their own productivity (Rosenberg[10], Kramer[11], Mueller[12]).
(5) Proving that an information service has contributed tangibly to the organization (e.g. by helping to win a contract, by revealing cheaper solutions to research or production problems, or by stimulating the development of new products).

It is probably true to say that most managers would like to be able to prove that the services they provide can be justified from a cost-benefit point of view, but the difficulties involved in such a study have discouraged all but a few attempts of this kind. Consequently, evaluations of information services tend to ignore benefit considerations (in effect, taking the benefit as a kind of 'given') and focus instead on one of the other levels. Most of the evaluations of operating systems have, in fact, been restricted to evaluations of their effectiveness (e.g. in terms of the number of users who express subjective satisfaction or the number of actual demands that are satisfied according to some more objective criteria). Few detailed cost analyses have been conducted or, at least, few are reported in the literature. And realistic cost-effectiveness analyses are even more scarce. This is a pity because it can be argued that a study of effectiveness has little real meaning unless related to costs and that, certainly, a cost analysis has little real value unless related to level of effectiveness. Managers of information services are, or should be, concerned with optimum allocation of the resources available (i.e. one that achieves the maximum quality of service possible within budgetary constraints) and optimum resource allocation is only likely to come from a true cost-effectiveness analysis.

A useful distinction, first made by King and Bryant[13], is that between **macroevaluation** and **microevaluation**. A macroevaluation of a system is one that measures its present level of performance (e.g. in terms of recall and precision or as a document delivery score) and is content to let the study rest there. A macroevaluation, then, merely establishes a benchmark. But a microevaluation goes much beyond this. It seeks to answer such questions as 'Why is the system operating at this level?', 'Under what conditions does the system perform well and under what conditions does it perform badly?', and 'What can be done to raise the level of performance in the future?' A microevaluation, then, is diagnostic while a macroevaluation is not.

Another possibly useful distinction in the information services environment is that between **inputs**, **outputs** and **outcomes**. Again, this is a sequence of increasing complexity. Inputs to an information service are the easiest things to measure. They can be expressed in purely quantitative terms: how many documents, how many people, how much money? Outputs are more difficult to deal with because output measures must take into account quality as well as quantity. For example, in the evaluation of a question-answering service the appropriate output measure is not the number of questions submitted. It is not even the proportion of questions for which an answer is supplied. It is the proportion of questions submitted for which a complete and correct answer is supplied. The outcome of an information service is the most difficult aspect to study for the notion of outcome brings us back to that of impact, effect or benefit. It is more difficult to evaluate outcomes than it is to evaluate outputs and it is more difficult to evaluate outputs than it is to quantify inputs. All types of information services will probably have reliable input data but few have meaningful qualitative output data and data on outcomes are likely to be non-existent. Where they exist in the information services field (e.g. applied to various types of libraries), standards tend to be entirely related to inputs. This is not because inputs are most important (far from it) but merely because inputs are easiest to look at, quantify and reduce to 'standard' form.

In the evaluation of an operating information service we should primarily be interested in its outcomes. After all, it is the beneficial outcomes that presumably justify the existence of the service. But it may not be possible to evaluate outcomes; or, at least, the evaluation of outcomes may be so complex as to discourage the attempt. On the other hand it should be possible to identify the desired outcomes of an information service and to select output measures that are at least predictors of the desired outcomes. Looked at in this way, an appropriate output measure may be regarded as at least a distant approximation of an outcome measure. To take one example, the desired outcome of an SDI service is presumably to make the users of the service better informed. The degree to which this outcome is achieved, however, is virtually impossible to measure. Nevertheless, it seems reasonable to suppose that an SDI service is more likely to make a user better informed if it brings to his attention documents that directly match his interest, and were previously unknown to him, than if it is unable to deliver any matching items. In this case, then, we have identified output measures (recall, precision, novelty) that can be regarded as approximations of the desired outcome measure. Likewise, in certain situations, we can identify input measures that can be regarded as predictors of outcomes. The size of the collection of a library, or its rate of growth, for example, might be regarded as a reasonable predictor of the document delivery capabilities of that library.

The input/output/outcome distinction may be considered related to the distinction between **long range** and **short range** objectives. Drucker[14] has pointed out that it is virtually impossible to evaluate any type of service institution against its long range objectives. Instead, we should back away from the long range objectives and identify short range objectives that are distant approximations of the long range objectives and that can be converted into meaningful evaluation criteria. As one example, Drucker points to the 'saving of souls' as the long range objective of the church. The extent to which this objective is reached by a particular church, however, is, to say the least, an unpromising evaluation problem. On the other hand, a short range objective of the church may be to encourage young people in the community to attend services and other church activities. The extent to which this is achieved is precisely measurable. If we accept that church attendance may contribute to the saving of souls, evaluation against the short range objective may be regarded as a distant approximation of evaluation against the long range objective.

Before leaving the subject of evaluation levels, it may be worth pointing out that, in certain information service applications at least, purely quantitative measures may relate only to successes but ignore failures completely. An obvious example is library circulation figures. A book borrowed by a user reflects, in some sense, a library success, but circulation figures tell us nothing about the library's failures—how many users are unable to find the items they seek. In this case a purely quantitative measure gives us a very incomplete picture of the library's performance. We need, instead, a qualitative measure, one that balances the successes against the failures, in this case some type of document delivery score.

6.2 Evaluation criteria

The users of services of any kind usually evaluate them, consciously or unconsciously, against cost, time and quality criteria. Users of information services also tend to judge them against these same criteria. The specific criteria that seem most important in the information service environment are listed in *Table 6.1*.

Cost factors are as important in the evaluation of information services as they are in the evaluation of other services and products. The service must be provided at a cost that the user feels is reasonable in relation to the benefits associated with it. Cost to the user involves more than direct charges. It includes the cost of his own time, that is, how much effort is involved in the use of the system. Studies of the information-seeking behaviour of scientists and other professionals have consistently shown that accessibility and ease of use are the prime factors influencing the choice of an information source. In general, the most convenient source of information is chosen, whether or not it is perceived by the user to be, in some sense, 'the best'.

TABLE 6.1. Criteria by which users will evaluate an information service

Level 1. Evaluation of effectiveness (considerations of user satisfaction)
  a. Cost criteria
    (1) Monetary cost to user (per search, per subscription, per document)
    (2) Other, less tangible cost considerations
      (a) Effort involved in learning how to use system
      (b) Effort involved in actual use
      (c) Effort involved in retrieving documents (through backup document delivery systems)
      (d) Form of output provided by the system
  b. Time criteria
    (1) Time elapsing from submission of request to retrieval of citations
    (2) Time elapsing from submission of request to retrieval of documents
    (3) Other time considerations—for example, waiting time to use an online system
  c. Quality considerations
    (1) Coverage of the data base
    (2) Completeness of output (recall)
    (3) Relevance of output (precision)
    (4) Novelty of output
    (5) Completeness and accuracy of data
Level 2. Evaluation of cost effectiveness (user satisfaction related to internal system efficiency and cost considerations)
    (1) Unit cost per relevant citation retrieved
    (2) Unit cost per new, that is, previously unknown, relevant citation retrieved
    (3) Unit cost per relevant document retrieved
Level 3. Cost-benefit evaluation (value of system balanced against costs of operating it)

Ease of use factors include ease of interrogating the system in the first place, that is, ease of making one's needs known, and ease of use of the output provided by the system, especially the ease with which the output can predict the relevance of the documents it refers to. A very important facet of the latter is availability of an efficient and convenient document delivery capability. A service that stops at the delivery of bibliographic citations goes only part of the way toward satisfying an individual's information needs. Such a service causes considerable frustration if the user is unable to obtain the documents cited or can do so only through procedures that he views as inconvenient and time-consuming.

The users of information services have various kinds of information needs, including the need for:

(1) A particular document whose identity is known.
(2) Specific factual information of the type that might come from some type of reference book or from a machine-readable data bank—for example, thermophysical property data on a particular substance.
(3) A few 'good' articles, or references to them, on a specific topic.
(4) A comprehensive literature search in a particular subject area.
(5) A current alerting service by which the user is kept informed of new literature relevant to his current professional interests.

These different needs have different response time requirements associated with them. The requirement relating to the current alerting service is that it should deliver regularly and frequently and that the information supplied should be as up-to-date as possible. The user needing a comprehensive literature search is usually engaged in a relatively long-term research project. Speed of response may not be critical to him, except that there may be some date beyond which the search results will have no value or, at least, greatly reduced value; he is willing to wait longer in order to achieve completeness; that is, completeness is more important to him than speed. For the other types of information needs, on the other hand, the user generally wants fairly rapid response.

The cost and time criteria relevant to the evaluation of information services seem fairly obvious and are relatively constant from one activity to another. But the quality criteria are perhaps less obvious and vary considerably with the particular service being evaluated. They may also vary with the kind of need that a particular user has in relation to a service.

There seem to be two major qualitative measures of success as applied to information services:

(1) Does the user get what he is seeking or not?
(2) How completely or accurately does he get it?

The first of these measures, which applies, for example, to the search for a particular item or the answer to a particular factual question, is simple and unequivocal. The second, however, is much more difficult to apply in practice because it implies both a human value judgement and the use of some graduated scale to reflect degree of success. The second type of measure is necessary, however, in the evaluation of most types of information retrieval activity. 'Recall' and 'precision' are two criteria frequently used to judge the performance of a search in an information retrieval system. Because these measures are well known and well accepted in the evaluation of operating information services, they will not be defined here.

The precision ratio and the recall ratio, used jointly, express the filtering capacity of the system—its ability to let through what is wanted and to hold back what is not. Neither one on its own gives a complete picture of the effectiveness of a search. It is always possible to get 100 per cent recall if we retrieve enough of the total collection; if we retrieve the entire collection, we certainly achieve 100 per cent recall. Unfortunately, however, precision would be extremely low in this situation because, for any typical request, the great majority of the items in the collection are not relevant.

The precision ratio may be viewed as a type of cost factor in user time—the time required to separate the relevant citations from the irrelevant ones in the output of a search. Consider, as an illustration, a search request for which there are 20 relevant documents in a particular database. Suppose that three different search strategies are used to interrogate the system and that each retrieves 15 of the 20 relevant items; that is, recall is 75 per cent. In the first search, the total number of items retrieved is 30, in the second it is 60, and in the third it is 150. The precision ratio in these three searches is 50, 25, and 10 per cent, respectively. In the first search the user has to examine only 30 citations to find the 15 of relevance; in the second, 60; and in the third, 150. All other things being equal, it takes him longer to separate the relevant from the irrelevant in the second search than in the first, and considerably longer in the third. It is in this sense that we can regard the precision ratio as a measure of user effort or cost. A search that achieves 75 per cent recall at 25 per cent precision is more efficient than one that achieves 75 per cent recall at 10 per cent precision.

Not everyone needs high recall all the time. Different users have different requirements for recall and precision, and a particular individual has different requirements at different times. The precision tolerance of the user is likely to be directly related to his recall requirements. At one end of the spectrum we have the individual who is writing a book, preparing a review article, or beginning a long-term research project. He is likely to want a comprehensive (high recall) search, and he may tolerate fairly low precision in order to assure himself that he has not missed anything of importance. At the other end, we have the typical user of, say, an industrial information service who needs a few recent articles on a subject and needs them right away. He does not need high recall but he expects high precision in the search results. Other individuals may prefer a compromise; they would like a 'reasonable' level of recall at an 'acceptable' level of precision.

It seems rather pointless to use the recall ratio as a measure of the success of a search in which high recall is unimportant. This has led some writers to suggest the use of some measure of proportional recall, or relative recall, in which the success of the search is expressed in terms of the number of relevant documents retrieved over the number of relevant documents wanted by the requester. For example, the requester specifies that he needs five relevant documents, but the search retrieves only three. The proportional recall ratio is, therefore, 3/5, or 60 per cent. This measure, although attractive on the surface, is rather artificial in that very few requesters are able to specify in advance just how many documents they want from the system.

Another limitation of the recall ratio is that it more or less assumes that all relevant documents have approximately equal value. This is not always true. A search may retrieve 5 relevant documents and miss 10 (recall ratio = 33 per cent), but the 5 retrieved may be much better than the 10 missed. They could, for example, be more up-to-date and might in fact make the other 10 items completely redundant. The recall ratio, although important, must therefore be used with some caution in the evaluation of information services.

The precision ratio also has its limitations. As we have already seen, it is actually an indirect measure of user time and effort spent at the output stage of the information retrieval process; that is, the higher the precision ratio, the less effort the user needs to expend in separating relevant items from those which are not. In a search of very low precision ratio in which, say, only 10 items among 80 retrieved are judged relevant, considerable user time and effort might be required to identify the relevant items in a printed or typed list, especially if it contains only bibliographic citations and the user must himself retrieve copies of many of the documents before he can decide which are relevant and which are not. But this measure of effort is really only appropriate to the evaluation of a delegated search—one conducted on behalf of the requester by an information specialist. In this situation the system is viewed as more or less a 'black box', into which a request is placed and out of which comes a group of documents or references to them. The precision ratio is a valid measure of the performance of any type of delegated search in which the information seeker submits a request to some 'system' and waits for the results, whether the search is manual or fully mechanized.

The precision ratio is not especially meaningful when applied to the non-delegated search. Here, the user conducts his own search and make relevance decisions continuously as he proceeds; that is, when he consults an index term in a printed index or an online system, he rejects irrelevant citations and records only those which seem relevant. A precision ratio could be derived for this type of search by counting the total number of citations the user consulted and the number he judged relevant, the precision ratio being the number of relevant citations found divided by the total number of citations consulted. This is a rather artificial measurement, however, because user effort in the non-delegated search situation can be expressed more directly in terms of the time required to conduct the search, and, for this, a unit cost in time, per relevant item found can be determined. Presumably, the higher the precision of a non-delegated search (proportion of relevant items examined to the total items examined), the less time it takes, all other things being equal.

Leaving aside direct costs, four performance criteria by which any type of literature search, manual or mechanized, may be evaluated from the viewpoint of user satisfaction have been discussed thus far: recall, precision, response time, and user effort. The salient points of these performance measures are as follows:

Recall. Important to all users of information services who are seeking bibliographic materials on a particular subject. In some cases, only a minimum level of recall is required—for example, one book or a few articles on a particular subject—and this is likely to be the most typical situation. In other cases, maximum recall is sought—for example, the user who wants a comprehensive search conducted in *Chemical Abstracts*.

Precision. A meaningful measure of the performance of a delegated search conducted in any form of system, manual or mechanized. It is an indirect measure of user time and effort and not particularly appropriate in the evaluation of non-delegated searches, including non-delegated searches in online retrieval systems.

User Effort. In a non-delegated search, effort is measured by the amount of time the user spends conducting the search. In a delegated search, it is measured by the amount of time the user spends negotiating his inquiry with the system and the amount of time he needs, when the search results are delivered to him, to separate the relevant from the irrelevant items, which is directly related to the precision ratio.

Response Time. In a delegated search, this represents the time elapsing between the submission of a request by the user and his receipt of the search results. In a non-delegated situation, it represents the time involved in the actual conduct of the search; in this case, it is also a measure of user effort.

Table 6.1 lists some further performance criteria that may be applied in the evaluation of information retrieval systems, including 'coverage' and 'novelty'. Coverage may be considered an extension of recall; it is expressed in terms of how much coverage of the literature on a specific subject is provided by a particular database. Suppose, for example, that a scientist wishes to find all possible references to the use of lasers in eye surgery. An obvious source would be the printed *Index Medicus* or, even better, the computer-based MEDLINE service operated by the National Library of Medicine (NLM). Suppose also that the search in the NLM database retrieves everything of relevance, that is, achieves 100 per cent recall—a rather unlikely situation. Even if the search is complete, so far as the database is concerned, the user who needs a really comprehensive search also wants to know the exact coverage of the database, that is, what proportion of all the literature on eye surgery using lasers is contained in the database. Searching a particular database may result in 100 per cent recall but may give a low overall coverage of the literature. Absolute coverage of the collection is only of direct concern to the person who needs a comprehensive search. It is probable that the user whose need is satisfied by finding, on the library shelves, one or two books on a subject of interest is quite unconcerned as to how complete the library's collection may be in this subject area. At a later time, however, he may require a comprehensive search on this or some other topic, and the coverage of the collection consulted would then be important to him. Coverage, like recall and precision, can be expressed as a percentage. If, for example, the results of a search conducted in *Chemical Abstracts* were being evaluated, it could be estimated, not very easily, that the recall ratio is 75 per cent; it could also be estimated, even less easily, that the coverage of *Chemical Abstracts* on the subject area of the search is 40 per cent. With an estimated coverage of 40 per cent and recall of 75 per cent the overall estimate of the comprehensiveness of the search is 30 per cent.

Another performance measure that may have some value is the novelty ratio, the proportion of relevant items retrieved in a search that are new to the requester, that is, brought to his attention for the first time by the search. The novelty ratio is particularly appropriate in the evaluation of literature searches conducted for current awareness purposes, that is, SDI, since, presumably, a good current awareness service brings documents to the attention of users before they learn of them by other means.

When cost criteria are related to quality criteria, cost-effectiveness criteria are derived. Some possible cost-effectiveness criteria applicable to information services include the unit cost per relevant item (document or document reference) retrieved and the unit cost per new relevant item retrieved. Cost can be measured directly in monetary units or in time and effort expended.

There is still one further evaluation criterion listed in *Table 6.1*, namely, accuracy of data. This criterion substitutes for recall and precision in the evaluation of information services designed to answer questions that have unequivocal factual answers. The answer to such a question as 'What is the melting point of . . .?' is either supplied completely and correctly or it is not. Question-answering services, whether the answer is supplied from a printed source or a machine-readable data bank, must therefore be evaluated in terms of the completeness and accuracy of the data supplied.

6.3 Some problems of evaluation applied to operating systems

In a book on experimentation in information retrieval a chapter on the evaluation of operating systems may be considered something of an interloper. Evaluation does not necessarily imply experimentation. In fact, the evaluation of an operating system will not usually involve any experimentation. It will merely be an analysis of the performance of the system at a particular point in time. This does not mean that experimentation is impossible within an operating environment. It is possible, but it might be quite difficult. Although most are not framed in this way, it is possible for an evaluation of an operating system to take the shape of a formal research project with a hypothesis that the investigators set out to test. An example of such a hypothesis might be 'Literature searches conducted for requestors who visit the centre in person produce better results in terms of recall and precision than those conducted for users who submit their requests by mail to the centre'. It is possible to build an evaluation upon a research hypothesis of this type. But, unlike the true experimental situation, in an operating environment it may be quite difficult to control all the independent variables that may affect the results.

This is not to imply that controls should not be sought. Variables extraneous to the focus of the study should be controlled as much as possible. In the evaluation of an operating system, established principles of experimental design, sampling, survey methodology, and other methodological issues, are as relevant and important as they are in any other evaluation situation. It must be recognized, however, that, when dealing with a real-life environment, some methodological compromises may need to be made.

While the evaluation may not be based on a formal research hypothesis, it should certainly have some clearly defined evaluation objectives.

The major steps involved in the conduct of an evaluation programme are the following:

(1) Defining the scope of the evaluation.
(2) Designing the evaluation programme.
(3) Execution of the evaluation.
(4) Analysis and interpretation of the results.
(5) Modifying the system or service on the basis of the evaluation results.

Definition of scope

The first step, the *definition of scope*, entails the preparation of a precise set of questions that the evaluation must be designed to answer. The purpose of an evaluation is to learn more about the capabilities and weaknesses of a system or service, and the definition of scope is really a statement of what precisely is to be learned through the study. The definition of scope must be prepared by the person requesting the evaluation, who is usually one of the managers of the system or one of those responsible for funding it. It is the responsibility of the evaluator to design a study capable of answering all the questions posed in the definition of scope. A sample work statement for an evaluation programme is given in *Table 6.2*. This statement is a list of the questions to be answered in the MEDLARS study, as reported by Lancaster[15]. It is a rather long list because the study was a comprehensive evaluation of a very large system. Evaluation studies of more modest scope would involve fewer questions. In fact, it is quite conceivable that an evaluation might be designed to answer only one or two important questions.

TABLE 6.2. Example of a 'Work Statement' for an evaluation of an operating information service

Overall performance

(1) What is the overall performance level of the system in relation to user requirements? Are there significant differences for various types of request and in various broad subject areas?

Coverage and processing

(1) How sound are present policies regarding indexing coverage?
(2) Is the delay between receipt of a journal and its processing in the indexing section significantly affecting performance?

Indexing

(1) Are there significant variations in inter-indexer performance?
(2) How far is this related to experience in indexing and to degree of 'revising'?
(3) Do the indexers recognize the specific concepts that are of interest to various user groups?
(4) What is the effect of present policies relating to exhaustivity of indexing?

Index language

(1) Are the terms sufficiently specific?
(2) Are variations in specificity of terms in different areas significantly affecting performance?
(3) Is the need for additional precision devices, such as weighting, role indicators, or a form of interlocking, indicated?
(4) Is the quality of term association in the thesaurus adequate?
(5) Is the present entry vocabulary adequate?

Searching

(1) What are the requirements of the users regarding recall and precision?
(2) Can search strategies be devised to meet requirements for high recall or high precision?
(3) How effectively can searchers screen output? What effect does screening have on recall and precision figures?
(4) What are the most promising modes of user/system interaction?
    a. Having more liaison at the request stage.
    b. Having more liaison at the search formulation stage.
    c. An iterative search procedure that presents the user with a sample of citations retrieved by a 'first approximation' search, and allows him to reformulate his request in the light of these retrieved items.
(5) What is the effect on response time of these various modes of interaction?

Input and computer processing

(1) Do input procedures, including various aspects of clerical processing, result in a significant number of errors?
(2) Are computer programs flexible enough to obtain desired performance levels? Do they achieve the required checks on clerical error?
(3) What part of the overall response lag can be attributed to the data processing subsystem? What are the causes of delays in this subsystem?

Designing the evaluation

The second step of the evaluation involves the *preparation of a plan of action* that allows the gathering of data needed to answer the questions posed in the definition of scope. The designer of the study must identify what data are needed to answer each question and what procedures could be used to gather the data in the most efficient and expedient way. For each question, the evaluator must decide whether (1) it can be answered simply by collecting data from the system as it presently exists or (2) some changes in the normal functioning of the system must be made in order to collect the necessary data. For example, the question 'What is the present response time of the system, expressed in ranges, means, medians, and modes?' can be answered from the system as it is now. It requires only the collection of data on the date and time a request is received and the date and time the results are submitted to the requester, for a representative sample of transactions. To answer a question of this kind, new records may need to be created for the purpose of the study, but, apart from record keeping, the existing system is not perturbed in any way. In contrast, consider the question 'What would be the effect on response time if action X were carried out?' This implies a change in the present system, and the question can be answered only by deliberately applying action X to a representative sample of transactions and comparing the response times with those of the system as it normally functions.

In some cases, then, the evaluator is primarily concerned with systematic and controlled observation of the system. In other cases, however, he needs to go beyond simple observation of this kind and into the field of experimental design. In the evaluation programme it is important that well-established procedures of experimental design be followed and that appropriate statistical techniques be applied to the analysis and interpretation of the results.

Execution of the programme

The third step, *execution of the evaluation*, is the stage at which the data are gathered once the evaluation design has been agreed on by all the parties concerned. This stage is likely to be the longest in terms of elapsed time. It may also be the stage in which the evaluator is least directly involved and perhaps the stage over which he has the least direct control. Although the execution stage can hardly begin before the design stage is completed, the analysis and interpretation stage should certainly begin before the execution stage is concluded; that is, the evaluator must ensure that he receives data continuously from the beginning of the execution stage, so that they can be reduced to a form suitable for analysis and interpretation. It should be fairly obvious what is involved in the analysis and interpretation stage of an evaluation project. Here the evaluator is concerned with reducing the data and manipulating it in such a way that it can answer, or at least contribute to answering, the questions posed in the work statement. It is not possible to present any precise guidelines for analysis and interpretation because they vary considerably from one evaluation application to another. In the case of the evaluation of an information retrieval system, this stage of the study is mainly concerned with the derivation and manipulation of performance results—for example, recall and precision ratios—and with the analysis of recall and precision failures. The failure analysis itself entails an examination of each document involved, the indexing records for the documents, the requests that caused the searches to be conducted, the search strategies, the system vocabulary, and the relevance assessments of the users. Through the examination of each of these it should be possible to determine which component of the system was largely responsible for the failures occurring. In addition to the analysis of the failures occurring in particular searches, the evaluator can use the recall and precision ratios, or alternative measures of search performance, as indicators of conditions under which the system seems to perform well and under which it seems to perform badly. For example, searches can be grouped by broad subject category, and an average performance figure, or figures, can be derived for each group. It would then be possible to identify subject areas in which unusually low scores occur. Through the joint use of performance figures, in this way, and analyses of failures in particular searches, the evaluator learns a great deal about the characteristics of the system, its weaknesses and limitations as well as its strong points. The joint use of the performance figures and failure analyses should answer most of the questions identified in the work statement for the evaluation. The final element in the analysis and interpretation phase is that in which the evaluator presents his report to the managers of the system, including in his report recommendations on what might be done to improve its performance. The fifth and final step of the evaluation programme is that in which some or all of the recommendations are implemented, that is, the step in which the evaluation results are applied to the improvement of the system.

Although not specifically mentioned in the discussion above, the value of a pretest should be recognized. Before the complete evaluation is carried out, it is important to follow through all the proposed procedures on a small sample of transactions, to ensure that the procedures are, in fact, viable and that they are capable of gathering the data needed to complete the study.

One obvious problem associated with an operating system is the fact that, because a rather high level of user co-operation may be demanded (e.g. in assessing the relevance of items retrieved), it is virtually impossible to conduct the evaluation unobtrusively. Not only will the users know they are participating in a study but, since we may need their co-operation in collecting and delivering various records, the staff may also be aware that an evaluation is taking place. How much effect this obtrusiveness is likely to have on the evaluation results is a matter of some debate. It could be argued that the overall performance figures for 'observed' searches may exceed, on the average, overall performance figures for unobserved searches. In this sense, the evaluation may be considered a study of the system under somewhat ideal conditions. On the other hand, the effect of the obtrusiveness may be considered to apply equally to all searches included in the evaluation. This being so, the obtrusive nature of the study may not have any significant effect for certain evaluation purposes. For example, the obtrusiveness of a study is unlikely to affect a comparison of performance across various subject fields since there is little reason to suppose that the obtrusiveness will affect one subject area more than another. This is somewhat similar to the findings of Lesk and Salton[16] that even major differences in relevance decisions, from one judge to another, had no significant effect on comparison of the performance of various processing options in SMART. Obtrusiveness may, however, have some effect on a diagnostic microevaluation since certain system components may benefit from the spotlight effect while others are unable to benefit. The obtrusiveness of a study cannot in any way minimize failures attributable to the database searched (i.e. indexing and vocabulary failures) but it might reduce failures relating to the exploitation of the database since a searcher, knowing he is observed, may put more effort into his interaction with the user and into the construction of the search strategy itself.

The evaluation of an operating information system usually requires many more compromises than the evaluation of an experimental system. To begin with, we probably don't want to evaluate all searches conducted (even all conducted within a restricted time period) but only a sample of these searches. Ideally we would like to draw these searches completely at random. But in a national system, with potential users spread over great distances, a purely random assignment may be impracticable. The difficulties of dealing remotely with many geographically dispersed users may be too great. Instead of drawing a completely random sample of users, we may have to be content with some compromise. In his evaluation of MEDLARS, for example, Lancaster[15] identified a number of organizations whose members, based on records of searches conducted in the past, might be considered to form a microcosm of the complete user population. Not only could these organizations, collectively, be expected to generate the required number of searches, but the distribution of their searches by subject could be expected to resemble rather closely the subject distribution of all requests from whatever source. Defining a search to be evaluated as one coming from a selected group of organizations greatly facilitated the conduct of the evaluation since contacts with the requesters, including distribution of the necessary evaluation forms and other materials, could be entrusted to librarians or other information specialists on the staff of these organizations. Moreover, with a limited number of organizations involved, it was possible to secure agreement to co-operate from the executive officer of each organization. This was an encouragement to the co-operation of the individual staff members without in any way influencing the type of requests they made to the system.

The problems involved in securing the co-operation of large numbers of users of informations services has encouraged the use of 'realistic simulations' of these services in certain evaluation applications. In such simulations a 'proxy' of a real user is employed. The proxy behaves in a way that is assumed to be typical of the behaviour of a real user and the performance of the system in relation to the needs of the proxy is evaluated. One example of such a simulation is the document delivery test (Orr *et al.*[3]). In this test, 300 citations, presumed representative of the document needs of the users of a particular centre, are checked against the centre on a particular day to determine (a) how many of the items are owned, and (b) how available each owned item is on that day. A similar test has been described by De Prospo *et al.*[17]. In essence, the document delivery test simulates 300 users walking into the centre on a particular day, each one seeking a particular document. Another form of simulation is the use of a set of questions for which complete and correct answers are known to test the question-answering ability of an information centre. The set of test questions can be applied to the centre obtrusively or unobtrusively (Bunge[18], Powell[19], Crowley and Childers[4], King and Berry[20]).

On the surface, simulations of this kind may be regarded as imperfect substitutes for real life studies. But not all advantages lie with the real life situation. To begin with, simulations do not disturb the users of the system; they are also likely to be considerably cheaper than the real life study. Moreover, it could be argued that a real life study tells us only how the system performs in relation to actual *demands* and tells us nothing about the potential performance of the system in relation to the latent needs that may never be converted into demands. There are obvious dangers associated with looking at demands only (Line[21], Lancaster[22]) since the demands (expressed needs) of users are likely to be influenced by their expectations of the capabilities of the system. Evaluation of a service in relation to expressed needs, with no concern for information needs that are unexpressed, may cause managers of an information service to move that service further towards the expressed needs and further away from the unexpressed needs. A simulation such as the document delivery test, insofar as this simulation can be assumed to reflect latent needs as well as expressed needs, offers certain advantages over an evaluation that looks at expressed needs only.

A major difference between the evaluation of an operating information service and the evaluation of a system in a laboratory environment is that the latter, lacking real users with real needs, must adopt some 'ideal' as the appropriate standard for performance. But the real service should not be evaluated against an ideal but only against the real needs of users, which may be much less than the ideal. An obvious example is the use of such measures as recall and precision ratios. In the experimental environment it is reasonable to evaluate the performance of a system in relation the ideal of 100 per cent recall and 100 per cent precision. But the ideal is not necessarily the best measure to use in the evaluation of an operating system. It makes no sense to regard as a failure a search that achieves only, say, 20 per cent recall if the requester does not require a high level of recall and is perfectly satisfied with 'a few good relevant items'. In an operating environment, evaluation measures must always be related to the precise needs of the users of the service.

Moreover, evaluation measures that seem perfectly appropriate in the experimental situation may be inappropriate to the operational situation because better (e.g. more direct) measures may exist. The precision ratio is a good example of this. In effect, the precision ratio is a user cost factor associated with achieving a particular level of recall (another way of looking at it is as a form of penalty associated with the attainment of a particular recall ratio). This ratio may be the only meaningful cost factor to use in an experimental environment. But in many operating environments there may be much more direct cost factors, such as the unit cost (in $$ or time) to the user per relevant item retrieved. Thus, if a user pays $25 for the results of an online search, and finds five relevant documents in search output, he is paying $5 for each relevant item retrieved. If he conducts his own search in a printed index, and finds six relevant items in 2 hours, we could say that the unit cost (in time) per relevant item retrieved is 20 minutes. These are much more direct measures of the cost of achieving a particular level of recall than is the precision ratio.

The evaluation of an operating information service is likely to involve many more compromises than the evaluation of an experimental system. In the latter case, for example, we might be able to determine the 'true' recall ratio for a search since it may be possible to have every document in the collection judged for relevance against every request used in the evaluation. In a real operating environment, however, it is impossible to establish true recall (Lancaster[5]) and we must instead be satisfied with some method of estimating the recall ratio of the search.

There is likely to be a difference, too, between the evaluation standards appropriate to the experimental and to the operating environments. In the former, it is impossible to evaluate the results of a literature search against information needs. Since there are no real users, there are no real information needs. The best we can do is to evaluate the results of a search against a request statement (relevance as opposed to pertinence, Lancaster[5]). But this is not good enough in the operating environment. The evaluation of a search against a request statement is an artificial situation. Since we have real users with real needs we must ask these users to evaluate the results of a search in terms of the degree to which they contribute to the satisfaction of the information need that prompted the request to the system.

The problems of controlled experimentation within an operating environment have already been mentioned. In a purely experimental situation it should be possible to control all extraneous variables so that one can be quite sure of what is affecting what. In a real life environment it is not so easy to experiment. One finds oneself always compromising between the experimental design and concern for the needs of the users. For example, if a completely new information service is introduced, one that promises to be much more effective than any of its predecessors, it is difficult to explain to members of a 'control group' why they are denied use of the service. Yet, if we really want to assess the impact of the service, some type of control group of this kind will be necessary. There may always be self-selected control groups (e.g. those people who choose not to use a new service) but a self-selected group is likely to be quite different from a group that is randomly selected to form a control.

A rare example of a true experimental design in the evaluation of various approaches to the provision of information services can be found in a recent paper by Olson[23]. The design used was a 3 × 3 factorial design as illustrated in *Figure 6.1.* Two levels of 'technical information intervention' and two of

		Behavioural interventions		
		Control	Level 1	Level 2
Technical information interventions	Control			
	Level 1			
	Level 2			

Figure 6.1. Factorial design: levels of interventions (from Olson[23]).

'behavioural intervention' are recognized in the design. The study was conducted in a research and development centre of an industrial organization. Level 1 of the technical information intervention was a conventional SDI service designed to support the work of selected project teams. Level 2 was a more personalized service in which technical information staff attended regular meetings of project teams and developed targeted information 'packages' to support the work of the teams. The behavioural interventions represented deliberate attempts to enhance the existing channels of formal and informal communication within the company. The basic hypothesis was that improved productivity within the centre would be more likely to occur through the interaction of both types of intervention—that improving the formal information service was not in itself enough to achieve improved acquisition, use and transfer of information. As the design indicates, some project teams received the highest level of behavioural intervention and the highest level of information service intervention, others received various combinations of the levels of intervention, and the control group received nothing in addition to the regular services offered by the technical information centre. This is a highly complex design that presents great problems in implementation. As Olson points out:

> 'Substantial difficulties were encountered. We had underestimated the difficulties of making the role changes in the information centre. In spite of the full support of the information centre manager and the directors of research and an agreement by the staff to carry out the interventions, there was substantial reluctance. The information centre staff saw some risks to beginning an experimental process which would change the existing patterns of information flow/non-flow. We succeeded in dealing with the staff's concerns by securing additional temporary help for the centre, by assisting the staff in developing information profiles, by top-level management assurance to the staff that their full participation in the project was highly valued, by meeting with the scientists and engineers to prepare them for the new role of the information centre, and by followup coaching and support meetings as necessary.'

A useful distinction has been made, by Giuliano and Jones[24], between **proof-oriented** and **insight-oriented** studies. Experimental research is concerned with proof and rejects insight. But the manager of an operating service may be satisfied with much less than the proving of a hypothesis within acceptable limits of statistical confidence. He may be perfectly satisfied with gathering enough evidence to give him insight into a situation and be willing to make decisions on the insight alone. It is frequently possible to gain useful insights with very small studies. Much larger and more expensive studies would be needed to provide 'proof'. Moreover, proving that option A performs better than option B does not in itself explain why A is better. Even with 'proof', the manager of a centre may not be willing to adopt A, or change to it, unless he understands the reasons why it appears to give better results, and so may still rely on some measure of interpretation or insight in making a particular decision affecting the future of the system.

Giuliano and Jones discuss the distinction in the following terms:

> 'One decision which has to be made is whether the effort is to be put into

> proof-oriented experiments or into insight-oriented experiments. By the former, we mean concentration of evaluator effort into large-scale systematic and statistically valid investigations of one type of variability with other conditions being relatively fixed; by the latter we mean a spreading of evaluator effort over a number of investigative forays designed to give not proof but insight as to how the experimental variables interact with one another.
>
> A proof-oriented experiment should lead to a well-defined statement of conclusion backed up with an analysis of variance of the results and identified confidence limits. However, such experiments are premature unless one knows exactly what one wants to prove and the conditions under which the proof is interesting. It is not of much interest to know that search option A is proven to be better than search option B under given conditions with confidence 0.9999—what is really of interest is whether the given conditions *are actually realistic, how much* better is A than B, and *is this better enough* to be of real concern?
>
> Insight-oriented experiments may or may not lead to well-defined conclusions, and one such experiment may or may not be sufficiently meaningful statistically to constitute convincing proof in the face of withering doubt. However, several such tests can sometimes be performed for the cost of one proof-oriented test, and the pattern of observed results might tell a lot more about the system being investigated than any single test, no matter how firm the conclusions of that one test are.' (Page 43)

Applying the results of an evaluation

Once an evaluation has been conducted, the results must be analysed and interpreted with a view to making improvements in the service. In the application of the results of an evaluation, also, the real life situation may differ substantially from the experimental one. In the latter, it should always be possible to make changes in any part of the system and to conduct further tests to assess the effect of such changes. In the real world, however, it may not be possible to make certain types of changes even though these changes have been shown to be highly desirable. To take one example, the results of an evaluation may strongly suggest that the personnel performing the indexing for a database should also be the personnel responsible for searching that database. The evaluation results have shown many examples of disagreement or lack of communication between the indexers and the searchers. But it may be practically impossible to integrate the two activities in a particular organization. In a government agency, for instance, it may be easier to get more money than it is to get authorization to increase personnel levels. The indexing is done outside the agency, under contract, and there is no possibility of increasing the staff so that both functions are performed by the same people. In the real world, also, there may be other types of constraints that are less likely to apply to the experimental situation. An operating information service must frequently make compromises. The theoretical ideal is not always attainable in practice. Some compromises may be necessitated by the fact that the centre operates as part of a larger enterprise—perhaps a network of some kind—and it may be willing to compromise on vocabulary, record formats and other things in order to

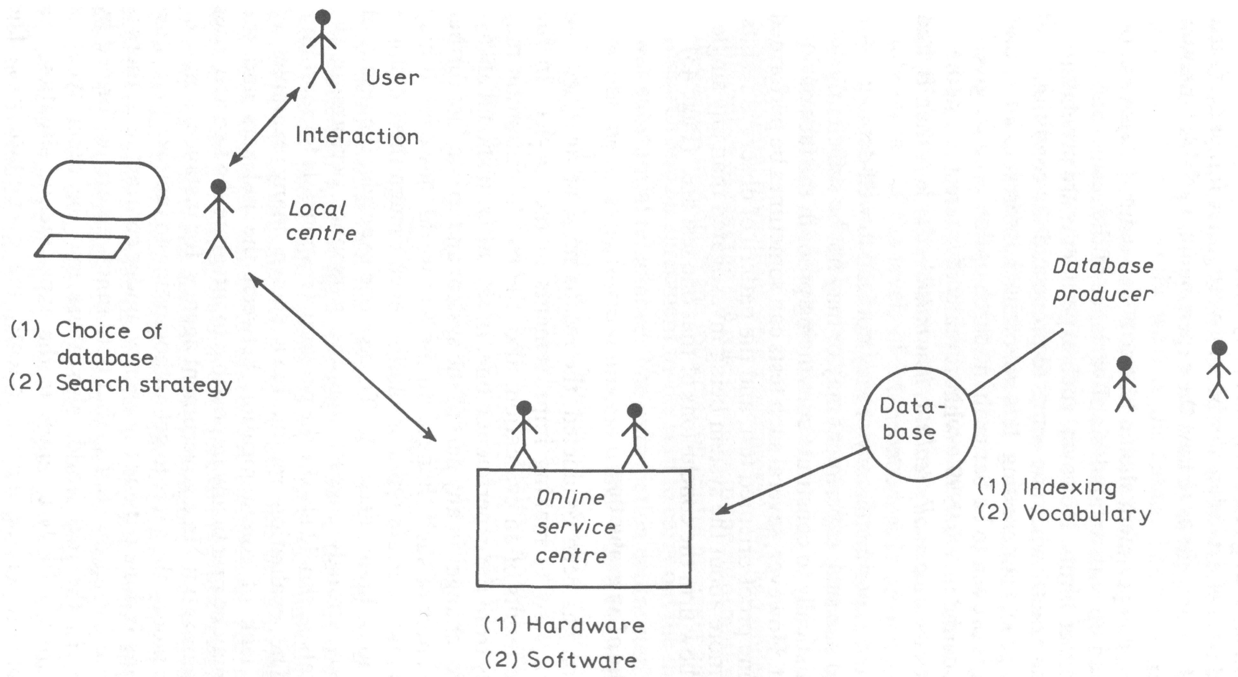

Figure 6.2. Factors affecting the literature searching performance of an industrial information centre

participate in the network. Information centres participating in an international programme, such as INIS or AGRIS, may be particularly prone to compromise. They may have to accept a thesaurus that is not ideal for their own needs and may even be required to work in English as a carrier language when they would much prefer to use Spanish or Portuguese.

Another problem in the real world is the sheer inertia of large organizations or those that have been operating for many years. A librarian may recognize the desirability of converting from one classification scheme to another but the 3 million volumes in the collection already are, to say the least, a discouragement to change. Likewise for changes in a large card catalogue. A change in cataloguing policy or practice is not only difficult to implement, but it will be many years before, applied to the catalogue as a whole, the change will have any significant effect. Another barrier is simply that of the cost of change. Change may be inexpensive in the experimental system but the cost of certain changes may be prohibitive in an operating service.

The experimenter in information retrieval is in an unusually fortunate position in that he/she will frequently have control of all the factors that affect the performance of the experimental system. Not so in the real world. It is rare for an information centre to have control over all the factors affecting the service it provides to its users. Consider, for example, an industrial library in the United States that is conducting online literature searches for its users (*Figure 6.2*) through some online service centre such as Lockheed. It can be seen from the diagram that the industrial information centre has direct control of only a few of the factors that determine the effectiveness of the services it provides. It has control over the way in which it interacts with its own users. To a very large extent it has the ability to influence the quality of interaction through a redesigned search request form, through improvements in the training of the information staff—for example, in the conduct of request interviews—or through insistence on a greater level of user involvement in the search process itself—for example, asking the user to be present at the terminal at the time the search is conducted. But the information centre does not have absolute control over even this stage of the information retrieval process, because it does not have complete control over its own users. The staff of the information centre can go only so far in leading and helping users. If a particular individual, for one reason or another, is completely unable to verbalize his information need, there may be very little the information specialist can do to help him.

Similarly, the local centre may be said to have control over its own search strategies. It does, of course, but this control is not absolute either. The characteristics of a search strategy are very largely determined by the characteristics of the software used to search online. The individual searcher must operate within the constraints of this software. For example, the extent to which he can use truncation in searching is obviously completely controlled by the truncation capabilities of the query language. Obviously, too, the search strategy must make use of the vocabulary of the database, and it must adapt to the indexing policies in use in the database, and these important factors affecting performance are under the control of only the database producer.

In fact, then, a typical information centre has control over only a few of the factors affecting the quality of the information services it provides. This situation, however, is not so gloomy as it first appears. The local information service has substantial control only over its approach to user–system interaction and the construction of search strategies, but these are in some ways the most important factors to control. They are most important because they are the factors that occur first in the whole operation. If a user's request statement is an inadequate representation of his real information need or if a search strategy is a very imperfect representation of a request statement, the search is virtually doomed to failure. In this case, it matters little whether the indexing is sufficiently exhaustive, whether the vocabulary is sufficiently specific, whether the query language is sufficiently flexible, and so on. Vocabulary, indexing, and other database characteristics may be close to perfect, but this does not help if the search strategies are seeking unwanted information.

There is a second reason why control over the processes of user–system interaction and search strategy construction can be considered particularly important. Changes made to these operations can take immediate effect. If today we introduce improved methods of interacting with our users or of constructing strategies, we can expect that the overall effectiveness of our services will improve immediately. But changes to a database tend to produce long-term rather than immediate effects, at least for the retrospective search situation. Changes made to indexing policy, or to the index language, are not going to have a very pronounced effect for some time. Consider a database of 500 000 documents growing at the rate of 100 000 items per year. Even if sweeping changes were now made to the vocabulary or indexing policies, it would be another five years before these changes would affect even half the total database.

Occasionally, of course, an information centre may have complete control over the entire situation: it develops its own database, applies its own hardware and software, prepares its own search strategies and interacts directly with its users. Such a situation is rare. An information centre that is completely in control of its own performance is in a fortunate position indeed.

References

1. ASHMOLE, R. F. *et al.* Cost effectiveness of current awareness services in the pharmaceutical industry, *Journal of the American Society for Information Science* **24**, 29–39 (1973)
2. DAVISON, P. S. and MATTHEWS, D. A. R. Assessment of information services, *Aslib Proceedings* **21**, 280–283 (1969)
3. ORR, R. H. *et al.* Development of methodologic tools for planning and managing library services, *Bulletin of the Medical Library Association* **56**, 235–267 (1968)
4. CROWLEY, T. and CHILDERS, T. *Information Service in Public Libraries: Two Studies*, Scarecrow Press, Metuchen, N. J. (1971)
5. LANCASTER, F. W. *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 2nd edn, Wiley, New York (1979)
6. MASON, D. PPBS: application to an industrial information and library service, *Journal of Librarianship* **4**, 91–105 (1972)
7. MAGSON, M. S. Techniques for the measurement of cost-benefit in information centres, *Aslib Proceedings* **25**, 164–185 (1973)
8. MARTYN, J. Unintentional duplication of research, *New Scientist* **21**, 338 (1964)
9. MCDONOUGH, A. M. *Information Economics and Management Systems*, McGraw-Hill, New York (1963)
10. ROSENBERG, K. C. Evaluation of an industrial library: a simple-minded technique, *Special Libraries* **60**, 635–638 (1969)

11. KRAMER, J. How to survive in industry: cost justifying library services, *Special Libraries* **62**, 487–489 (1971)
12. MUELLER, M. W. *Time, Cost and Value Factors in Information Retrieval*. Paper presented at the IBM Information Systems Conference, Poughkeepsie, N. Y., September 21–23 (1959)
13. KING, D. W. and BRYANT, E. C. *The Evaluation of Information Services and Products*, Information Resources Press, Washington, D.C. (1971)
14. DRUCKER, P. F. Managing the public service institution, *The Public Interest* **33**, 43–60 (1973)
15. LANCASTER, F. W. *Evaluation of the MEDLARS Demand Search Service*, National Library of Medicine, Bethesda, Md (1968)
16. LESK, M. E. and SALTON, G. Relevance assessments and retrieval system evaluation, *Information Storage and Retrieval* **4**, 343–359 (1968)
17. DE PROSPO, E. R. *et al. Performance Measures for Public Libraries*, Public Library Association, Chicago (1973)
18. BUNGE, C. A. *Professional Education and Reference Efficiency*, Doctoral Thesis, Graduate School of Library Science, Urbana, University of Illinois (1967)
19. POWELL, R. R. *An Investigation of the Relationship Between Reference Collection Size and Other Reference Service Factors and Success in Answering Reference Questions*, Doctoral Thesis, Graduate School of Library Science, Urbana, University of Illinois (1976)
20. KING, G. B. and BERRY, R. *Evaluation of the University of Minnesota Libraries Reference Department Telephone Information Service*, Pilot Study, University of Minnesota, Library School, Minneapolis, ERIC Microfiche ED077517 (1973)
21. LINE, M. B. The ability of a university library to provide books wanted by researchers, *Journal of Librarianship* **5**, 37–51 (1973)
22. LANCASTER, F. W. The tip of the iceberg, *Bulletin of the American Society for Information Science* **4**, 32 (1978)
23. OLSON, E. E. Experiments to improve information transfer and the effectiveness of R & D in industry. Paper presented at the Annual Meeting of the American Association for the Advancement of Science, Houston, January (1979)
24. GIULIANO, V. E. and JONES, P. E. *Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems*, Arthur D. Little Inc., Cambridge, Mass. (1966)