

## Selected Papers

### Fifty Years After – Almost

#### F. Wilfrid Lancaster

I began my library career in 1949. Being associated with libraries for more than 60 years does give one a somewhat unique perspective on the profession. In this talk, I will look at changes in the way library and information services have been evaluated within the context of changes in the services themselves within the broader context of changes in the environment in which these services have operated.

The library where I began my career was a public library in the north of England. It was a unique experience because it was the last public library in the country to remain a closed-access library. Using this library was rather like using a research library such as the Library of Congress. Users had to find call numbers for books in the public card catalogs or in library book lists and present these at a service desk. A staff member looked for the books and presented them to the user, for borrowing, if they were found. If they were not found, of course, the user had to go through the process again, a very frustrating experience.

Looking at this library service as an evaluation problem, it is clear that one part of the service, the document delivery function, would have been easy to evaluate – a user either left the library with a book, or books, he requested or he did not. The prerequisite evaluation question – was the user able to find what he was looking for in the catalog – would have been almost impossible to deal with.

The library converted to a conventional open-access library a year or so after I began work there. This undoubtedly made users happy. They could browse the shelves and find their own books. Of course, evaluation had become more complicated. There was no easy way of knowing whether a user found what he wanted in the library or not. A book borrowed could be considered a kind of “success” but how many people came to the library and left without finding what they were looking for?

Library and information services have changed very much in the sixty years since I entered the profession. From my perspective, the most fundamental change has been the relentless move by managers of library and information services to make users do things for themselves. As libraries have become increasingly self-service institutions, their activities have become increasingly difficult to evaluate. Moreover, I believe that the managers of these services have, at the same time, become less and less concerned about evaluating them.

Librarians were not much concerned with evaluation back in 1949 either. If you had raised the evaluation issue at a professional meeting back then, the audience would probably have laughed at you. Librarians did not evaluate. They did not need to evaluate. Libraries were universally accepted as “good” for the communities they served, so why did they need evaluating?

Certainly, librarians did quantify. They knew how many books were borrowed, how many reference questions were received and answered, perhaps how many people entered the library. But quantification is not the same as evaluation.

---

*New Trends in Qualitative and Quantitative Methods in Libraries*

A. Katsirikou and C. H. Skiadas (eds)

© World Scientific Publishing Co (pp. 1-7)

NEW TRENDS IN QUALITATIVE AND QUANTITATIVE METHODS IN LIBRARIES - Selected Papers Presented at the 2nd Qualitative and Quantitative Methods in Libraries - Proceedings of the International Conference on QQML2010  
<http://www.worldscibooks.com/compsci/8144.html>

And the statistics collected represented only events presumed to be successes. Nothing was known about possible failures.

Of course, not all librarians of the 1940s, or even earlier, lacked interest in quantification or scientific management. Most obviously, S. C. Bradford, a librarian at the Science Museum in London, had already made a major contribution to the field that we now know as “bibliometrics” with his pioneering paper on the scattering of periodical articles over periodical titles. This was first published in an engineering journal in 1931 but it appeared again in his book Documentation in 1948. Wyndham Hulme, another British librarian, had virtually invented the field of bibliometrics, which he called “statistical bibliography,” as early as the 1920s.

Although librarians were not much concerned with evaluation in the 1950s, evaluation criteria did exist. The most obvious were the Five Laws of Library Science, first put forward in 1931 by the Indian librarian Ranganathan. These “laws” were:

1. Books are for use
2. Every reader his book
3. Every book its reader
4. Save the time of the reader
5. The library is a growing organism.

These may seem rather superficial at first but, when you study their implications, they are actually quite profound because they offer essential criteria for the evaluation of library services from the perspective of their cost-effectiveness, as well as their effectiveness, and even point to important cost-benefit considerations.

Ironically, interest in the objective evaluation of library services may well have been stimulated by work on the objective evaluation of information retrieval systems, which began towards the end of the 1950s. The pioneering study was the ASLIB Cranfield Project, with which I was associated for a period of time. This was really a study of indexing systems – i.e., methods of representing the subject content of documents – and was initiated originally to compare the effectiveness of different methods for arranging entries for books in subject catalogs. While this project, which continued over a period of years, yielded many interesting results, it was probably most valuable for highlighting two basic criteria for evaluating the results of a search in any database, recall and precision, the extent to which a search finds the material you were looking for and the extent to which it is able to avoid retrieving a whole lot of material you were not looking for. The ASLIB Cranfield Project had no real information service to evaluate. It was based entirely on simulation. The research questions for which searches were performed were actually based upon documents known to offer good answers to the questions and known to exist in the database.

The Cranfield study occurred at a very opportune time. Computers were just beginning to be used in information retrieval applications and some systems and services of significant size began to emerge. It was natural that the evaluation criteria identified at Cranfield should be applied to these new systems. In the United States, The National Library of Medicine took the lead here. Its Medical Literature Analysis and Retrieval System (MEDLARS) was established in the 1960s and I was fortunate enough to be hired by the Library to undertake an evaluation of its service. Unlike the Cranfield studies, this was not a simulation. It was an evaluation of a real service – computerized searches of a large biomedical database performed to satisfy the information needs of biomedical professionals. The study was unique in its size: the results of 300 searches were evaluated by the biomedical professionals requesting the searches. In those days, computer searches were performed offline. The requesters received printouts of the bibliographic references retrieved by the searches performed

for them. The 300 requesters participating in the study evaluated the search results from these printouts and from a random sample of the actual biomedical articles, referred to in the printouts, supplied to them by the library. The results of this evaluation were published in 1968.

At about the same time, the National Library of Medicine also made a major contribution to the development of evaluation methodologies for more conventional library services. The Library was engaged in establishing a Regional Medical Library Program. This involved selecting libraries in different parts of the United States to receive contracts to provide various types of service essentially on behalf of the national library. To help them select these regional representatives, the Library gave a contract to the Institute for the Advancement of Medical Communication to develop criteria and procedures for the evaluation of various basic library services. These investigators came up with a number of valuable evaluation tools. One was a type of inventory which could be used to compare the range and scope of services offered by various medical libraries. The libraries received a numerical point score based on the range and scope of their services, a perfect score being 1000 points. The procedure was quite discriminating. For example, when tested on six academic medical libraries, it produced a high score of 721/1000 and a low score of 533/1000.

An even more important outcome of this contract was the “document delivery test.” This simulation tested the ability of academic medical libraries to satisfy the document needs of users. The evaluators enter the library with a sample of 300 bibliographic references for items that library users could reasonably be expected to be looking for at that time. Each item is given a score reflecting immediacy of availability. The best possible score is given if the book or article is immediately available on the shelves; the worst score for the item is given if it is estimated that it could not be available in less than a week (e.g., by borrowing from another library). Based on the sample of 300 items, a capability index for the library could be derived, and this could be used to compare different libraries.

The document delivery test pioneered the objective evaluation of the most basic of library services – its ability to deliver printed items to users at the time these items are needed. The description of this test in the literature, in 1968, inspired a great many related studies in the next two decades. Some of these were very similar simulations, modified for applicability in different types of libraries. Other studies were based on a kind of user survey. Typically, a user entering the library was given a form on which to record details of the items he was looking for on that visit. For each item recorded, he was also asked to indicate success or failure: could he find a call number for it in the catalog and, if he could, was he able to find it on the shelves. Library professionals could later check the failures to determine whether they were catalog use failures or “shelf failures.” For items known to be in the collection, a determination could be made of where the item was at the time the user was unable to find it.

The late 1960s and early 1970s also saw the beginning of attempts to evaluate the ability of library reference services to answer factual-type questions completely and correctly. These studies were usually performed by having volunteers pose questions for which the answer was already known and documented. The overwhelming majority of these studies were performed with public libraries and with the questions posed by telephone. Much later, in 1991, I was heavily involved in a study that evaluated the quality of reference service in several departmental libraries in a large academic institution. In this case, we trained students to walk into these libraries with their test questions and to record how the librarians approached the

question, what sources they used, what answer they came up with, and the time elapsing.

By the early 1970s, the searching of databases by computer was already beginning to move from an offline to an online mode of operation. The evaluation of online database searches was not significantly different as long as the librarian was still performing the searches on behalf of the user (i. e., they were delegated searches). The nondelegated search, the situation in which a library user performs an online search for himself, is a completely different proposition. The most obvious problem is that of finding out what the user does in the searches and to what extent the user finds the information sought. Even the criteria most often used to evaluate the delegated search, recall and precision, were not really appropriate to the nondelegated situation because these were really secondary measures of success rather than primary measures. The precision ratio was an indirect measure of user effort. In the nondelegated search, the more direct measure was obviously how much time the user spent on the search. The recall ratio was not so relevant either because a user would naturally stop searching when he felt he had found what he was looking for or, alternatively, gave up the search as a lost cause. Nevertheless, the evaluator could not afford to neglect the fact that a search may have missed items that would be much more valuable to the user than those actually found. As online searching developed, it became possible to do some unobtrusive monitoring of use. Online monitoring could be used to discover, for example, what terms, in what combinations, were used in the search, which records were retrieved, and which of these were selected in some way by the user, but such monitoring could not determine the value of the search to the user. This could only be determined by asking him. Even this was not fully satisfactory because the user could only judge success on the basis of what was retrieved, knowing nothing of what was not retrieved. In some cases, the items missed could make those found virtually superfluous. To fully evaluate the success of a nondelegated search really required searches on the same topic to be performed by search specialists with the results compared with the user results and any items not retrieved by the user submitted to him for his evaluation. This was a tedious evaluation process and not completely satisfactory because of a time lapse – an item that might have been highly valuable to a user on January 15 may have no value at all to him on January 20.

The evaluation of library and information services in the 1960s and 1970s was mostly concerned with their effectiveness – the extent to which they were able to satisfy user demands. By the 1980s, however, those responsible for the funding of libraries became increasingly concerned about the costs of these services. This led to new levels of evaluation. Cost-effectiveness evaluations tried to relate the effectiveness of services to their costs. This led to new evaluation criteria such as the cost per item borrowed, the cost per item consulted in the library, and the cost per useful item retrieved in a database search. Some studies began to look at the cost-effectiveness of different methods for the delivery of information services. By the 1990s such studies included comparisons of the cost-effectiveness of providing service from printed versus electronic journal collections.

By this time, too, the very existence of library and information services was threatened, especially in industrial, government and international organizations. Some libraries and librarians found themselves having to justify their existence. Cost-benefit analysis, undertaken to prove the worth of the service to the community, had existed for some time, but studies of this kind became more common in the 1990s, and the approaches used became more sophisticated. The usual approach was to estimate what would happen to the organization if the information service was

eliminated – for example, how much it would cost to obtain needed information in other ways. Outstanding cost, cost-effectiveness and cost-benefit analyses were performed by Griffiths and King, mostly in the industrial or government library environments. Their most important analyses looked at the benefits of exposure to information through the library in terms of the impact on the individual of being deprived of this service – loss of productivity, duplication or other waste of effort, and time and costs of obtaining needed information elsewhere. When costs were calculated for such losses, they were able to conclude that potential savings to the organization associated with exposure to library services could exceed the costs of providing these services in ratios ranging from a low of 7.8:1 to a high of 14.2:1.

In the 1990s, library associations in the United States began to concern themselves more actively with the assessment of the quality of the services provided by libraries. One leader was the Association of Research Libraries, which spearheaded the application to libraries of a measure of quality already in existence in the business community. The method, which is now well-known and has been widely applied, involves the measurement of the gap between the service levels expected by library users and the service levels they perceive to exist. Later this methodology was modified to apply more clearly to libraries in digital form,

From the time I first got involved with information service evaluation, in 1962, I have strongly believed that the main purpose of such activities is diagnostic – to identify problems and failures in a particular service and to identify ways to improve the situation. Probably the most important evaluations of this type are those that are able to document information failures and the results of these. Such studies fall into two broad categories: (1) undiscovered public knowledge, and (2) unintentional duplication of research.

Don Swanson used the terms “disconnected” or “noninteractive” to refer to two bodies of literature that are unknown to each other and not linked by conventional bibliographic means (e.g., not indexed in a similar way in databases, not citing each other and not citing a common antecedent literature). Clearly, most disconnected literature pairs are completely unrelated in the sense that the finds of one research area have no possible relevance to the other. In some cases, however, two disconnected literatures may interrelate (i.e., literature A may make a contribution to research area B, B to A, or both). Swanson refers to such literatures as “complementary.”

He performed extremely valuable pioneering research on disconnected biomedical literatures for several years. In some cases he was able to discover scientifically significant connections that were previously unknown. For example, one body of literature describes how dietary fish oil may bring about certain changes in properties of the blood while another suggests that changes of this kind could be beneficial in the treatment of Raynaud’s disease. Yet, the “fish oil” research (literature) was disconnected from the Raynaud’s disease research area. In another example, the literature on magnesium metabolism was shown to have potential relevance to migraine research: magnesium deficiency can bring about neurological changes that may lead to a migraine attack.

There is also much evidence to suggest that the amount of undesirable duplication in science research is not inconsiderable. The best study of its kind was that conducted by John Martyn in 1964 in England. In an investigation of 647 current research projects in government, industry, and academia, Martyn gathered documented evidence of 43 cases of unintentional duplication of research, and 106 cases in which information discovered from the literature, while research was underway, would, if found earlier, have saved time, money, or effort in the research project. Martyn found that, in about 9% of all the projects studied, money could have

been saved through an improved awareness of research reported in the literature. An extrapolation to the total expenditure for scientific research in the United Kingdom in 1962 led Martyn to conclude that at least 6 million pounds a year could be saved in U.K. research through more effective use of the literature. This was likely to be a very low estimate, however, because it was based only on cases of duplication or suboptimum approaches to research discovered by the scientists through literature searches that were very probably quite incomplete. Martyn hypothesized that this was only the tip of an iceberg and that much greater waste would be uncovered by more exhaustive searches of the literature. More than 20 years later, Martyn repeated his study, using the same survey methodology. The later investigation revealed that, although literature searching had been greatly facilitated through widespread online access to databases, and researchers were more aware of the importance of such searching activities, the number of cases in which they discovered relevant information too late to be of maximum value to them seemed to be increasing rather than decreasing.

Returning more specifically to the evaluation of library services, I believe that – with the exception of selected work in the areas of cost-effectiveness and cost-benefit evaluation - this field has taken steps that have been mostly backward since the 1980s. Rather than performing diagnostic evaluations of specific services in a single library, the profession preoccupied itself more with comparisons – usually one or more libraries against some norms – and with general subjective surveys of user satisfaction. Using interviews or questionnaires to determine attitudes towards library services in general has much less value than studies that determine the success or failure of the service on a particular incident (the critical incident). Although library user contributions to a general impressionistic survey may give anecdotal information on particular events, this is not a true critical incident study and has little diagnostic value. People best remember the events that are either exceptionally satisfactory or exceptionally unsatisfactory.

To consider what has happened to the evaluation of library and information services in the last decade or so, we need to consider what has happened to library services in this period, and even what has happened to society in general. What has happened in society, of course, is that technology has been used to replace public service. Technology allows us to do things for ourselves that we used to expect others to do for us. The most pernicious aspect of this new self-service society is that technology replaces people. When I make a telephone call to a business or government agency, the chance of actually talking to a human being is decreasing rapidly from year to year. Usually I am subjected to a computerized menu, and computerized voice output, with categories that never seem to match my needs. My satellite television provider even expects a menu of computerized output to diagnose my reception problems. The business and other entities that provide such facilities seem to consider that my time is free. My time is not free. If I were not wasting my time on fruitless telephone calls, I could be doing something more productive – like watching soccer on television. The fact that these entities consider my time to be free is all part of a more general malaise, the fact that public service is a thing of the past. Corporations and government agencies no longer care whether the people they supposedly serve are served or not.

Library and information services in general share in this social malaise. In my experience, only public libraries retain any vestige of public service. Libraries in academia, government and industry have increasingly used technology to replace service. There exists a pervasive assumption that using technologies to empower users – to access databases for themselves, to obtain materials from other libraries, to

check out books for themselves, and suchlike, means that things are now better for library users than they once were. Moreover, library managers, just like managers of other enterprises, no longer care.

In the library and information service arena, evaluation is no longer a matter of concern or even interest. Take a look, for example, at the Annual Review of Information Science and Technology. Nine of the ten issues from 1966, when it was first published, to 1975 contained chapters dealing in some way with evaluation. In the next decade, 1976-1985, five of the ten issues contained evaluation-related chapters, and the following decade, 1986 to 1995, saw only three. There has not been an evaluation-related chapter since 2001, and now the term “evaluation,” or its synonyms or related terms, rarely even appears in the indexes.

But perhaps the compilers of these indexes deserve considerable credit. In actual fact, there have been very many studies in the last decade that claim to be evaluation studies. The indexers just felt that they did not deserve to be given the evaluation label. This is because they usually collect rather gross data through web log analysis or, alternatively, collect general impressionistic data from library users that produce results such as (a real example):

Undergraduates in all ten colleges were reasonably satisfied with the library.

Faculty overall were dissatisfied with information control but reasonably content with service affect and facilities.

Undergraduate students used the physical library facilities much more frequently than did faculty.

Is this evaluation? I don't think so.

In the good old days of evaluation, we first made a list of all the questions we wanted to answer about a service. Then we set about designing a study that would answer these questions. Today, it seems, those claiming to be evaluators have a different approach. The main question of their concern is “What data can we collect in large quantities as automatically and painlessly as possible?” Gross data lead to gross conclusions.