Chapter 3.

METHODOLOGY

The obvious methodology for the comparison of two distinct systems would have been matched pairs - performing the same search twice, once on each system. Jamieson and Oddy, however, suggest that there might be problems, in particular the "learning effect". This is caused by the user being present at the searches on both systems, and his or her perceptions and relevance judgements of the second search being influenced by what had been learnt in the execution of the first search. An additional problem might be that the learning effect is unlikely to be symmetrical between the systems and randomising the order in which the systems are used would not necessarily provide a reliable way of eliminating this effect. Therefore it was decided that a comparison of the two systems would be based on independent samples, i.e. for each query in the experiment, a random choice would be made as to which system to use. This would necessitate a fairly large sample size in order to satisfactorily determine the significance of any observed differences.

3.1. Sample size

The following discussion of sample size starts from the arguments used by Jamieson and Oddy in their proposal (6). As we shall see below, however, there are some problems with the Jamieson/Oddy argument.

We want to compare two systems, on two independent samples of requests (each of size N), on various measures or variables as discussed below. We propose further to use a statistical significance test: following Jamieson and Oddy, the suggested test is that using the Mann-Whitney U statistic. Suppose, then, that the distributions of values of the measure in question for the two systems are of similar shape, but differ in location (and in particular in mean) by d. Then we can calculate the expected value of U, as a function of d, N and perhaps some other parameters of the distributions. Given also a chosen level

of statistical significance, we can determine whether the experiment is expected to yield a correct and significant result, i.e. whether the sample means will be in the right order and the U value will be in the significance range. We can, indeed, use the relation to determine a minimum value of N to achieve this result.

Of course we do not know d in advance. Therefore we have to start by specifying the minimum d we wish to be able to detect (if there is little difference between the systems, a large sample size will be required). Further, other distributional parameters are likely to be involved. Jamieson and Oddy discuss d in relation to the mean value m for the conventional system: they set themselves the target of detecting a difference correctly if d/m exceeds 5%. Choosing also a significance level of 5% (95% confidence), and assuming a particular shape of distribution, they come to the conclusion that a minimum N of 190 (for each of the two samples) is required. In our proposal for this project, we appealed to that argument, and therefore set ourselves a target of 500 searches in total (allowing some leeway). For a variety of reasons discussed below, we have fallen far short of this target.

It now appears, on closer examination, that there are problems with the Jamieson/Oddy argument. One problem lies in the use of d/m as a parameter: a little thought shows that d should be related to the spread or standard deviation rather than the mean, and it turns out that in the specific analysis they actually used something other than d/m. A second problem lies in the choice of distribution and a third lies in the assumption of a continuous variable. Unfortunately, these problems conspire in the wrong direction: to suggest that the sample size required may be larger than Jamieson and Oddy claim. A technical analysis of the problem is presented in Appendix A2.

3.2. Variables

The variables to be examined are divided into eight categories and will be evaluated by the questionaires, logs and relevance evaluations. The eight categories include :

3.2.1. Retrieval effectiveness

This is based on relevance judgements on a maximum of 50 full format offline prints, made quite separately from the online evaluation for relevance feedback purposes, which was usually based on titles alone. Relevance judgements are made on a three-point scale, ie. "Relevant", "Partially relevant", "Not relevant". Relevancel indicates

that the middle group has been classified as not relevant. Relevance2 that it has been included with the relevant group.

The parameters to be evaluated will be: total number of items retrieved; number of relevant items retrieved; and precision; the latter two being calculated both for Relevance1 and for Relevance2.

Because of the difficulties presented earlier with match pair evaluations, and because the experiment is being conducted in an operational environment, there is no recall base. Originally a subsequent "broad search" executed after the initial search was suggested in order to provide this information, but the "broad searches" were abandoned and substituted with a limited number of matched pairs. The matched pairs were designed to help acquaint intermediaries with the new system (by providing a comparison between Boolean and Weighted techniques), and as a backup if the results on one system proved inadequate or it was thought better results could be achieved on the other system.

Additional relevance data was gathered during the evaluation of offline prints. This information was concerned with whether the user had seen the documents prior to the evaluation. For the purposes of this experiment this information will not be analysed, but may prove useful in future projects for determining novelty ratios.

3.2.2. User effort

This is based on interaction between the intermediary and user at the presearch stage, and between machine, intermediary and user during the search. It includes: time to prepare the search, presearch terms, number of terms added or amended during the search, online time, online citations and number of relevance judgements made.

3.2.3. Cost

Essentially connect time, pss packets, on and offline citations. Items contributing to the overall cost such as postage and overheads will not be considered.

3.2.4. Subjective user reactions.

Mostly concerns the user's and intermediary's overall reactions to the search, impressions of effort involved, and reaction to search results - not offline prints. This also includes variables from other categories such as how close was the search to the original/intended

enquiry; was the expected number of references retrieved and the intermediary's contribution.

3.2.5. User characteristics

Personal data about the individual which would relate to the search process such as areas of subject expertise, the level of their work and number of previous online searches either with or without an intermediary.

3.2.6. Request characteristics

Subject area of the request, nature of enquiry (e.g. accurate or vague), type of search required (i.e. broad or narrow), and number of presearch terms.

3.2.7. Intermediary's contribution

Time taken in preparing the search, terms added or amended during the search, number of relevance judgements, intermediary's assessment of the difficulty of the search and the user's asssessment of the intermediaries contribution.

3.2.8. Search process characteristics

The idea behind this category was to investigate the circumstances under which the two systems perform differently, and to try to isolate the factors influencing these differences in the search process. The scope for such investigations in the present project was very limited; nevertheless, some questions in this category were included in the questionnaire. The questions concerned with the search process asked for subjective impressions of the search process from both user and intermediary, e.g. why the search was terminated; the number of terms added or amended during the search; and an assessment of the effects resulting from judging the relevance of titles during the search.

3.3. Data collection instruments

3.3.1. Questionaires

The questionnaires provided a qualitative assessment by the user and the intermediary of a range of variables. There were three questionnaires: two completed by the enquirer and one by the intermediary. In addition there was an introductory form briefly explaining the project, what would be required of the user and stressing

that all the information given was strictly confidential and data protected. The questionnaires are reproduced in Appendix A7.

The first questionnaire given after the presearch interview dealt mostly with user and request characteristics. It enquired about contact details such as name, address and telephone number. It then proceeded to ask for combined user and search information such as status (post-graduate, consultant, researcher, etc.); what they hoped to use the search results for; whether they had done any online searches before, either on their own or with an intermediary. There were also two questions regarding request characteristics: firstly whether the user wanted a broad or narrow search, and secondly whether they viewed the nature of their search request as precise or accurate, general, vague or waffley. This last question was trying to ascertain how the request and the terms used in the search related to the subject domain of the query.

The next two questionnaires were completed immediately after the search, while the user and intermediary were still sitting at the terminal. Remaining at the terminal became important because the user often asked how could she/he answer these questions without first seeing the offline prints. It was explained that this questionnaire was concerned with the search process and the offline print results would be separately evaluated at a later stage.

These two questionnaires were concerned primarily with the user's and intermediarie's overall satisfaction with the search, the amount of effort involved and a consideration of the results based upon what they had seen during the search. The user was separately asked how close was the search to the original enquiry and did they retrieve the number of references they had anticipated. The last two questions for the user were conditional on whether a Weighted or Boolean search had been allocated. Their object was to determine the influence of relevance feedback to the search process. The user was asked whether they had seen any titles (or references) displayed during the search, and if so did viewing these titles appear to make the search more effective or change the course of the search from the information supplied.

The intermediary, on the other hand, was asked three separate questions relating to request characteristics, user effort and search process. The questions asked were "what was the number of presearch terms", "how long did it take to prepare the search" and "what was the reason for finishing?"

3.3.2. Evaluation of offline prints

Offline prints were obtained for the purpose of relevance evaluation for the effectiveness measurement. In the case of the Boolean searches, whichever sets were finally selected as the results of the search were used; in the case of the Weighted searches, any sets selected for offline printing plus any documents viewed online and judged relevant at that stage were included. In either case, each user was asked for a maximum of 50 offline prints; if the combined set was larger than 50, then the first 25 and the last 25 were used.

The primary consideration here was whether or not the references retrieved by the search were relevant. In order to give the user as much information about the reference as possible it was decided to supply full format offline prints which included all the information provided by the data-base producers. The evaluations were to some extent complicated by the user's previous knowledge of any of the retrieved documents. An attempt was made to overcome this problem by asking two questions about each reference to be evaluated. Firstly "From the information given is the document an answer to or about your subject query" to which the user would reply "Yes", "Partially" or "No". Secondly, the user was asked to select any one of the following four categories which best applied to the document under consideration. The categories were:

1) Seen the document itself before and it was useful
2) Seen the document itself but it was not useful
3) Have not seen the document represented by this reference but would like to see it
4) Have not seen the document represented by this reference and would not like to see it

Therefore each had two responses, a letter Y, P or N and a number 1-4. The evaluations were marked directly on a separate copy of the offline prints. As indicated above, the number was restricted to 50; it was thought that this woudl ensure that the time taken to evaluate the prints would not prove to be either prohibitive or offputting. The user was subsequently given their own complete copy of the offline print set to keep.

3.3.3. The Logs

Complete logs of all searches were kept automatically by Cirt. In the case of weighted searches, both the user / Cirt interaction and the Cirt / Data-Star interaction were logged.

The logs provide the quantitative assessments such as number of terms, number of online relevance judgments, number of online/offline prints, pss packets and online time etc.

There are two types of logs, the logs of the searches and the logs of all the communication between the front-end and the host. For Boolean searches the two logs are virtually identical. Weighted searches, alternatively, have two totally different logs: the search logs print out the searcher's transactions with Cirt, and the net logs show Cirt's transactions with Data-Star.

Examples of the logs are in Appendix A3.

3.4. The participants

Originally searches for the experiment were intended to come from the University of London Central Information Services and from other institutions in the University. The institutions which could take part were limited to those which had a connection to JANET and whose areas of expertise were compatible with the databases available on Cirt i.e. Medline, Psychological Abstracts and Inspec. Because of the decline in the number of searches at CIS, the project became reliant on other University of London Institutions, more particularly Medical Schools. It was decided to approach these schools at intervals and in limited numbers because of the time involved in getting them operational, and the possibility of greatly reduced systems efficiency resulting from more than two intermediaries using the system simultaniously. CIS provided the information relating to the number of searches per annum for the medical schools, and the heaviest users were approached first, in the hope that they would be able to contribute a minimum of 50 searches to the data set.

Eleven libraries were contacted. They were:

1. Charing Cross and Westminister Medical School Library
2. The City University: Skinners Library
3. Guys Hospital Medical School: Wills Library
4. Guys Hospital: Paediatric Research Unit
5. Imperial College: Lyon Playfair Library
6. Institute of Cancer Research: Royal Cancer Hospital Library
7. London School of Hygiene and Tropical Medicine
8. Middlesex Hospital: Medical School Library
9. St. Bartholomew's Hospital: Medical School Library
10. St. George's Hospital: Medical School Library
11. St. Thomas' Hospital: Medical School Library

Of the eleven libraries contacted, six showed willingness to

participate. Of those six only St. Bartholomew's, St. George's, and Imperial College have in the event furnished us with data.

The inability of the participating intermediaries to substantially increase the data set became clear in the summer of 1986. Only 18 searches had been collected since February. At this stage the methods for data gathering were reconsidered and it was decided to alter the strategy and offer free searches as well as subsidised searches.

The free searches would require the user to come to the Information Science Department at The City University for two appointments, one for the search, the other to evaluate the offline prints. The free searches were performed by the project information scientist. Advertising was distributed in August and September (see Appendix A4). The first completed data set was added on the first of October 1986. Complete returns include three questionnaires, search logs and evaluated offline prints. This method proved very successful and has provided the additional benefit of reduced wastage. The offline prints are evaluated on site one week after the search thereby reducing the possibility of users not returning them if the prints are taken away for evaluation.

3.5. Procedure for data collection

Once the intermediaries had decided that they were willing to participate each one was given a copy of the Cirt Users Manual (see Appendix A5). All the intermediaries had a tutorial session with the project information scientist and one or two searches on their own in order to become first acquainted and then confident with Cirt.

There was a list of guidelines to help the intermediaries and provide a consistent method for data collection. When a user came to an intermediary asking for a search the intermediary would make sure three criteria were met: firstly that it was a subject search (as opposed to an author or source search); secondly that the user would be present during the search; and thirdly that the enquiry suited one of the databases available on Cirt.

The users were then given the introductory form containing a short explanation of the project and asking for their cooperation, which they were asked to sign. Hopefully by signing and dating the bottom the users realised their commitments and were more likely to completely fulfill them.

The intermediary then drew a random allocation card. This was a set of cards (see Appendix A6) half of which were allocated to Boolean and

half allocated to Weighted. They were shuffled to create a completely random order and then numbered sequentially with a query id. This id provided the means of keeping all the data (questionnaires, logs, and offline prints) for each query together and easily distinguishable. The card method of allocation was decided upon so that a check would be kept on the number of Boolean and of Weighted searches done and each type of search would be allocated an equal number.

The next step was the presearch interview conducted in the usual way followed by the purple presearch questionnaire (see Appendix A7). The search was then executed using either Boolean or Weighted retrieval depending on what had been allocated. After the search while still at the terminal the user and intermediary would complete the blue and green questionnaires respectively. An appointment was made for the next visit to evaluate and collect the offline prints.

Logs of the searches were kept automatically by Cirt on the LSI 11/23 so neither the user nor the intermediary were concerned with their collection.

3.6. Discussion

As with any project there are certain things that would have been done differently with the benefit of hindsight. It was regrettable that the data collection from the medical schools was so slow in coming and so limited in number. Technical problems, coming to grips with the front-end and the limited time that the very busy intermediaries could spend on the project provided a disappointing number of results. Free searches done by the project information scientist, on the other hand, proved an excellent method for data collection. It greatly reduced wastage because most users willingly returned to collect their own copy of references as well as evaluating those for the project. It also provided technical advantages because there was no need for an additional telecommunications link to City. All transactions with Cirt were direct. In addition programming and technical support was on site and problems could be dealt with immediately.

It may be argued that the free searches were potentially less realistic than those undertaken at the medical schools. In the sense that at least some of the searches would not have happened without the free search offer, this is the case. However, the users concerned had to be sufficiently motivated to come to City on two separate occasions, the enquiry had to be a legitimate subject query (as opposed to a quick author or journal search) and the requests certainly represented genuine

current interests or problems, which they would have tackled (if not with an on-line search) then some other way.

With regard to the questionnaire, some minor changes in wording and construction of certain questions would have been desirable. In particular when enquiring as to the intermediaries contribution to the search (Blue questionnaire number four) it would have been useful to split this up into two questions. Firstly could the user have done the search on their own without an intermediary? Secondly how helpful was the intermediary. Also on the Blue questionnaire in questions seven and eight, the word references would have been better as titles. In the same question not every one knew what was meant by modify. It might have been more helpful to say "If yes did you add new terms or limit with other terms on the basis of the titles you saw online?" Lastly it would have been better to have asked the intermediary as well as the user what they thought about the nature of the subject enquiry (Purple questionnaire number three), i.e. Precise or Accurate, General, Vague or Waffley. The reason for this is the intermediary is more likely to understand how the terms relate to the subject domain. Then this question perhaps could have provided a more sound appraisal upon which to base a priori judgements relating to the search. Nevertheless these changes are minor and perhaps would be of little significance, given the results discussed below.