A Multidimensional Approach to the Measurement of Human–Computer Performance

A. P. JAGODZINSKI* AND D. D. CLARKE†
*Department of Computing, Plymouth Polytechnic, Drake Circus, Plymouth PL4 8AA
†Department of Experiment Psychology, University of Oxford, South Parks Road, Oxford

An approach to the problem of performance evaluation of human–computer systems, which (a) was designed for a comparative evaluation of different versions of the same system; (b) regards user job satisfaction and system comprehension as being as important as system efficiency; and (c) presents a coherent strategy for human–computer system evaluation; is described.

Received September 1986, revised May 1987

1. INTRODUCTION

The performance evaluation of human–computer systems has been recognised as a complex problem: 'Techniques are required that take account of the multidimensional nature of human–computer dialogue in the context of both task and user variables'.[7]

This paper presents one approach to the problem, derived with a good deal of retrospective rationalisation, from a 3-year research project based on a live on-line system implementation in a university library. The approach was designed for the comparative evaluation of different versions of the same system. It is orientated towards situations in which the users' job satisfaction and comprehension of the system are regarded as being at least as important as the system's efficiency in processing transactions. The criteria and techniques described here are intended to be suitable for use by the computer systems analyst who, in a commercial system implementation, is responsible for testing the overall performance of the human–computer system's performance before it is released to the users. The aim is to present a coherent strategy for human–computer system evaluation which can be applied by the systems analyst as a normal part of the implementation process, at a reasonable cost.

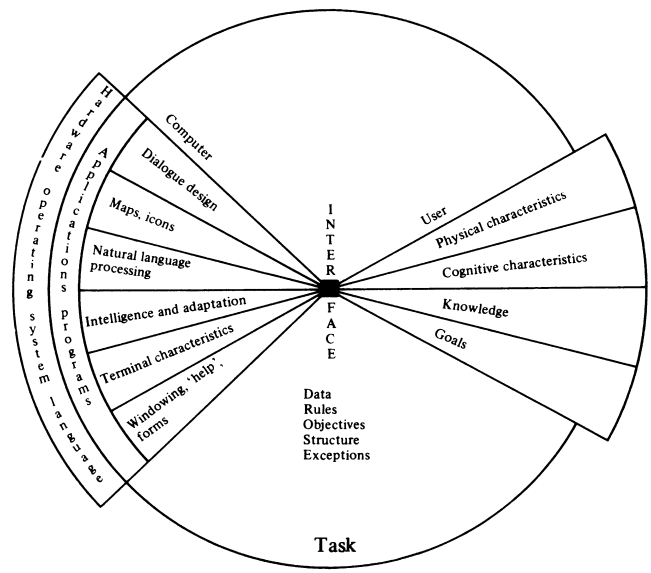

Figure 1. The central elements of human–computer interaction.

2. DIMENSIONS OF INTERACTION

In recognising the multi-dimensional nature of human–computer dialogue it is necessary to identify and define the important dimensions so that they can provide criteria for evaluation.

The terminal dialogue is the central and most intensive medium of human–computer interaction, but should be seen in the context of the task which the system performs.

The quality of terminal dialogue depends on the match between the technical elements of the computer system and the cognitive characteristics of the user. Some of the more important facets of system and user are summarised in Fig. 1.

Fig. 2 adds the job effects, those elements of organisation and the larger world in which the task and system are set.

The users' attitudes, beliefs and personal objectives are in complex interaction with the host organisation's objectives and norms, with its task structures, and with

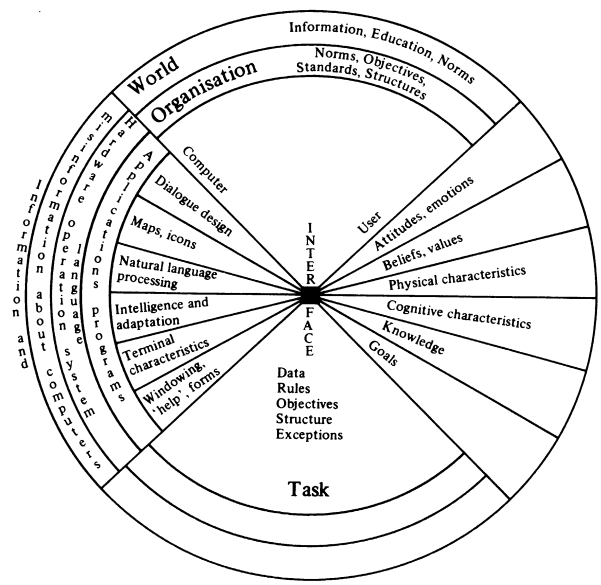

Figure 2. A broader perspective of human–computer interaction.

the general information and misinformation about computers that exists in the world in general. These affect the broader acceptability of the computer system to the user and cover such problems as deskilling, loss of status, job security and so on.[10,13]

A complicating factor is that the three aspects of interaction, that is job effects, task effects and terminal dialogue, can often be inextricably linked in their effect on users' attitudes. There are explanations for this, for example the principle of cognitive consistency which describes how individuals who perceive some disadvantage in a situation then find it hard to recognise or acknowledge its possible advantages.[11]

Although the terminal dialogue provides the most direct and concentrated medium of interface, the more diffuse, amorphous dimensions of interaction can be at least as significant for the acceptance of a computer system by its users. Thus task and job should probably be seen as providing the essential context of the terminal dialogue, with no one element being meaningful in isolation from the others. For this reason, methods for the evaluation of interfaces were also designed to comprehend the attitudes and perceptions of users with regard to the computer system's effects on their tasks and jobs.

3. EVALUATION CRITERIA

From these dimensions of the interaction it was possible to identify the performance characteristics by which its quality can be measured. Overall, four categories of performance were distinguished and are summarised below.

3.1 Technical performance

This includes factors such as response time, screen refresh rate, ambient lighting and anthropometric details. There is already a considerable amount of guidance for design and evaluation of these factors, which were among the first to be identified by ergonomists. It is undoubtedly important to get this part of the design right, but the techniques can be found elsewhere and were not included within this research.[4]

3.2 Task efficiency

In traditional approaches to systems analysis and design, for example that of Thierauf,[18] the rate of accuracy of operators processing a representative cross-section of transaction types is generally considered to be the best test of the system's effectiveness. Such measures are, in addition, clearly indicative of the ease of use of the system and of its value as an aid to productivity. For this reason they were included, but were also interpreted in the light of the more holistic analyses which follow, rather than being taken in isolation.

3.3 Quality of users conceptual model (UCM)

This criterion is taken to be an indicator of how well the users' cognitive characteristics, knowledge and goals are recognised in the system design. It is based on T. P. Moran's explanation of UCM: 'the whole conceptual organisation of the computer system from the user's point of view – the user's conceptual model of the system – is an integral part of the user interface'.[12]

This was inferred partly from performance measures as in (3.2) above, partly from tests of the users' perceptions of how easy the system was to use and partly from tests of users' comprehension. The measurement of users' comprehension was based on tests of the accuracy of their conceptual models of the system, and on observation of their navigation from one task to the next.

3.4 Job satisfaction

Job satisfaction is hard to identify although its opposite, alienation, can be all too obvious.

> Alienation exists when workers are unable to control their immediate work process, to develop a sense of purpose and function which connects their job to the overall organisation of production, to belong to integrated industrial communities, and when they fail to become involved in the activity of work as a mode of self-expression.[2]

Job satisfaction, then, can be taken to exist when the user has control of his work process, a sense of purpose in relation to organisation goals, a variety of tasks, when his skills are used appropriately and when there is a sense of social community.

These elements of performance are probably best interpreted in terms of users' subjective perceptions of effects, rather than, for example, by some external yardstick of task size or job status which may not equate with the users' criteria. Users' perceptions of task and job effects were assessed by the use of a 34-item attitude questionnaire.

The content and form of this questionnaire were drawn from an earlier 60-question investigation into the attitudes of all the staff towards the prospect of computerisation in the organisation.[9] This provided a clear indication of the aspects of job satisfaction which were thought to be most liable to change as a result of computerisation. These included task comprehension, task size, both vertical and horizontal, deskilling, job status, job security and job prospects. The preliminary investigation also included pilot testing of the questions to ensure that their wording was clear to respondents.

Sophisticated measures of users' perceptions of on-line systems' quality have been used in other contexts. For example, Dzida, Herda and Itzfeldt[5] derived seven dimensions of quality from an initial set of 100 system requirements. However, their evaluation was based on the views of specialist computing staff and would have been too technical for naïve users. Bailey and Pearson developed a semantic differential scaling method for measuring users' perceptions of the quality of computerised information systems.[1] However, it too was orientated towards experienced users of long-established systems, rather than towards the comparative evaluation of prototype systems during development.

Both of these approaches provide excellent models of thorough, multi-faceted subjective assessment for different purposes. However, some of the techniques they use may not fall easily within the practical limits of a commercial computer system implementation.

In many other cases references to evaluation in the literature of HCI have been found to be derived from

objective measures in experimental situations. They have used subjects employed only for duration of the tests, so that subjective measurement on dimensions such as task and job satisfaction would not have been realistic.[6,17]

4. THE TEST SYSTEM

The criteria described above may be applied in a variety of ways. Their use is illustrated here by the examples which follow from a journals control system in a university library. Four prototype interfaces were created, performing functionally identical tasks, but with distinctly different dialogue designs. The basic functions of the system were as follows.

(1) Registration: recording the arrival of new issues of journals.

(2) Holdings enquiries: finding out whether or not a particular issue is recorded on the system.

(3) Changes to issue details: modifying details of issues already on file.

Routes between these functions are shown in Fig. 3.

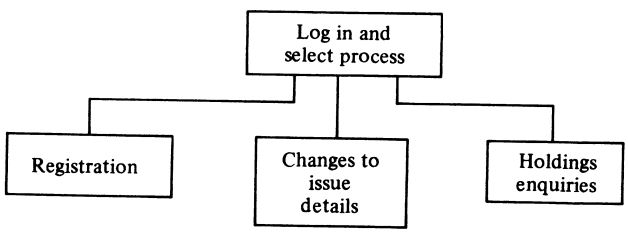

Figure 3. The functions of the journal control system.

The four versions of the system had the following interface features.

Version 1. Standard, menu-driven dialogue; continuous scrolling (Fig. 4*a*).

Version 2. 'Chunked processes' – also menu-driven, but with dialogue divided up into functional groupings separated by page breaks (Fig. 4*b*).

Version 3. 'Ancillary maps' – in addition to the standard interface, a second screen showing pictorial maps of the various stages of processing, for example Fig. 4*c*.

Version 4. 'Chunked processes' and 'Ancillary maps' – pictorial maps as in Version 3 with functional grouping of processes, as in Version 2.

5. TEST PROCEDURES

5.1 Test design

The purpose of the evaluation was to measure the effect of enhancing the systems with 'chunked processes', 'ancillary maps' and a combination of both features. In addition a control test (version 1) with neither enhancement was needed for comparison. Thus the experiment was set up as a 2×2 matrix of tests as in Fig. 5.

5.2 Arrangement of subjects

Volunteers were enlisted from the population of potential users of the system. They were full-time library staff, and none had more than one hour's previous experience of using any computer system. However, all had experience of routine clerical library tasks such as the use of a journals register.

With two hours available per subject and a test consisting of nine separate stages, it was necessary to expose each subject to only one version of the system.

This constraint prevented the study of transfer effects between different versions of the system. However, it also avoided the possibility of subjects becoming confused by exposure to the different versions, with the ensuing difficulties in measuring attitudes.

Volunteers' job grades and approximate ages were known before they were assigned to test groups. Eight subjects were assigned to each of the four test groups. Extra subjects were also tested for groups A, B and D; these permitted a small amount of adjustment of group membership, so that the groups' mean ages could be almost equalised (age proved to be a significant covariate with some aspects of subjects' performance).

Job grades were assigned equally to groups and as closely as possible to the proportions occurring in the whole population of staff of the library as follows. Job


```
                Please input EITHER the search-key for the journal's title
                        OR the journal's ISSN:

AAPGBU

        1. DEFINITIVE TITLE: 'AAPG Bulletin'

        SPONSORING BODY: 'American Association of Petroleum Geologists'

        The search-key you have input refers to the above title.
        You may EITHER input the number of the title you select,
                        OR re-input the search-key or ISSN:
1
        SELECTED TITLE: 'AAPG Bulletin'

        HOLDING (1)     LIBRARY  :  RSL              SOURCE : Copyright
                        STATUS   :  Ordered            COPY : 1
                        SHELFMARK :  PER 1253d 392     RANGE : 1980-
```

Figure 4*(a)*. Standard version of holdings enquiries, showing part of continuously scrolling screen dialogue. Scrolling takes place from the bottom upwards, continuously, with no break between sub-functions.

```
***** IDENTIFY TITLE *****

Please input EITHER the search-key for the journal's title
OR the journal's ISSN:

AAPGBU

1. DEFINITIVE TITLE: 'AAPG Bulletin'

SPONSORING BODY: 'American Association of Petroleum Geologists'

The search-key you have input refers to the above title.
You may EITHER input the number of the title you select,
OR re-input the search-key or ISSN:
1
SELECTED TITLE:  'AAPG Bulletin'
```

```
***** IDENTIFY HOLDING *****

SELECTED TITLE: 'AAPG Bulletin'

HOLDING (1)    LIBRARY  : RSL          SOURCE : Copyright
               STATUS   : Ordered        COPY : 1
               SHELFMARK: PER 1253d 392  RANGE : 1980-

HOLDING (2)    LIBRARY  : RSL          SOURCE : Purchase
               STATUS   : Available      COPY : 2
               SHELFMARK: PER 1253d 960  RANGE : 1982-

Please select one of the preceding holdings by inputting its
number:
2
SELECTED HOLDING: COPY 2
```

Figure 4*(b)*. 'Chunked processes' version of the holdings enquiries function, showing two successive screens, each one containing the chunk of dialogue for one sub-function. Each chunk would be preceded by a 'clear screen' action, and scrolling of text would proceed from the top on a blank screen.

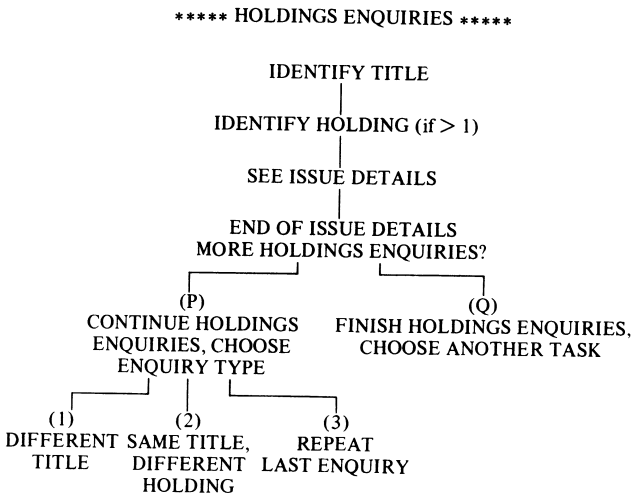

Figure 4*(c)*. Ancillary map showing the holdings enquiries function. Full details of this system and its alternative interfaces are given in Ref. 9.

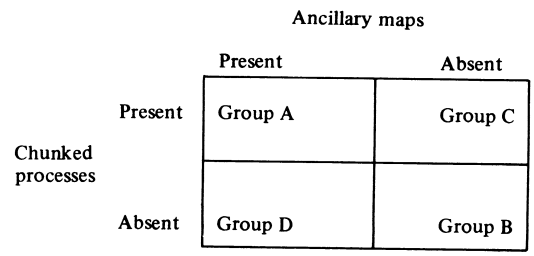

Figure 5. The set-up of the test groups.

grades per group: secretarial staff, 1; library assistants, 3; principal/senior library assistants, 3; academic related staff, 1.

Thus as far as possible the composition of test groups was equalised by stratification. The only aspect remaining which could not be assumed to be equally distributed was the subjects' previous experience of serials control work. This was taken into account by step 4 of the tests (see Section 5.4). In practice, there was no significant correlation between this measure and any other aspect of performance.

5.3 Physical layout of tests

The stages of the test involving the use of the computer were conducted using the room layout shown in Fig. 6. The observer was able to see the actions of the subjects and read the screen dialogues, but could not easily be seen by the subject. A separate version of the updated file was used for each subject so that the effect of their

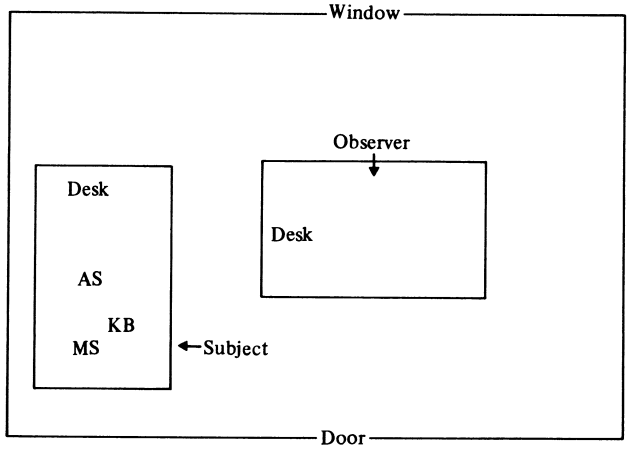

Figure 6. The layout of test equipment, subject and observer. AS = ancillary screen, MS = main screen, KB = keyboard.

updates could be permanently recorded and examined later to check their scores.

The tests were scheduled in conjunction with the co-users of the CTL 8046 computer to run at prearranged periods of low demand for service, so that response times would be consistent between tests.

All subjects received identical instructions and training, and performed functionally identical tasks. Each subject's test lasted about two hours including training and comprised the following stages:

(1) An introduction to the exercise.

(2) An introduction to the general principles of serials control.

(3) A training set of guided exercises in manual serials control.

(4) A timed set of exercises in manual serials control. This was included to measure previous experience in serials registration in case this was a covariate with performance on the computerised registration system (10 minutes).

(5) A training set of guided exercises in serials control with the computer system.

(6) A timed set of exercises in computerised serials control (30 minutes). This was to provide a measure of task efficiency (Section 3.2) and an indication of the quality of the UCM (Section 3.3)

Examples of tasks in manual and computerised serials control are given in Fig. 7.

(7) A test of comprehension of the system. (Based on the principles laid down by Bloom.)[3] See Fig. 8 for examples. This was to measure the quality of the UCM.

(8) A questionnaire to elicit the users' attitudes following their use of the system. See Fig. 9 for examples. This was to investigate the users' perceptions of the effect of the system on task and job factors, described in Section 3.4.

(9) A questionnaire on subjects' personal details, included to ensure that ages and job grades of subjects were equally distributed between groups.

Stages 1–3 and 5 provided the subjects with the necessary training. Stages 4, 6 and 7 provided objective measures of manual and computer transaction processing performance and comprehension, respectively. Stage 8 provided a subjective measure of users' attitudes.

This test design enabled the subject to carry the required operations with little reference to or help from the experimenter in order to reduce the risk of bias being introduced.

An observer was on hand throughout each test to give help to the subject if required. An experimental design in which no help would be given to the subjects was considered. This would have made scoring somewhat easier, as it is difficult to equate one form of help with another. However, as the subjects were real members of the library staff and as the tests represented their introduction to the use of computers its was thought to be highly undesirable to risk alienating them by offering no help when they became stuck.

(18) Register volume 4 of *Medicine and Sport*. It is damaged. Search-key is MEDSZZ.

(19) Check the status of *Medical School Admission Requirements, 1969*. If it is not 'binding' or 'on loan' change it to 'lost'. Search-key is MEDSAR.

(20) Record the arrival of parts 1, 2 and 3 of *Medicines Act 1968 Information Leaflets*. The are *not* combined. Search-key is MEDAIL.

(21) Register volume 73 part 432 of *Medicina Española*. Search-key is MEDEZZ.

(22) One privileged reader has asked to borrow part 17 of *Medicine in Ireland* and another has asked to borrow part 18. Can both requests be satisfied simultaneously? Search-key is MEDIZZ.

(23) The 1965 *Members' Handbook* has been lost. Change its status accordingly and add the note 'photocopy requested from BL'. Search-key is MEMHZZ.

Figure 7. Some examples of tasks in serials control. Up to 70 such tasks were included in manual and computer system tests, that is stages 3, 4, 5 and 6.

(2) When registering an issue it may be the next-expected part in the sequence What are the other four possibilities for its sequence?

(a) .....
(b) .....
(c) .....
(d) .....

(3) When you have identified the sequence of an issue during registration what do you do next? Tick *one* of the following list.

(a) Record the details of the issue.
(b) Examine the issue to see if it is damaged.
(c) Register the next issue.

(4) If you are registering an issue but there is no indication of its sequence is it possible to use the system to record its details? Tick one:

YES
NO

Figure 8. Some examples of comprehension questions. All subjects were required to answer 10 of these in stage 7 of the test.

variable DEVL

11. How do you view the effect of a computer system at the Library on your opportunities to develope your skills and knowledge?

I'd look forward to better opportunities			I wouldn't expect any change to be significant			I'd worry that opportunities would be worse
7	6	5	4	3	2	1

variable SECU

15. What is your reaction to its effect on your job security at the Library?

I'd expect more job security			I would not expect any change			I'd expect less security
7	6	5	4	3	2	1

variable JUDG

27. If your work involved the use of a computer, do you think this would lead to more use of judgement and making decisions or more simple application of standard procedures?

More judgement and decisions with a computer			About the same as now			More simple application of standard procedures with a computer
7	6	5	4	3	2	1

variable ENJO

33. Would you expect to enjoy your job more or less if it involved the use of a computer?

I'd expect to enjoy it much more with a computer			About the same as now			I'd expect to enjoy it much less with a computer
7	6	5	4	3	2	1

Figure 9. Some examples from the questionnaire on users' attitudes. Each subject completed 34 such questions. The left/right orientation of favourable/unfavourable responses was randomised in the full questionnaire.

6. SCORING

Score sheets were kept for each subject. For example, the scoring for the computer system exercise is shown in Fig. 10.

The 'RESULT' column was coded 1 for a correctly completed transaction or 0 for an error or uncompleted transaction. The 'HELP' column was filled in with the number of separate requests for help in any one transaction. In practice this never exceeded 1.

The 'NUMBER OF DETOURS' column was coded with the number of sub-optimal branches chosen by the operator for each transaction. For example, at the end of question 19 the screen prompts the operator as follows.

'This enquiry is now completed.'
Please input EITHER 'P' if you wish to make another enquiry, OR 'Q' if you wish to finish.

If the operator responds with 'Q' he is given 1 point for a detour as the next transaction is another enquiry.

The 'NAVIGATION RESULT' column was not used in practise, but was intended for calculation of the proportion of detours to possible detours.

NAME: VERSION: TIME:
GROUP: DATE:

(1) QUESTION NUMBER	(2) RESULT	(3) HELP	(4) NET RESULT (2)-(3)	(5) NO. OF DETOURS	(6) NAVIGATION RESULT (3-(5)	
1						
2						
3						
4						

Figure 10. The score sheet for the computer system tests.

Comprehension, stage 7 of the test, was simply scored on the number of correct answers, with a possible maximum of 1 point per question, 1 being deducted for each missing or wrong response.

The attitude questionnaire was designed according to the principles suggested by Oppenheim (1966), with a bipolar scale divided into seven equal intervals, each of which is subsequently given a numeric score (see Fig. 9). The subject rates each item by assigning to it the position on the continuum which his perception of the issue merits. It is described fully by Jagodzinski.[9]

7. ANALYSIS OF SUBJECTS' PERFORMANCE SCORES

7.1 Correlations between variables

Prior to the tests being carried out it was suspected that there may be important covariance between scores on performance with the computer system and scores on the manual registration task, a measure of previous experience with the manual system. In practice there was none. Consequently, no analysis of covariance with scores on manual registration was carried out.

Correlations between other aspects of performance in processing transactions were investigated using two-tailed tests of significance and the PEARSON CORR option of the SPSS package.[14]

There were interesting significant negative correlations between some variables across all subjects. These are shown in Table 1.

There were significant negative correlations between several aspects of performance and the subjects' ages.

Table 1. Significant correlations between performance scores (all groups)

Variable 1	Variable 2	Pearson *r*	P
Age	Help given	0.4089	0.020
Age	Comprehension	−0.3755	0.034
Age	Net score	−0.4687	0.007
Comprehension	Net score	0.5963	0.001
Comprehension	Errors on 1st 10 questions	−0.4187	0.017
Comprehension	Reported of ease of understanding	−0.2137	0.240
Errors on first 10 questions	Navigation detours on first 10 questions	0.3661	0.039
Ease of use	Reported ease of understanding	0.6058	0.001
Ease of use	Reported ease of relating system to task	0.3630	0.041

Table 2. Group means for subjects' ages

Chunked processes	Ancillary maps: Present	Ancillary maps: Absent
Present	Group A 32.5	Group C 32.5
Absent	Group D 31.1	Group B 32.875

However, groups were stratified to ensure that ages and grades were balanced, so that the effect of age as a covariate may be ignored. The group means for age are shown in Table 2.

Comprehension scores from stage 7 of the tests seem to reflect closely the subjects' performance score from stage 6, as might be expected. However, subjects' reported ease of understanding from stage 8 had an insignificant but negative correlation with comprehension scores, suggesting that subjects were generally not able to judge how well they understood the system.

The significant positive correlation between navigation detours and errors on the first 10 questions in stage 6 of the tests suggests that these two aspects of performance are related, although not necessarily causally.

Predictably, subjects' reported ease of use of the system from stage 8 correlated positively with their reported ease of understanding and reported ease of relating the system to the task. These three correlations of variables from the attitude questionnaire are included with the performance variables because they reveal the disparity between actual and reported ease of use, even though reported ease of use is consistent with reported ease of relating the system to the task and understanding. No other correlations between attitude variables were considered, these being examined by means of discriminant function analysis and analyses of variance described in Section 8.

7.2 Analyses of variance

Two-way unrelated ANOVA calculations were used to see if there was any significant variance between the results of the four groups of subjects on the performance variables.

Initially it was thought that a simple comparison of total numbers of correctly processed transactions might be all that was needed. However, the large within-group

variances of such a gross score effectively obscured inter-group difference.

It was also realised that this result would have been distorted by the fact that subjects were given help by the observer when they could make no further progress on their own (these occasions were recorded too).

Assuming that subjects' abilities are distributed equally between the groups, for a given set of tasks the best indicator of the differences between the user interfaces was judged to be the number of occasions on which help had to be given by the observer.

Accordingly, the number of requests for help during the first 10 transactions in stage 6 of the tests was used as the score for the first ANOVA. The group totals are shown in Table 3.

Groups A and C clearly required far less help than groups B and D.

The significance of this finding is shown by the results of the ANOVA in Table 4.

This result shows that an interface with chunked processes enables the user to process transactions with significantly less help than otherwise. Presumably process chunking makes the operation easier to grasp (as was predicted by Rasmussen and Jagodzinski,[16,8]) so that less external help is necessary.

In practical terms this improvement would have important benefits for installations with naïve users. If, as would be the case in the library system, the users never got the chance to develop fluency with the system, the benefit would materialise as a significant reduction in the need to call in the system supervisor. If there was no help conveniently available the benefit might turn out to be a significantly reduced error rate.

The second aspect of performance in stage 6 which was chosen as having potential for distinguishing between groups was the number of navigational detours (i.e.: sub-optimal choices at branch-points) made in the first 10 transactions. Navigation performance can probably be taken as an indicator of the quality of the subjects' overall view of the system and the routes available to them.

The group totals are shown in Table 5.

Both ancillary maps and chunked processes appear to have a beneficial effect on the subjects' navigation. The exact nature and significance of this effect is shown by the ANOVA results in Table 6.

This result shows clearly the value of the combination of chunked processes and ancillary maps in assisting the users' navigation through the system. The two facilities were designed to be complementary, with matched

Table 3. Group totals of help given during first 10 transactions

Chunked processes	Ancillary maps: Present	Ancillary maps: Absent
Present	Group A 4	Group C 6
Absent	Group D 11	Group B 12

Table 4. ANOVA of help given on first 10 transactions

Sources of variance	Sums of squares	Degrees of freedom	Mean squares	*F* ratios	Significance
Variable A (ancillary maps)	0.0313	1	0.0313	0.0279	—
Variable B (chunked processes)	5.2813	1	5.2813	4.7154	< 5%
A × B interaction	0.28	1	0.28	0.25	—
Error	31.3763	28	1.12	—	—
Total	36.9689				

Table 5. Group totals of detours made on the first 10 transactions

Chunked processes	Ancillary maps: Present	Ancillary maps: Absent
Present	Group A 7	Group C 12
Absent	Group D 11	Group B 24

Table 6. ANOVA of navigational detours on the first 10 questions

Sources of variance	Sums of squares	Degrees of freedom	Mean squares	*F* ratios	Significance
Variable A (ancillary maps)	0.0625	1	0.0625	0.0016	—
Variable B (chunked processes)	0.5	1	0.5	0.0126	—
A × B interaction	19.5625	1	19.5625	13.7803	< 0.1%
Error	39.75	28	1.4196		
Total	59.875				

Table 7. Comparison of group means for tests which showed no significant benefit from the system enhancements

Description of score	Group means: Maps and chunks	Chunks only	Maps only	No maps, no chunks
Comprehension test scores	8.5	7.5	8.1	7.25
Errors on first 10 transactions	0.625	0.75	0.625	0.875
Total transactions attempted	16.5	13.9	12.4	14.25
Total correct transactions excluding those where help was given	16	12.6	10.9	12.5

headings indicating the relationship between current process and overall position in the function. It appeared that this aspect of the interface enhancements was successful.

In practical terms fewer navigational detours show that these subjects are finding the optimum routes through the system more quickly and may ultimately perform faster and with greater confidence.

However, it is interesting to note that this enhanced version of the system was less popular than some of the others (see Section 8, Analysis of attitudes).

ANOVAs were also carried out on other aspects of subjects' performance, and although these generally showed better results with the enhanced versions of the system, the advantages they were not significant at less than 10%. These results are summarised in Table 7.

8. ANALYSIS OF SUBJECTS' ATTITUDES

Stage 8 of the tests, described in Section 5.4, was an attitude questionnaire (see Fig. 9) designed to elicit subjects' perceptions of a range of issues affecting the quality of the dialogue, the tasks they were asked to perform and the larger context of their jobs.

Discriminant function analysis[14] was used in an attempt to identify functions which would effectively discriminate between the groups. Wilk's lambda was used as the criterion of discriminating power on which variables were to be selected for the analysis. With 32 subjects and 34 variables there was a danger that an uninterpretable solution, tending towards one variable per case, could have been reached. Consequently the maximum number of steps in the selection of variables was set to nine, so that only the nine variables with the most discriminating power would be selected.

The results from the discriminant function analysis were not particularly revealing. Briefly, the most strongly defined variable emerged as question 11, 'How do you view the effect of a computer system at the library on your opportunities to develop your skills and knowledge?' (variable DEVL).

Group A and group C, both of which had the 'chunked processes' feature, appeared to be most popular in this respect. However, the nine variables identified as having the most discriminating power were then examined in more detail using analysis of variance, as described in Section 7. Of these the variables which proved to be significant are those obtained from the questions shown in Fig. 9. Some caution must be used in evaluating the significance of the results of such a wholesale approach. At a significance level of 10% one would expect 1 in 10 tests to appear significant by chance. The results which follow include only those with results significant at the 5% level of probability or less. They do appear to be consistent with each other and with the results of the discriminant function analysis.

Note that all four questions are shown in full in Fig. 9, question numbers 11, 15, 33 and 27, respectively.

The results for variable DEVL are shown in full below, followed by a summary of the next most significant variable, JUDG.

Groups A and C clearly have more optimistic expectations for their opportunities following computerisation than groups D and B. The significance of the finding is shown in Table 9.

Table 8. Group means for variable DEVL (opportunities to develop skills and knowledge)

Chunked processes	Ancillary maps: Present	Ancillary maps: Absent
Present	Group A 6.125	Group C 5.75
Absent	Group D 5.25	Group B 4.375

Table 9. ANOVA of DEVL (opportunities to develop skills and knowledge)

Sources of variance	Sums of squares	Degrees of freedom	Mean squares	*F* ratios	Significance
Ancillary maps	3.125	1	3.125	2.5925	
Chunked processes	10.125	1	10.125	8.3997	< 1%
Interaction	0.5	1	0.5	0.4148	
Error	33.75	28	1.2054		
Total	47.5				

Table 10. Results of ANOVA for question 27 JUDG (use of judgement and decision making)

	Ancillary maps	Chunked processes	Interaction
F ratios	4.8	3.3	10.1
Significance	< 5%	< 10%	< 0.5%

The high degree of significance of this result, coupled with the fact that this variable was the most significant in the discriminant function analysis, shows it to be highly important. An explanation of the effect could be that subjects felt better able to cope with the system when it was provided with chunked processes, and therefore viewed their development under computerisation more optimistically than those who did not have the benefit of the enhancement.

Analysis of variance also showed significantly greater optimism for the chunked-process versions on subjects' perceptions of job security ($P < 0.1\%$) and job enjoyment ($P < 5\%$), with the use of a computer system.

Again the interpretation of these results was that subjects felt better able to cope with the computer system.

Table 10 summarises the results for question 27 on use of judgement and decision making.

The group means for this variable (Table 11) reveal two effects.

First, group C with chunked processes but no ancillary maps stands out as expecting more use of judgement and decision-making than any of the other three groups. (Use of judgement and decision making were identified in an earlier survey as desirable characteristics of a computer system.)[9]

Secondly, this expectation is absent in group A, which

had ancillary maps as well chunked processes. Thus it seems that the presence of a second screen generates pessimistic expectations which cancel out the advantages of chunked processes.

The low scores of group A obscure the high scores of group C, so that, overall, chunked processes do not appear to have a significant effect. The presence of ancillary maps appear to have the overall effect of reducing scores, although a look at Table 11 shows that this is only the case between groups C and A, and not between groups B and D.

Table 11. Group means for question 27 (use of judgement and decision making)

Chunked processes	Ancillary maps: Present	Ancillary maps: Absent
Present	Group A 2.875	Group C 4.75
Absent	Group D 3.375	Group B 3.00

This undesirable interaction effect between the two system enhancements was not expected. Section 7.2 shows that there is a significant improvement in users' performance in navigation when both enhancements are present, but perhaps this is gained at the cost of the user feeling spoonfed by the interface and not being able to exercise his own judgement. Alternatively, it may be a reaction against having two screens to look at, with a possible feeling of pressure from having to operate in the data domain and functional domain simultaneously. Rasmussen explains:[16]

> no mental task should be forced into a level of consciousness higher than the task itself justifies (due to some inappropriate coding of information or choice of strategy in the computer). If this principle is not followed, the operator may have to time-share the main task with the extra irrelevant task of data recording (page 85).

To summarise, of the two enhancements to the system's interface the chunking of processes was clearly the more successful. On the measures of transaction-processing performance it significantly reduced the amount of help required by the users, and on the measures of attitudes it was found to improve significantly several aspects of the users' expectations of the effects of computerisation.

The most plausible explanation of this outcome is that the effect predicted by Rasmussen (1980), that is, an improvement in the users' capacity to grasp the processes of the system, has occurred. This manifests itself directly in that the users require less help in operating the system, and indirectly in that their confidence as computer users, and thus their optimism about their future under computerisation, increase.

Ancillary maps appeared to improve navigation performance but at the same time reduce the optimism of users' expectations. This effect was not expected by the systems designers (although maybe it should have been in the light of Rasmussen's work).

9. SUMMARY AND CONCLUSIONS

This approach to evaluation was based on the assumption that it is important to assess the performance of a human/computer system not just on technical factors but also on job factors, task factors and its match with users' cognitive characteristics. This assumption arose from a preliminary investigation of users' perception of the likely effects of computerisation. In other situations it is likely that such dimensions of interaction may be different, or at least have a different order of priority. For example, Bailey and Pearson's study of user satisfaction with management information systems identifies the most important factors as accuracy, reliability and timeliness.[1]

The aim of the approach described here was to compare the effectiveness of four different prototype versions of the same system. This again narrows the focus of some of the techniques so that they would not, for example, be directly transportable to a post-implementation review of the acceptability of a single system.

For the reasons given above, this approach to evaluation cannot be regarded in any way as universal. However, it can be assessed in terms of its aim of distinguishing on several dimensions the relative effectiveness of the four prototype systems.

On the dimension of cognitive fit or quality of UCM, characterised by objective measures of navigation performance, requests for help and user comprehension, results clearly supported the process-chunking prototype, and to a lesser extent the use of ancillary maps.

On the dimension of task efficiency, results were not statistically significant but still supported the process-chunking prototype.

On the dimension of job satisfaction obtained by subjective measures, significant preference for the process-chunking version was revealed on four counts, namely users' perception of opportunities to develop skills and knowledge, job security, job enjoyment and use of judgement. Significant dislike of the ancillary maps prototype was also revealed on this dimension.

These results demonstrate that the combination of objective and subjective techniques used is capable of revealing quite fine distinctions between different versions of a system, including the unexpected aversion to ancillary maps. The results also support the belief that system performance on one dimension, such as quality of UCM, significantly affects performance on other dimensions, such as job satisfaction.

Thus the results of the evaluation were sufficiently clear to satisfy the overall objective of the exercise, that is, to guide the systems analyst in his interface design. The choice made was for the process-chunking version, with ancillary maps to be available only if specified during a 'help' request.

Most of the costs incurred by this example of evaluation arose from time spent in reviewing the techniques of experimental psychology and statistics required for the design of tests, questionnaires and analyses. The cost of implementing similar tests for a different system would therefore be considerably lower. It is estimated that 10 working days spread over 2 man-months for a systems analyst would be sufficient to set up the tests including preliminary questionnaires, run them

and analyse the results, with about 2 hours per subject in addition. The total labour costs of the whole project, including systems analysis, design and programming, were of the order of 35 man-months.

In this academic library, as in many other organisations, the staff, with their thousands of man-years of combined experience and expertise, represent the institution's most valuable asset. In such circumstances it must be important to pay attention to the acceptability of any computer system which they may be required to use. The approach adopted should recognise the priority of user acceptability at all stages of the computerisation project from systems analysis and design through to implementation. System evaluation using methods such as those described here can provide a valuable indication of acceptability both relatively between alternative designs and, to some extent, absolutely as a measure of users' perceptions of how it will affect their working lives.

Acknowledgements

The research described here was funded by the SERC and ICL Ltd through a CASE studentship. The system described was implemented and tested at the Radcliffe Science Library of Oxford University with the kind permission of Dr Dennis Shaw CBE, keeper of Scientific Books.

REFERENCES

1. J. E. Bailey and S. W. Pearson, Development of a tool for measuring and analysing user satisfaction. *Management Science* **29** (5), 530–545 (1983).
2. R. Blauner, *Alienation and Freedom: the Factory Worker and his Industry*. University of Chicago Press, Chicago (1964).
3. B. S. Bloom, *Taxonomy of Educational Objectives*. Longmans. New York (1956).
4. K. B. De Greene, Systems and psychology. In *Systems Psychology*, edited K. B. De Greene, pp. 3–50. McGraw-Hill, Maidenhead (1970).
5. W. Dzida, S. Herda and W. D. Itzfelt, User-perceived quality of interactive systems. *IEEE Transactions on Software Engineering*, *SE-4*, 270–276 (1978).
6. G. J. Hitch, G. Alistair, J. Sutcliffe, J. M. Bauers and L. M. Eccles, Empirical evaluation of map interfaces: a preliminary study. *Proceedings of the 2nd Conference of the BCS HCI Specialist Group*, edited M. D. Harrison and A. F. Monk. CUP, Cambridge (1986).
7. S. Howard and D. M. Murray, An outline of techniques for evaluating the human–computer interface. *Fourth Symposium on Empirical Foundations of Information and Software Sciences*. Atlanta, Georgia (1986).
8. A. P. Jagodzinski, A theoretical basis for the representation of an on-line computer system to naïve users. *International Journal of Man–Machine Studies*, **18**, 215–252 (1983*a*).
9. A. P. Jagodzinski, Some applications of cognitive science in the analysis, design and implementation of interactive computer systems. *D.Phil. Thesis*, Oxford (1983*b*).
10. A. P. Jagodzinski, The interaction between electronic storage systems and their users. *Proceedings of the 11th meeting of IATUL*, Gothenburg, 1985, pp. 133–138 (1985).
11. A. P. Jagodzinski and D. D. Clarke, A review of methods for measuring and describing users' attitudes as an essential constituent of systems analysis and design. *The Computer Journal* **29** (2), 97–102 (1986).
12. T. P. Moran, An applied psychology of the user. *ACM Computing Surveys* **13** (1) (1981).
13. E. Mumford and M. Weir, *Computer Systems in Work-design – The ETHICS Method*. Associated Business Press, London (1979).
14. N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Bent, *Statistical Package for the Social Sciences*, 2nd edition. McGraw-Hill, New York (1970).
15. A. N. Oppenheim, *Questionnaire Design and Attitude Measurement*. Heinemann, London (1966).
16. J. Rasmussen, The human as a systems component. In *Human Interaction with Computers*, edited T. R. G. Smith and H. T. Green, pp. 67–96. Academic Press, London (1980).
17. Y. Rogers, Evaluating the meaningfulness of icon sets to represent command operation. *Proceedings of the 2nd Conference of the BCS HCI Specialist Group*, edited M. D. Harrison and A. P. Monk. CUP, Cambridge (1986).
18. R. J. Thierauf, *Systems Analysis and Design*. Merrill, Columbus, Ohio (1986).