

EVALUATING FEATURE CHECKLISTS AS A MEASUREMENT INSTRUMENT IN HUMAN- COMPUTER INTERACTION

Edward Andrew Edgerton

Submitted for the degree of Doctor of Philosophy (Ph.D.) to the University of
Glasgow, Faculty of Social Science.

Research carried out in the Department of Psychology

Submitted September 1994

© Edward Andrew Edgerton 1994

ProQuest Number: 13818404

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13818404

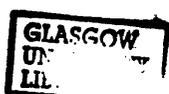
Published by ProQuest LLC (2018). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Ariv
10036
Copy 1



Dedicated to my parents: Edward and Catherine

Abstract

This thesis is concerned with the development and assessment of a measurement tool for use in the field of Human-Computer Interaction (HCI); the instrument is known as the Feature Checklist (FC). The FC consists of a list of features of the user interface such as menu commands, against which are a few columns each asking a particular question.

A series of seven studies were conducted in which the development of FCs progressed in a logical manner.

Study 1 demonstrated that FCs were a more accurate and valid instrument compared with simple open-response questionnaires for asking users about their recent usage of menu commands, and had an accuracy of 87%. Studies 2 and 3 attempted to increase the accuracy of FCs by improving their visual layout. Study 4 demonstrated that FCs could provide additional information (i.e. other than frequency of usage) and that this additional information was also accurate. Study 5 replicated the findings of study 4 in an HCI setting and also provided evidence to suggest that command names are a more suitable way of listing features on the F.C. than semantic descriptions of commands' functions. Study 6 demonstrated the way in which FCs could be applied to HCI evaluation and assessed the cost to the user of completing a FC. Finally study 7 employed FCs in a "real-life", industrial setting.

Throughout the thesis an attempt is made to relate the findings of each study to important research on human memory, in order to understand more fully the processes involved in FCs; the relevance of different theories of human memory are discussed.

The results suggested that FCs provide accurate and valuable information about such things as: usage levels of interface features; user knowledge of the existence and function of interface features; and user estimates of the usefulness of interface features. As such, it is proposed that FCs are a useful addition to the area of HCI evaluation.

Contents

Abstract	ii
Acknowledgements	vii
Chapter 1	1-1
Introduction	1-1
1.1 Overview of thesis	1-1
1.2 Outline of studies conducted	1-3
Chapter 2	2-1
The case for feature checklists	2-1
2.1 What are feature checklists?	2-1
2.2 Where are feature checklists used at present?	2-2
2.3 Is there a need for the feature checklist in HCI?	2-4
2.3.1 What type of information do feature checklists obtain?	2-5
2.3.2 What are the alternatives to feature checklists in HCI?	2-6
2.3.3 Review: reasons for using feature checklists	2-13
2.4 Where can feature checklists be applied in HCI?	2-15
2.4.1 Types of users	2-19
2.4.2 Where should feature checklists be used in the development cycle?	2-20
2.5 Summary	2-22
Chapter 3	3-1
Literature review	3-1
3.1 HCI literature	3-1
3.1.1 Command usage	3-2
3.1.2 Textual and graphical features of interfaces	3-5
3.1.3 HCI design methodology	3-8
3.2 Human memory literature	3-12

Chapter 4	4-1
A comparison of the feature checklist and the open response questionnaire in HCI evaluation	4-1
4.1 Introduction	4-1
4.1.1 The importance of assessing FC validity	4-2
4.2 Method	4-2
4.3 Results	4-4
4.4 Discussion	4-7
4.4.1 “Generation-Recognition” models of memory	4-7
4.4.2 FCs: order of presentation and accuracy	4-9
4.4.3 FCs and “state-dependent memory”	4-11
4.4.4 Electronic data logging	4-12
4.4.5 Unprompted recall (ORQ) and command usage frequency	4-15
4.4.6 Recognition (FC) and command usage frequency	4-17
4.4.7 “How often?” estimate and command usage frequency	4-21
4.5 Summary	4-24
Chapter 5	5-1
Visual realism in feature checklist design: implications for validity (“Brickles”).	5-1
5.1 Introduction	5-1
5.2 Method	5-3
5.3 Results	5-4
5.4 Discussion	5-6
5.4.1 “How often?” estimate and command usage frequency	5-7
5.4.2 FC accuracy: recognition vs. recall	5-8
5.5 Summary	5-9
Chapter 6	6-1
Visual realism in feature checklist design: implications for validity (“MacPaint”).	6-1
6.1 Introduction	6-1
6.2 Method	6-3
6.3 Results	6-4
6.4 Discussion	6-5
6.4.1 FCs and the “encoding-specificity hypothesis”	6-6
6.4.2 An evaluation of the “encoding-specificity hypothesis” and “generation-recognition” models of human memory	6-7
6.4.3 Current views on recall and recognition	6-8
6.4.4 The “transfer-appropriate processing” account of memory	6-10
6.5 Summary	6-11

Chapter 7	7-1
Using feature checklists for discovering user knowledge of the Glasgow underground railway system	7-1
7.1 Introduction	7-1
7.2 Method	7-3
7.3 Results	7-4
7.3.1 “Did you know this underground station existed?”	7-4
7.3.2 “Do you know what this underground station might be used for?”	7-6
7.3.3 “Do you think you would ever have a need to use this underground station?”	7-8
7.4 Discussion	7-10
7.4.1 “Existed?” column accuracy	7-10
7.4.2 “What for?” column accuracy	7-13
7.4.3 “Need?” column accuracy	7-15
7.4.4 Memory processes in different FC columns	7-17
7.4.5 FCs and memory: an explicit reconstructive process	7-18
7.5 Summary	7-20

Chapter 8 **8-1**

Employing feature checklists for measuring users’ knowledge and need for menu commands	8-1
8.1 Introduction	8-1
8.2 Method	8-3
8.3 Results	8-6
8.3.1 Standard FC accuracy	8-6
8.3.2 Standard FC accuracy of extra columns	8-7
8.3.3 Standard FC reliability	8-8
8.3.4 Standard FC accuracy vs. combined FC accuracy	8-9
8.3.5 Standard FC accuracy vs. semantic FC accuracy	8-10
8.4 Discussion	8-11
8.4.1 “What for?” column accuracy	8-12
8.4.2 “Need?” column accuracy	8-13
8.4.3 FC reliability	8-16
8.4.4 Using semantic descriptions of command function as memory prompts	8-16
8.5 Summary	8-19

Chapter 9	9-1
Using feature checklists to evaluate menu commands in a word-processing package	9-1
9.1 Introduction	9-1
9.2 Method	9-4
9.3 Results	9-5
9.3.1 System bug detection	9-5
9.3.2 Individual user performance	9-6
9.3.3 Feature checklist user cost	9-11
9.4 Discussion	9-13
9.4.1 Using FCs for system bug detection	9-13
9.4.2 Using FCs to improve user performance	9-14
9.4.3 Completing a feature checklist: the cost to the user	9-16
9.5 Summary	9-17
Chapter 10	10-1
Applying feature checklists in an industrial setting: monitoring usage of a “new” system.	10-1
10.1 Introduction	10-1
10.1.1 EasyReader: an electronic documentation system	10-1
10.2 Method	10-2
10.3 Results	10-4
10.4 Discussion	10-5
10.4.1 Cost to the investigator	10-9
10.4.2 Implications for future research	10-10
10.5 Summary	10-11
Chapter 11	11-1
Using feature checklists in HCI	11-1
11.1 Feature checklists: a critical review	11-1
11.1.1 Accuracy of feature usage data	11-1
11.1.2 Accuracy of feature knowledge data	11-3
11.1.3 The cost of employing FCs	11-3
11.1.4 Transferring laboratory validity to field validity	11-4
11.2 Feature checklists and human memory	11-4
11.3 Guidelines for feature checklist design and implementation	11-7
11.4 Possible future research	11-8
References	12-1
Appendices	A-1

Acknowledgements

This thesis reports research conducted from 1990 to 1993, whilst I was employed as a research assistant at the University of Glasgow on SERC/DTI grant GR/F/39171/IED4/1/1109 "Measurement of user interface performance".

Thanks are due to my supervisor Dr. Steve Draper for comments on my work, for ideas raised during discussions and for reading successive drafts of my thesis. I am indebted to Mr. Paddy O'Donnell for his assistance on statistical analysis and for general guidance on the layout of this thesis, and to Dr. Frances Duffy for reading and commenting on the sections of my thesis relating to human memory. I would also like to thank my colleagues in the Department of Psychology for advice and assistance given, in particular Pat Jordan and Judy Ramsey.

Finally, I would like to thank my wife Carole for her support, patience and understanding.

Chapter 1

Introduction

This thesis is concerned with the development of a measurement instrument for use in Human-Computer Interaction (HCI) evaluation and design. The instrument is called a “feature checklist” (FC) and relies on users’ memories to obtain information about their interaction with the features in a system e.g. menu commands. FCs obtain different types of information including:

- Whether users know that a feature exists.
- What features are used and how often.
- Whether users know what a feature is for.
- Whether users have a need to use a feature.

The thesis discusses the need for FCs in HCI and conducts a number of studies aimed at examining their accuracy. Practical applications of FCs are presented along with guidelines on how to use them. Finally, recommendations on future research on FCs are outlined.

1.1 Overview of thesis

In order to demonstrate the need for FCs and their usefulness, it is necessary to discuss alternative measurement instruments that currently exist. Chapter 2 discusses the

advantages and disadvantages of FCs in relation to these alternatives. This chapter also includes a discussion of where and when FCs should be used in the design process, and with what types of users.

A literature review of research relevant to FCs is presented in chapter 3. Since FCs are a new measurement instrument, this chapter reviews the most nearly relevant HCI literature, and includes sections on: command usage, textual and graphical features of interfaces, and HCI design methodology. Chapter 3 also includes a review of human memory literature since FCs are dependent on users' memories.

Given that research has often shown human memory to be unreliable, it is essential to assess the accuracy (validity) of the information that FCs obtain. Chapter 4 (study 1) demonstrates the accuracy of FCs and discusses this in relation to research on recall and recognition.

In chapters 5 and 6 (studies 2 and 3 respectively) attempts are made to increase the accuracy of FCs by making them more "visually realistic", i.e. closer in appearance to the actual interface under investigation. The second of these studies demonstrates that visual realism is an important consideration and discusses this with respect to the "encoding-specificity hypothesis" (Tulving and Thompson 1973).

Chapter 7 (study 4) attempts to assess the validity of additional columns on the FC but in a non-HCI setting. This chapter demonstrates that FCs can obtain valuable data on users' knowledge of the existence and function of features, and to a lesser extent their need for these features.

In chapter 8 (study 5) an attempt is made to replicate the findings of chapter 7 in an HCI setting. This chapter also investigates an alternative design of FCs where features are listed by a description of their function rather than their appearance (i.e. command name or icon). The results do not support the use of these descriptions.

An example of how FCs could be used for system bug detection in "real-life" evaluation is given in chapter 9 (study 6); this chapter also looks at how FCs can be used to improve user performance. Finally, this chapter examines quantitatively one of the main advantages of FCs that has been proposed namely, the cost to the user of completing a FC (in both time and effort).

The application of FCs in a “real-life” industrial setting is presented in chapter 10 (study 7). This chapter describes how FCs were used in a monitoring programme aimed at assessing the use of a newly installed electronic documentation system. The issue of the cost to the user is again addressed here along with the importance of defining the user type.

Finally, chapter 11 summarises what the thesis has and has not achieved. A discussion of important issues relating to FCs is given along with guidelines for using FCs and recommendations for future research.

1.2 Outline of studies conducted

In total seven studies were conducted; these followed on logically from each other in the development of FCs. Table 1.1 lists these studies along with a brief description of their primary aim and the system in which they were conducted:

Table 1.1: Description of studies conducted

Study Number	Description of Primary Aim
Study 1 (Chapter 4)	To compare FC accuracy with Open-Response Questionnaire accuracy, for measuring command usage (“Microsoft WORD 5.0”)
Study 2 (Chapter 5)	To try and improve FC accuracy for measuring command usage, through increased visual realism, (“Brickles 7.0” - a computer game for the “Apple Macintosh”)
Study 3 (Chapter 6)	A further attempt at improving FC accuracy through increased visual realism, (“MacPaint”)
Study 4 (Chapter 7)	To assess the accuracy of additional columns on the FC in a non-HCI setting (The Glasgow Underground Railway System)
Study 5 (Chapter 8)	A further attempt at assessing the accuracy of these additional columns in an HCI setting, (“Microsoft WORD 5.0”)
Study 6 (Chapter 9)	To demonstrate how FCs can be used as a bug detection instrument and to explore the cost to the user (“Microsoft WORD 5.0”)
Study 7 (Chapter 10)	To examine the cost of employing FCs in a “real-life” industrial setting (“EasyReader”)

Chapter 2

The case for feature checklists

2.1 What are feature checklists?

In its most basic form, the term “checklist” is simply a list of items (features) that may be checked (ticked) off by the respondent. However, there are many different meanings and uses for checklists; examples include:

- Pre-flight Checklist - To support behaviour by being an external memory aid e.g. aircraft pilots use one for pre-flight checks, to supplement recall of a large set or list (sequence) of items. The items are familiar and are instantly and reliably recognised; the checklist is to ensure none are forgotten. It is used to support an individual’s actions and uses recognition to overcome the unreliability of recall.
- Usability Checklist - To elicit information not from the respondent’s memory but via their behaviour, i.e. a checklist of information-getting behaviour, the results of which are recorded. For example, Ravden and Johnson’s checklist for usability evaluation in HCI is intended for use by those not familiar it; each item is not recognised but contains a full description of what to look for (Ravden and Johnson 1989).
- Feature Checklist - Unlike the previous 2 kinds of checklists, the feature checklist gathers information from the respondent’s memory. They consist of a list of features/items of the system/behaviour under investigation; the intention is that the respondent will immediately recognise and remember these features which will then act as a cue for asking questions about them. The usual layout of a feature checklist

is a list of features (e.g. menu commands) against which are a few columns, each asking a particular question about that feature (e.g. “Have you used this feature?”, “How often do you use this feature?”, etc.). In a sense therefore, feature checklists are a specialised form of questionnaire.

This thesis is concerned with feature checklists (FCs) and their use in HCI. At this point however, it is useful to examine in more detail, areas in which checklists, and FCs in particular, have actually been used.

2.2 Where are feature checklists used at present?

Although the term “checklist” appears quite frequently in HCI literature (e.g. Ravden and Johnson 1989, Oppermann, et. al. 1989, Reiterer 1992) it should be noted that these checklists are very different from the “feature” checklist that we are proposing.

In the Ravden and Johnson example, users filled in detailed “usability” checklists about the acceptability of various aspects of the interface, thus highlighting particular types of problems (Ravden and Johnson 1989). This type of checklist was more concerned with users’ opinions on various aspects of the system that they had just used; a section of this checklist is shown in figure 2.1.

Figure. 2.1: Example of section from Ravden and Johnson's usability checklist

Interface aspect	Unclear Clear				
“How legible was the text on the screen?”	1	2	3	4	5

Users had to work through lists of questions similar to this (concerned with different aspects of the interface) and answer by circling the appropriate answer.

The checklist described by Oppermann et. al. and Reiterer, known as “EVADIS” (evaluation of dialog systems), required usability specialists to evaluate the usability of the system for pre-defined tasks by assessing whether it met detailed requirements given in checklists (Oppermann et. al. 1989, Reiterer 1992). An example of the EVADIS checklist is shown below:

Figure. 2.2: Section of two-dimensional framework of the test items in EVADIS

ergonomic criteria technical system components	avail- ability	suitability for the task	clearness	self- descriptiveness	conformity with user expectation
1. input/output interface 1.1 info presentation 1.2 input media 1.3 input 1.4 output media 1.5 speech			1		
2. dialogue interface 2.1 dialogue techniques 2.2 dialogue control 2.3 messages 2.4 error handling 2.5 help-system, manual, tutorial					2
3. functional interface 3.1 functionality of the software 3.2 functionality of the user interface 3.3 response time			4		
4. organizational interface 4.1 tech. organizational interface					

This checklist consists of about 150 items which check the various properties of the user interface. All the items are embedded in a two-dimensional framework, with the dimensions being “technical system components” and “ergonomic criteria”. The task of the evaluator is to compare the analysed ergonomic quality with the attainable ergonomic quality and gave a rating (1 = full satisfaction, 5 = non satisfaction).

The “feature” checklist that we have defined is a very rarely used measurement instrument; this is especially true for HCI research, where they have never been applied or studied in any empirical manner so far as we are aware. However, one study that did employ a FC (albeit outside HCI) has been conducted (Belson and Duncan 1962).

In this study, the researchers compared the FC with an “open-response questionnaire” (ORQ) for measuring subjects’ memory of the publications (i.e. newspapers, magazines, etc.), that they had looked at the previous day and also the television programmes that they had seen the previous day. The FC listed a range of publications and television programmes and asked subjects to tick the ones that they could remember reading or watching. The ORQ on the other hand, simply asked subjects to write down the publications and television programmes that they could remember reading or watching.

The results showed a substantial difference in yield between the FC and the ORQ and led the researchers to conclude that, “the sharp differences in yields from the two methods makes it quite clear that at least one of them can be seriously in error when used to assess the previous day’s behaviour and ... in the circumstances validation of each of these methods becomes a pressing issue” (Belson and Duncan 1962). The FCs gave a higher yield. If this yield is of accurate answers (which they did not show), then this would be consistent with the general pattern of superiority of recognition over recall, but suggesting a practical application of this feature of human memory.

The question arising from this study that is of interest to this thesis is, does their result transfer to HCI contexts, i.e.:

- Is the higher yield of FCs also obtained in HCI applications?
- Is that yield trustworthy, i.e. are FCs reliable?

However, before attempting to answer these questions, possible applications of FCs to HCI, should they be shown to be valid, are considered.

2.3 Is there a need for the feature checklist in HCI?

FCs may be a useful method for obtaining information. To examine how this information could be used in HCI, we should look at:

- (i) The types of information that FCs can obtain.
- (ii) The alternative HCI methods that could obtain this information.

2.3.1 *What type of information do feature checklists obtain?*

The type of information that FCs can obtain will depend on the questions (columns) that the researcher decides to use. We propose that FCs can obtain three types of information:

(i) “Usage Information” - information about which features in a system people use, and how frequently these features are used.

It can be seen that FCs are in one sense an alternative to other measurement methods that obtain information about users behaviour/interaction with a system. These alternatives will be discussed later, but include video recording, direct observation, electronic data logging, etc. All of these methods can obtain detailed, quantitative information about the actions that users perform; this may lead to a rapid identification of such things as user difficulties, and where usage is concentrated and hence where design improvements will do most good.

(ii) “Knowledge Information” - information about people’s knowledge of certain aspects of the features in a system, e.g. knowing whether features existed and what these features are for.

This information can be both revealing about the system itself and users’ conceptual model of it. This may be used for diagnosing shortcomings in the system design or simply to gain a better understanding of the cognitive processes that users employ. Many HCI studies have been conducted in order to obtain this type of information, which can be qualitative or quantitative (e.g. Mayes et. al. 1988, Lewis et. al 1990, Molich and Nielsen 1990, etc.).

In order to obtain this “knowledge information”, researchers have employed a variety of measurement methods including think-aloud protocols, cognitive walkthroughs, interviews, etc. FCs may therefore, be a useful alternative to this aspect of these methods, which again will be discussed later.

(iii) “Opinion/Attitude Information” - behavioural information about people’s attitudes to the features that a system contains, e.g. opinions on whether the features are useful or ever needed.

A major aim of HCI research and HCI evaluation in particular, involves measuring user’s attitudes towards different systems and trying to understand why users hold

these opinions. This information may be useful for guiding system design which may in turn lead to greater user acceptance.

Although this type of research often focuses on “global” features of a system e.g. “how easy was the system to learn?”, it may also be useful to obtain information about users attitudes to more “low-level” features such as menu commands, which may impact on users overall attitude to the system. FCs may be useful for measuring users attitudes/opinions to these “low-level” features; in this sense, FCs can be seen as an alternative to methods such as questionnaires, semi-structured interviews, etc.

Rather than put forward the case that FCs are simply an alternative to other HCI measurement methods, it is the intention to show that FCs have many advantages over other methods, and in some cases may be the only feasible method to employ. The following section outlines the advantages and disadvantages of a number of HCI measurement methods already mentioned, including FCs.

2.3.2 *What are the alternatives to feature checklists in HCI?*

In order to demonstrate the need for FCs it is useful to discuss the advantages and disadvantages of FCs and the possible alternative methods already mentioned. This can be done by describing all the methods against three important criteria:

(1) “Cost” - the amount of time involved in using that measurement method.

Table 2.1 (below) compares the “cost” of employing a number of different HCI measurement methods including FCs. “Cost” can be broken down into four components, namely:

- Preparation (i.e. the time taken by the investigator to prepare the method).
- Administration (i.e. the time taken by the investigator to administer the method).
- Analysis (i.e. the time taken by the investigator to analyse the data obtained).
- User (i.e. the time taken by the user to complete the method).

It can be seen that the first three components relate to the investigator’s time, whilst the last component relates to the user’s time.

Table 2.1: Cost (in time) of employing various HCI measurement instruments

Method	Preparation	Administration	Analysis	User
Data logging	low or high	low	high	low or high
Video recording	medium	medium	high	high
Direct observation	low	high	medium	low or high
Think-aloud protocols	medium	high	medium	high
Cognitive walkthroughs	high	high	medium	N/A
Questionnaires	med/high	low	medium	variable
Semi-structured interviews	med/high	high	medium	high
Feature checklists	low/medium	low	medium	low/medium

Electronic data logs require little preparation by the investigator provided the software is available, otherwise it is unlikely to be a feasible method. If it is available the investigator does not need to be present, except that is, to start/stop the log recording; there may be some scope for automating this, e.g. the electronic data log may start/stop recording when the user initiates some specified action such as opening/closing an application. Perhaps the greatest amount of time involved in employing this method is that spent by the researcher analysing the resulting data. Many electronic data logs record at a low level (e.g. mouse positions) and so create huge log files which require a substantial amount of time in tasks such as re-coding, translation, etc. (Lindgaard and Millar 1989). Others, such as “UNIX acct”, record processes which have a variable and indirect relationship to user commands. An example of an electronic data log recording is shown below; this was obtained using the macro-recorder application “Tempo II plus” in study 1 (chapter 4). N.B. in this study, subjects did not use a mouse but instead selected commands using the keyboard; even with taking this into account, it should be seen that in the example shown only one menu command was identified.

Figure. 2.3: Example of electronic data log print out (“Tempo II plus”) from study 1

```
type enter
type clear
type enter
type "2"
type clear
type enter
type "3"
type clear
type enter
type "5"
type clear
type enter
type "4"
type clear
type enter
type "6"
type down-arrow
type down-arrow
type down-arrow
type down-arrow
type enter
```

Ideally, electronic data logging could be used in “real-life” situations where users would be performing their normal, everyday tasks with computers; this would obviously not require any more user time. In reality however, data logs are rarely used in “real-life” situations for various reasons (discussed later) and are instead used frequently as part of experimental investigations conducted in usability laboratories; in these situations, the cost in time to the user may be significant, (i.e. attending the session).

Video recording is a commonly used alternative to electronic data logs (e.g. Jordan and O’Donnell 1992, Neal and Simons 1983), and are very similar in terms of cost. The preparation time is spent by the investigator organising a suitable location, setting up the video camera, checking that the recording is clear and appropriate, etc. Often the

investigator needs to be present to ensure the recording is appropriate and therefore the time spent administering the method can be high. Again the major cost is in the time spent by the investigator analysing the data (recording); this can be laborious if the intention is to record command usage. Since it is often not practical to use video recording in “real-life” situations, the method is often employed in experimental settings; again, the cost in time to the user may then be significant (i.e. presence required for session).

Direct observation involves the investigator observing the tasks/actions that a user executes whilst using a system; the investigator may make notes on these actions (and possibly record timings). Little time is needed preparing the method except perhaps for designing lists/scales that the investigator may use and possibly arranging a suitable time/location. However, since the investigator needs to be present throughout, it is a costly method in terms of investigator time. The amount of time required by the investigator to analyse the data will vary depending on the purposes of the evaluation (level of detail needed). Since direct observation can take place either in “real-life” situations or in experimental, laboratory based studies, the amount of time required of the user can be negligible or significant respectively.

Think-aloud protocols require subjects to speak their thoughts out aloud while engaged in some task or action. As far as preparation is concerned, the investigator will need to spend some time planning the tasks that the user will engage in and preparing the recording method e.g. video, audio, note-taking, etc. Since the investigator is required to be present (e.g. for prompting the user), it is a very costly method in terms of investigator time. One alternative however, is to video record the users actions on the screen and replay these at a later date for the user to comment on; this technique is called “delayed protocols” (Draper and Barton 1993). Depending on the recording method used, the time spent analysing the data can be high (video recordings) or medium (note-taking). Regardless of whether “normal” or “delayed” think-aloud protocols are used, the cost in time to the user is significant.

Cognitive walkthroughs involve the developers of an interface examining the interface in the context of core tasks a typical user would need to accomplish; actions and feedback of the interface are compared to the goals and knowledge the user would be expected to have to identify discrepancies. Cognitive walkthroughs are a very expensive instrument for the investigator both in terms of administration and

preparation; in some cases they have been shown to be even less cost-effective than empirical testing (Karat et. al. 1992).

Questionnaires are an established measurement method in HCI and are often used as a substitute for behavioural observations because of their cheapness. Although a significant amount of investigator time is required in designing the questionnaires, they can be administered to large samples of users without the need for the investigator to be present, thus reducing administration time to a minimum. The amount of investigator time required to analyse the data and the time spent by the user to fill in the questionnaire will vary depending on its size and level of detail.

Semi-structured interviews have a fixed agenda and response categories, but allow the investigator flexibility in paraphrasing and follow-up probe questions. Although they require similar amounts of investigator time in preparing the method and analysing the data as questionnaires, they are often found to be more reliable. However, they are a costly method to employ both in the time spent by the investigator administering them and the time spent by the user completing them.

Feature Checklists can be relatively cheap to prepare depending on the system under investigation; they basically involve creating tables that list the features of the system (e.g. icons) and have columns (containing questions) alongside. Like questionnaires the experimenter need not be present and are therefore cheap in administration time. It is likely that the data analysis time and the user involvement time will vary depending on the number of features concerned and the questions asked. However, since the feature checklist only requires the user to tick or cross off answers, the user involvement time should be considerably less than that required for most questionnaires.

(2) "Situation/Control" - the situations in which the method can be used.

An important criterion that has already been touched upon concerns the situations in which the methods can be applied i.e. laboratory/experimental studies vs. "real-life" or "in-the-field" studies.

The laboratory based approach is commonly used in HCI research because the investigator can have careful control over the situation. As a result of this control the investigator can manipulate a number of factors associated with interface design and study their effect on various aspects of user performance, or simply observe and record

the interaction in great detail. This approach has been employed with much success by a variety of researchers using a variety of measurement methods (e.g. Baxter and Oatley 1991; Karat et. al. 1986; Sutcliffe and Springett 1992; etc.). Recently, interest has focused on the use and design of laboratories for conducting HCI “usability” testing (Nielsen 1994). However, there are a number of problems with laboratory based research that has led many researchers to look at alternatives; the major problems are:

- **Cost-effectiveness**: because laboratory research is expensive in terms of investigator time and equipment costs it is often impractical in many situations, or conversely must involve small numbers of users; this has led to the claim that “measurement is a time consuming and thus costly task ... Usability specialists are therefore often faced with a choice between doing more limited usability evaluation, or doing no usability evaluation at all.” (Molich 1994).
- **Artificiality**: as a consequence of the rigid control employed, there is a danger that the behaviour of the users may not be typical and therefore findings may not transfer to “real-life” situations. Some researchers have went as far as claiming that “It is not meaningful to talk simply about the usability of a product, as usability is a function of the context in which the product is used. The characteristics of the content (the user, tasks and environment) may be as important in determining usability as the characteristics of the product itself” (Bevan and MacLeod 1994).

Of the methods already discussed, a number have problems associated with use outside of laboratories. Electronic data logs are seldom used outside of laboratories because of the infrequent and unpredictable nature of “real-life” user interaction. In laboratory studies, users normally have pre-defined tasks to conduct and the investigator is present for switching the data log on and off; however, in “real-life” evaluations the electronic data log may have to be left on for a long time in order to collect the required information, this causes software memory problems as well as the problems of translating huge and often irrelevant information from the logs (already mentioned). Of course logging software could be written that avoided these problems and recorded more useful data, but such software is seldom available when an investigator needs it. Potentially logging is of great value; in practice so far it is poorly matched to investigator needs.

Video and direct observation methods also suffer from similar problems and in addition have the unwelcome problem of being obtrusive i.e. user performance levels may be

altered because they are aware that their performance is constantly being monitored; this is known as the “Hawthorne effect” (Mayo 1933; Roethlisberger and Dickson 1939). Think-aloud protocols and cognitive walkthroughs are also obtrusive methods which require a high investment of time from the user (as well as the investigator); as a result they are often employed with small groups of users to obtain detailed information.

Apart from the concern that laboratory findings may not apply in “real-life” cases (due to artificiality and obtrusiveness), there is the additional concern relating to the fact that these studies involve small groups of users, i.e. “is the behaviour of a small group of users with a large system, containing many features, similar to that of the population as a whole?” A study investigating the behaviour of users of the “UNIX” operating system highlights this problem (Draper 1985). One of the findings of this study is that users typically only use a small, but diverse, personal subset of the many features contained. The implication of this is that in order to evaluate the system properly large samples of users should be involved. As we have seen this is not feasible using methods such as video recordings, think-aloud protocols, etc, in laboratory settings. The obvious alternative would seem to be questionnaires and/or semi-structured interviews, however these themselves may be far from ideal in evaluating large systems with large numbers of users (e.g. questionnaires can be lengthy, whilst semi-structured interviews are expensive in investigator and user time). If it can be established that FCs are cheap to design and administer, and relatively cheap to complete, then they may be a very useful method.

(3) “Data” - the accuracy of the data obtained.

One of the proposed advantages of FCs (e.g. over questionnaires), is that users need only respond by placing a tick or a cross beside features to indicate their answer. As a consequence the feature checklist will produce quantitative data on user behaviour, knowledge, etc. (i.e. data that can be recorded in a numerical form). As we have already seen, FCs are an alternative to electronic data logs and observation methods for measuring what features are used and how frequently; however with respect to data on users’ knowledge of the functionality and existence of features, and their need for using them, FCs are an alternative to questionnaires and semi-structured interviews. The important question concerning the data is “do FCs have a comparable accuracy to these alternative methods?”

One possible problem with FCs is that they are a retrospective measurement and as such their accuracy may be affected by the known unreliability of human memory (e.g. Nickerson and Adams 1979; Mayes et. al. 1988). Despite the problems that have already been discussed with electronic data logs and video recordings, they will obviously not be affected by human memory and could be vastly superior (in accuracy) to FCs. However in one sense their accuracy may also be suspect. They record user actions rather than intentions, and therefore record events such as menu commands that were actually invoked even although the user may have been unaware of doing so (e.g. accidents such as the mouse slipping on pull down menus). Thus logging techniques cannot be wholly trusted for measuring which commands are known and intentionally selected by users. Studies 1, 2 and 3 (chapters 4-6) investigate the accuracy of the data obtained from FCs on feature usage.

Despite the fact that retrospective measurement methods are often suspect, it should be pointed out that in many “real-life” situations, evaluations are frequently not conducted until after the system has been up and running for a period of time; as a consequence retrospective methods are often the only option. Study 7 (chapter 10) is a good example of this situation.

Although we have no reason to believe that the data obtained from FCs on users’ knowledge of the functionality and existence of features, and their need for using them, will be inferior to that obtained from questionnaires, it is still important to assess the accuracy of this type of information; studies 4 and 5 (chapters 7 and 8) investigate this issue.

2.3.3 *Review: reasons for using feature checklists*

Having now discussed the types of information that FCs can obtain and the advantages and disadvantages of alternative measurement methods, we are now in a position to assess the potential contribution of FCs to HCI.

From the preceding discussion it was seen that the strength of empirically based, laboratory research is that it allows the investigator to statistically test explicit hypotheses relating to user interaction under controlled conditions. Despite this, empirical/laboratory methods are not always the most feasible or desirable method to employ due to their high costs and the constraints of many “real-life” situations; this view is best summed up by Nielsen and Molich: “the empirical or laboratory test method... consistently appears to be one of the more effective techniques for identifying

software usability problems. Unfortunately, in many practical situations empirical evaluations are not conducted because of lack of time, knowledge or resource” (Nielsen and Molich 1990). One study highlighted late involvement in the software development process as one of the most common complaints amongst interface designers (Grudin and Poltroch 1989). A “real-life” example that demonstrates this point is given in study 7 (chapter 10), where FCs are used to evaluate an electronic documentation system after it had already been installed and used within an organisation.

Of course FCs are not the only “real-life” measurement method available. However, it has been shown that there are many problems associated with the numerous, possible alternatives that exist. We propose that the three most likely alternatives to FCs for measuring users’ usage, knowledge and opinions about system features are electronic data logging, video recording, questionnaires and semi-structured interviews; these methods and their disadvantages (compared to FCs) are shown below in table 2.2.

Table 2.2: Most likely alternatives to FCs and their associated disadvantages.

Alternative method	Disadvantages compared to FCs
Electronic data logging	(i) often unavailable (need to be specially written or use other applications e.g. macro-recorders); (ii) often unsuitable (e.g. record overly detailed actions); (iii) translation/analysis problems; (iv) hardware problems (need to run for long time); (v) “real-life” evaluation is often retrospective; (vi) don’t record user intentions; (vii) only record usage information
Video recording	(i) obtrusive; (ii) expensive to analyse and code; (iii) often poor quality; (iv) only record usage information
Questionnaire	(i) more time consuming to prepare; (ii) can be time consuming and difficult for user to complete
Semi-structured interview	(i) very costly for investigator to administer; (ii) user participation is very time consuming

An important aspect of this table that has not been dealt with in great detail, is that since FCs obtain different kinds of information (see section 2.3.1), the alternative is to employ some combination of the methods listed in the table e.g. an electronic data log to record feature usage and a semi-structured interview to measure user opinions and

knowledge of features. If we ignore the problems already mentioned about electronic data logs, then it could be argued that this would be the most useful and cost-effective combination if the sample size for the interview was small. However, as we have already discussed, findings from other research has shown that users have diverse usage and interviewing a small sample will not obtain this information nearly as well (Draper 1985). Feature checklists can be used at little extra cost (to user and investigator) with large samples of users.

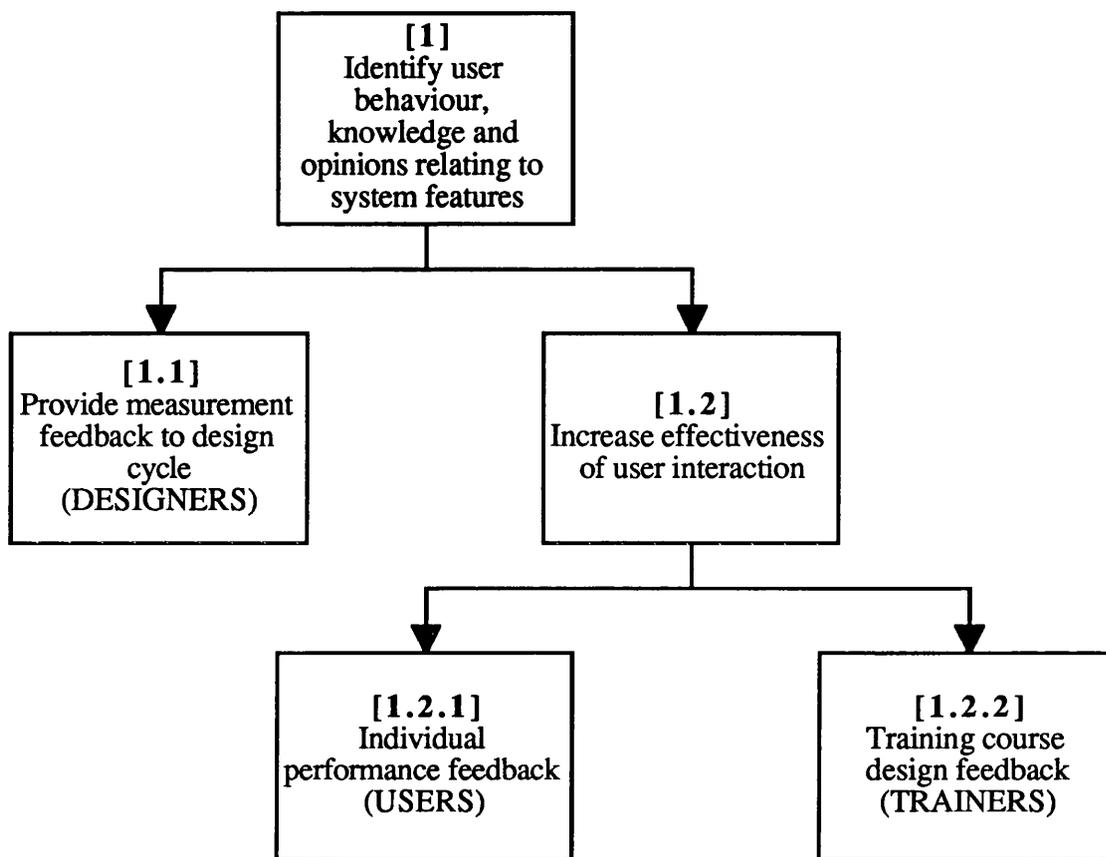
Due to the practical problems associated with laboratory evaluations, an important question arises, namely: “Is there an alternative method that is as effective as a lab test but less costly and time consuming?” (Smilowitz et. al. 1994). We propose that in some cases, FCs may be just such a method.

Having now addressed the need for FCs, we are now in a position to discuss where we feel FCs can be used in HCI.

2.4 Where can feature checklists be applied in HCI?

So far FCs have been discussed in relation to the evaluation of interfaces, system software, etc. However, we propose that in addition to this, FCs can be effectively used in other areas of HCI. Figure 2.3 shows a hierarchical structure of how we envisage FCs being used.

Figure 2.4: Hierarchical structure showing general aim and uses of FCs.



The general aim of FCs is given at the first level, i.e. to identify user behaviour, knowledge and opinions relating to the system features (box [1]). The second level shows that this information can then be used in two main ways, i.e. for system bug detection and to increase the effectiveness of user interaction (boxes [1.1] and [1.2]). The latter of these uses can be broken down into individual user feedback (box [1.2.1]) and training course design feedback (box [1.2.2]). These use can now be looked at in greater detail.

“System bug detection” - in order to understand how FCs might be useful for detecting possible bugs in a system, it is necessary to look at the questions (i.e. columns) that we propose to use; these are:

- “Existed?” - did you know this command existed?
- “Used?” - have you ever used this command?

- “How Often?” - how often do you use this command?
- “What for?” - do you know what this command does?
- “Need?” - do you think you will ever have a need to use this command?

By analysing the answers to these questions it should be possible to identify the following types of interface bugs:

- “Information flood” - i.e. cases in which users know that the command exists and what it does, yet expressed little need for using it. If there are many cases of this, the interface may be swamped by features unwanted by the user.
- “Guessability” - i.e. cases in which users know that the command exists but don’t know what it does.
- “Reminding” - i.e. cases in which users know that the command exists, and what it does, yet expressed that their need to use is greater than their actual usage. This is perhaps a problem of being reminded at the right point in the relevant task.
- “Memorability” - i.e. cases in which users have used the command at some point but can’t remember what it does.
- “Information delivery” - cases in which users know what the command does and judge it to be useful, yet are unaware that it existed in this interface.

Table 2.3 (below) summarises these bugs and the answers that would indicate their existence.

Table 2.3: Summary of bugs that FCs may detect and answers that would indicate the existence of these bugs

Bug Type	FC Column Answers
Information flood	(“Existed?” ✓ + “What for?” ✓) - “Need?” ✗
Guessability	“Existed?” ✓ - “What for?” ✗
Reminding	(“Existed?” ✓ + “What for?” ✓ + “Need?” ✓) > “How often?” ✗
Memorability	“Used?” ✓ - “What for?” ✗
Information delivery	(“What for?” ✓ + “Need?” ✓) - “Existed?” ✗

It is reasonable to expect that information flood and guessability would be the most common type of bug; studies 6 and 7 (chapters 9 and 10) investigate the issue of bug detection. If FCs are a suitable instrument for detecting the existence of the design bugs described, then they may provide system designers with useful information on the appropriateness of command names, the location of icons, etc.

However, it is unlikely that major design changes would be based purely on feature checklist information. Instead we propose that after possible bugs have been highlighted, these should then be followed up by interviewing a small, representative sample of users. It should be noted that this does not contradict the criticism made earlier about using a combination of electronic data logs and semi-structured interviews (section 2.3.3); in the present example, possible bugs have already been identified by applying FCs to a large sample and the interviews are merely to confirm/discount these bugs (rather than use the interviews to obtain user information about knowledge and opinions of commands). In fact this represents a crucial aspect of applying FCs. Because they have the potential of being a cheap survey instrument they can gather the frequency information (of both usage and some problems) that can focus attention on where the important aspects of a design are, and so direct other instruments that yield better detail but could not be applied across many users, tasks, and situations.

However, a key issue is not only to detect problems, but also to:

- (i) Grade them by cost to the user.
- (ii) Weight them by frequency of occurrence.

Most methods entirely ignore the second of these. This is because until recently, much of HCI assumed that user tasks were known, and systematically ignored the workplace and actual use and practice. A practical first step in remedying this is to obtain survey information on tasks, or at least command usage, from as complete a population as possible. FCs are the best chance of this and in addition may detect some of the bugs already described.

“Increase effectiveness of user interaction” - since FCs obtain information about a number of aspects relating to users’ interaction with the features in a system, they may also be useful for assessing the effectiveness of this interaction. If this is the case then it may be possible to increase the effectiveness of users’ interaction; this could be done in two ways:

(i) “individual performance feedback” - after administering FCs to a group of users of a particular system it will be possible to measure their own individual usage, knowledge and opinion of each feature in that system. From this we can identify a number aspects of their performance that may be usefully highlighted to users; these include:

- commands not used by the user but judged useful by the investigator e.g. in a word-processing application users may never use a “Table of Contents” and instead manually type this in.
- commands that users claim they never have a need to use, e.g. font styles etc.

Having highlighted performance aspects such as these, it may be useful to give individual feedback to users that could increase the effectiveness of their interaction i.e. feedback illustrating how to use commands such as “Table of Contents” and suggestions to remove commands that are not needed and may cause distractions. Study 7 (chpt. 10) looks at this issue.

(ii) “training course design feedback” - one other possible application of FCs concerned with increasing the effectiveness of user interaction, relates to computer training courses. Feature checklists may be usefully employed at the end of training courses to assess what users have gained (learned) from the course; this will obviously relate to the purposes of the training course. For example, if the aim of a training course is to teach users the basics of word-processing, then the feature checklist could be used to assess whether users had used the basic commands and were aware of their function. With detailed training courses, that allow users to explore the system more freely, FCs could be used to identify areas of the system where users had a poor knowledge of feature functionality; this could then feedback into future training courses.

2.4.1 *Types of users*

An important aspect of FCs that needs more discussion, concerns the types of users that feature checklist should be used with; this is obviously related to the purpose of using them and the situations in where they are used.

In order to be a useful bug detection instrument, FCs must be completed by experienced users that have used the system frequently over a period of time, so that information can be gained about a large number of features used in normal, everyday interactions. What

this is really saying is that in order to evaluate the features in a system fairly, the system itself must be used and explored.

With respect to identifying and increasing the effectiveness of an individual's performance with a system, it is not necessary for the individual to be an experienced user. However, the individual will have to have used the system on at least several occasions in order to have explored some of the interface.

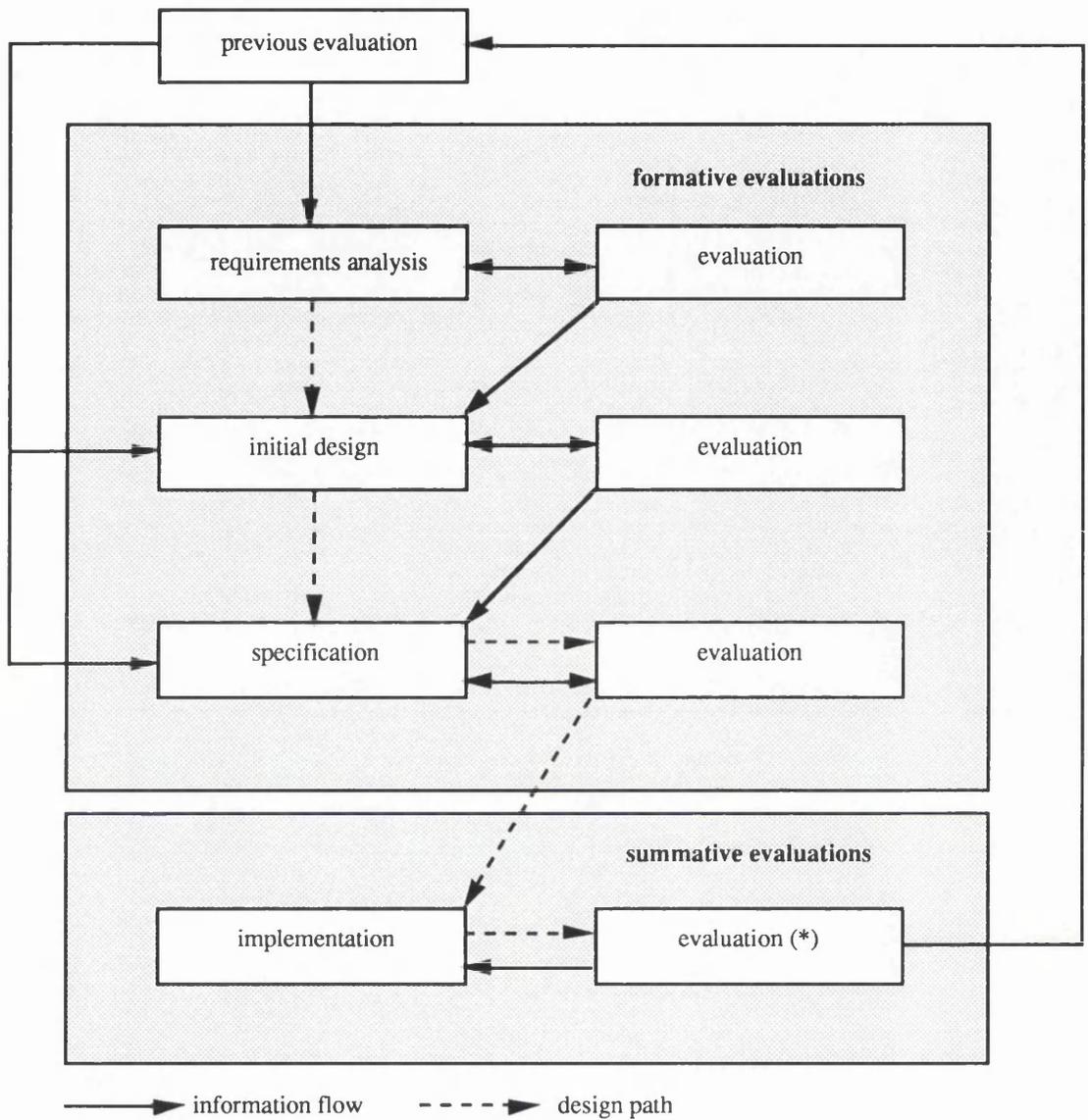
Finally, with respect to increasing the effectiveness of training courses, it is obvious that the feature checklist will be used with users that have limited experience of the system.

The first of these three user types relate to measuring aspects of the system, whilst the last two relate to measuring aspects of the users performance (N.B. the system and user performance are of course heavily inter-related).

2.4.2 *Where should feature checklists be used in the development cycle?*

Based on our proposals concerning the use of FCs as a bug detection method (i.e. experienced users; actual working system, etc.), it should be clear that FCs will be applied towards the end of the typical HCI development cycle (Sharratt 1990); this is shown below in figure 2.5:

Figure 2.2: Evaluation in relation to the development cycle (Sharratt 1990)



In this development cycle, FCs will be applied at the evaluation marked by the asterisk (*), i.e. a summative evaluation that leads back into a re-designed or up-dated system.

2.5 Summary

To summarise therefore, FCs are at present an unused method in the field of HCI. We propose that there is a need for FCs in HCI research because they are likely to have the following advantages:

- They rely on users' memories and so can draw upon interactions in normal work situations.
- They are a cheap method for the investigator to design and administer (in terms of time).
- They should be a relatively cheap method for the user to complete (in terms of time and effort).

These properties mean that FCs have the following advantages:

- They are well suited to "real-life" HCI, evaluation research.
- They can be used with large samples of users at little extra cost, and hence survey populations not a few individuals of unknown representativeness.
- They can obtain various types of information about users' interaction with a system and as a consequence have various uses.

We have suggested that FCs can be used to:

- Detect possible design bugs which can then be focussed on in subsequent evaluations.
- Identify and improve the level of user performance with the system.
- Identify possible weaknesses in training courses and suggest possible improvements.

However, before FCs can be used for any of these purposes, their accuracy must first be established; this is the main concern of this thesis.

Since FCs are a retrospective measurement method and since human memory has in many cases been shown to be unreliable, it is important that we look at the relevant

literature on human memory. In addition it is worthwhile looking at relevant HCI research relating to user behaviour and evaluation of system features.

Chapter 3

Literature review

In this chapter, we provide a review of the most nearly relevant literature to FCs. This literature review is broken down into two main sections:

- (1) An HCI literature review.
- (2) A human memory literature review.

3.1 HCI literature

In chapter 2 it was shown that FCs are as yet, an unused measurement instrument in HCI. As a consequence, this literature review will look at studies in which we think FCs could have been effectively employed. In these studies FCs should not be seen as a direct replacement for the actual methods used, since they obtain different types of data (e.g. number of times a feature was selected instead of time taken to make correct selection). The intention is to show that FCs may have been a useful alternative, relating to the “overall” purpose of the studies described e.g measuring users memory, improving system design, etc.

The HCI literature can usefully be broken down into three sections:

- Studies concerned with command usage.
- Studies concerned with evaluating textual and graphical features of interfaces.

- Studies concerned with HCI design methodology.

3.1.1 *Command usage*

“The nature of expertise in UNIX” - Draper (1985)

This paper discussed the nature of expertise in “UNIX”; it challenged the common sense view that there are experts and novices and the accompanying set of beliefs, e.g. experts know more than novices and also know things that novices do not, etc.

Over a period of 8 months the researcher collected data on the commands used in the system, specifically “the number of distinct commands that that person used at least once”. The data was collected from a laboratory computer, using an in-built logging facility which recorded every process that was run and who ran it; this data was collapsed every night and a cumulative record was kept of individuals versus commands. In total data was collected from 94 users. The “UNIX” system under investigation had around 570 commands and of these, “only 394 were recorded as used at least once by at least one person ... the largest vocabulary recorded for a single individual was 236”.

The main conclusion of the study was that “while there are certainly some users with a larger command vocabulary than others, experts’ real skill seems to lie less in familiarity with the whole command set than in discovering skills that allow them to find answers to the questions they cannot answer from memory”. The implication of this finding for providing help to users was discussed.

The major aspect of this paper that is of interest to this thesis is the drawbacks of the data collected as a result of employing this technique. Draper states that “there are twin potential problems with estimating command vocabulary from such data, that is with equating observed use with the vocabulary known by the person”. The first of these problems involves cases where users know a command and what it does but don’t use it during the period of the study; however since this study had a relatively long data collection period (8 months) this problem would have been compensated for. The second problem involves the opposite scenario, i.e. cases where users have used a command but do not know what the command does (perhaps because it was invoked accidentally). Although there was no adequate method for dealing with this problem,

the researcher concluded that there was no reason to suspect that a systematic error of this kind occurred.

From the discussion of FCs in the previous chapter it can be seen that neither of these potential problems would have been a concern if FCs were used as an alternative to the in-built logging in this study. Using the FC it would have been possible to identify what commands users had used and whether they knew what the command was for; as a consequence, FCs would have been a useful method for measuring users command vocabulary.

Employing FCs in this study would also have significantly reduced the amount of investigator time and effort. Potential problems associated with employing FCs in this study however, would be the amount of subject time and effort required (i.e. completing a FC with 570 features), and the accuracy of FC data (i.e. would subjects remember what commands they had used etc. out of a maximum of 570, over an 8 month period?).

“Information flow in a user interface: the effect of experience and context on the recall of MacWrite screens” - Mayes et. al. (1988)

This paper was concerned with discovering and characterising what users know i.e. what their expertise consists of. Rather than take the view that expertise means “knowing the commands a system offers to users: their names, what they do, and how to use them to carry out useful tasks”, the authors of this paper argued that expertise “consists in the fluid use of a flow of information rather than in its permanent retention as knowledge”.

In order to investigate this theory, this study was designed to probe what users of varying experience could recall of the “MacWrite” interface. The study contained 15 subjects split into 3 groups of 5 (occasional users, intermediate users and frequent users). Subjects worked through a questionnaire that led them through the whole process of using “MacWrite”; at various points through this questionnaire subjects were asked “to recall exactly what would be on the screen and to record this on paper in as much detail as possible”. The main finding from this study showed that “overall recall performance of even our frequent users was surprisingly poor”.

A follow-up study was conducted to investigate to what extent a failure to recall some feature of the interface was reflected in the subjects’ use of that feature in actual

performance. This study contained 5 subjects in total (2 occasional, 1 intermediate and 2 frequent users). Subjects were asked to create a short document using “MacWrite” by performing the tasks listed on a questionnaire. The results showed that all subjects had very little difficulty in performing the tasks; more importantly “items that could not previously be recalled, or items that were put in the wrong place, or confused in some other way were generally found without hesitation and used with ease when creating the document”.

The results were discussed in relation to:

- The accessibility of information (e.g. Tulving’s “encoding-specificity hypothesis”).
- The information flow theory of human action (i.e. action organised around a flow of information picked up from the environment during execution).

Yet again we can argue that FCs could have been usefully employed in this study. Since the “real-life” task in using “MacWrite” is to select commands from a set of possible alternatives (i.e. recognition) it is unsurprising that recall in this study was poor. A more appropriate method of discovering what commands users could remember in this study would have been the FC. Using FCs may have also help to discover the role that visual knowledge plays in users’ performance as opposed to semantic and procedural knowledge i.e. memory may be context-dependent on visual aspects of the interface. The issue of the visual context provided by the interface and the implications for FCs is explored more fully in chapters 5 and 6.

“Errors in training computer skills: on the positive function of errors” - Fries, et. al. (1991)

This paper investigated the effects that errors had in training users computer skills. The authors argued that errors had many positive effects and that training courses should incorporate errors and explain their benefits rather than trying to avoid errors altogether.

Two groups of subjects received a 6 hour training programme on a word-processing system (“WordStar”) followed by a 2 hour testing session. The first group which contained 9 subjects, received “error-avoidant-training” (i.e. training with explicit instructions designed to reduce the chances of errors occurring), and the second group which contained 15 subjects received “error-training” (i.e. no detailed instructions on how to proceed).

Both groups were compared on a number of performance variables, however, the variable of most importance to this thesis was “free recall” i.e. recall of commands known and what they could be used for. The results showed that the “error-training” group was significantly better on a number of these variables including free recall (13.2 commands versus 7.66 commands) and suggested that these subjects had a better organised mental model.

In our opinion, further support for the use of “error-training” might have been found if FCs had been employed to measure subjects knowledge of commands in this study rather than the free recall method. As we have already mentioned in the Mayes et. al. paper, recall is perhaps not the most appropriate way of measuring users’ knowledge of commands since it is not analogous to the “real-life” task, i.e. selecting commands from a set of possible alternatives. FCs on the other hand are a closer match to the “real-life” task since they display the list of possible alternatives and users have to recognise which ones they used, what they are for, etc. Given that this is the case it is likely that using FCs in the present study may have revealed even greater differences between the “error-avoidant-training” and “error-training” groups for memory of commands.

3.1.2 *Textual and graphical features of interfaces*

This section discusses HCI research concerned with evaluating and designing interfaces. The first part of this section focuses on textual features of interfaces whilst the second part deals with graphical features; at the end of both of these sections the role that FCs could have played in this research is discussed.

Evaluating textual features of an interface

The increase in the use of computers generally, has led to a subsequent increase into HCI research on usability, user friendliness, etc. Much of this research has focused on the design of textual features in different systems and the problems associated with the ambiguity of language and the way in which this is interpreted by users, (Rosenberg 1982; Hammond et. al. 1983; Lindegaard and Perry 1986; Perry et. al. 1986).

Two types of ambiguity have been identified in this field; these are:

- Context ambiguity (a person’s expectations provide the frame of reference within which the content of a statement is interpreted).

- Vocabulary ambiguity (confusion resulting from multiple meanings of a single word).

In the former, “research concentrates on gaining an understanding of the so-called user model; that is, the user’s conception of the system and of how it works - the user model thus represents the context in which the user interprets system actions”. In the later, “research focuses on the actual language used between the individual and the computer when engaging in interactive dialogue”, (Lindegard and Perry 1986).

The thrust of this research is to identify words (command names) that have the best “goodness of fit”, i.e. the degree to which a given name suggests the system function it executes. By using various selection and rating methods it was possible to agree on names that users preferred and also show that the least preferred names led to poorer user performance, (Lindegard and Perry 1986).

Given these results, the design of features such as command names is clearly an important HCI consideration, as Lindegard states, “the decision to present a given system to users via menus does not in itself eliminate all the problems that users may face when seeking to operate an unfamiliar computer system” (Lindegard 1987). The design of such menu systems requires evaluation not only during the initial design of a system but also during the design of subsequent versions.

Designing subsequent versions of a system is particularly relevant to FCs. As we explained in section 2.4.2, we propose that FCs are most usefully employed towards the end of the typical design cycle (i.e. to evaluate an existing system and identify enhancements for subsequent versions).

In the research described, FCs would be a useful instrument for identifying the features (textual commands) that were poor at conveying their function to users (i.e. features with low “guessability” - see chapter 9). FCs would also be useful for identifying cases where users were unaware of features’ existence even although they were judged to be useful. This could be due to textual features being located on inappropriate menus.

A final issue about FCs that is worth pointing out here, relates to the research conducted by Lindegard and Perry (Lindegard and Perry 1986). In this paper the authors discuss the usefulness of pen/paper studies in the design of user interfaces, (e.g. easier,

faster, less expensive, etc.). We concur with this view and propose that FCs are an appropriate pen/paper technique.

Evaluating graphical features of an interface

Developments in HCI have resulted in the design and use of graphical user interfaces (GUIs), i.e. interfaces that allow users to select objects or specify operations by directly manipulating objects using mouse or keyboard operations. There have been a number of reasons for the success and growth of GUIs including, “ease of use, ease of learning and increased user productivity” (Rheingold 1989). Of particular importance in this area has been the use of icons in computer environments to represent underlying concepts, objects or tasks e.g. Microsoft Windows (Microsoft 1990). Much of the research in this field has aimed at improving user performance and reducing errors by attempting to “develop guidelines and standards for constructing and using icons” (Kacmar and Carey 1991).

Research conducted on users’ memory for icons and the functions they perform is particularly relevant to this thesis. One conclusion that has been made is that “subjects exhibited increased recall if they are able to form a meaningful association with the message being conveyed by the icon” (Lansdale, et. al. 1990). The tasks involved in this type of research typically involve recognition (matching icons with functions), selection (associating concepts with icons), and recall. We can now look at how FCs might be a useful addition to this type of research.

Once a system has been installed and used in the working environment for a period, it would be possible to use FCs to obtain information on users’ knowledge and usage of the features (icons) in that system. Using this information it would be possible to identify cases where users knew that an icon existed but did not know what it was for (i.e. what it did). If this was found to be the case across a number of users then this might indicate a “guessability” problem with that icon; this could be backed up by interviewing these users. This method would be an alternative to experiments where users are asked to perform a task by selecting the appropriate graphical feature from a possible set and are tested on their accuracy of selection and time required to make a selection (e.g. Kacmar and Carey 1991).

FCs could also be used to identify cases where users know an icon exists and judge it to be useful, yet very rarely or never use it. In chapter 9 we term this bug type “reminding”. If such cases exist then they may indicate that the location of icons may

be an important factor affecting whether users remember to use them, e.g. due to inappropriate grouping with other icons or because the icon is located at the bottom of the list, etc. This idea has been investigated using alternative methods to FCs, e.g. one study using reaction times and selection accuracy showed that “the use of some icons in traditional rectangular menus may not be as effective as when they are used in pie menus” (Callahan, et. al. 1988).

3.1.3 *HCI design methodology*

Many of the evaluation instruments that we have described in chapter 2 relate to interface design that occurs early in the design process where the interface designer is starting from scratch. As we have already mentioned in section 2.4.2, we propose that FCs should be used towards the end of the typical design process i.e. to obtain information about existing systems that can then feedback into subsequent designs/versions. This type of system design is often referred to as “usability testing” (Open University 1990).

Improvements or updates to existing systems are a common and important aspect of HCI design and are necessary for a number of reasons including: changes in technology, changes in interface standards, market competition, etc. As a result “a different set of needs confront the designer who is updating a programme that has an existing, well-known interface” (Telles 1990). In many cases however, re-designing existing systems presents additional challenges to those encountered in designing from scratch. The following list describes the challenges that the re-design (or updating) process faces (Telles 1990):

- The existing user base - “old” systems already have an experienced and opinionated user base; any changes will impact on these users. (N.B. some of these users may not want change since they have invested substantial amounts of time learning the existing system).
- Learning new commands for old functions - to the user, functionality and commands are difficult to separate; these users are extremely reluctant to learn new commands for existing functions.

- The value of the current interface - no matter how bad the old interface appears, it contains many excellent elements. Existing users therefore need to be consulted to identify what works in the existing interface.

It is our view that FCs are well suited to these additional challenges. FCs obtain information directly from users of the system; any design bugs detected or system enhancements identified will be done so on the basis of this information. Since the users are the main drivers behind design changes, it is therefore unlikely that popular or well understood features will be significantly changed. Instead it is likely that FCs will highlight:

- Features that users have no need to use (“information flood”).
- Features that users do not know exist (“information delivery”).
- Features that users do not understand i.e. don’t know what the feature is for (“guessability” or “memorability”).
- Features that users forget to use when needed (“reminding”).

As we have already mentioned once the FC has highlighted these bugs, other instruments such as interviews should then be used to assess the validity and seriousness of these bugs before design changes are made.

Two aspects of HCI design methodology that are particularly relevant in usability testing are: incorporating user defined feedback, and conducting iterative designs (Good et. al. 1986). As we have already discussed, FCs are well suited to the former of these aspects; we can now discuss the role of FCs in iterative design.

In chapter 2 we discussed a number of problems involved in conducting HCI evaluation in “real-life” settings, e.g. time constraints, artificial settings, samples sizes, etc. As a consequence of these problems (or resource constraints) it is impossible to fix all design bugs immediately and design must therefore be iterative. Early stages of the design cycle focus on avoiding serious problems/bugs that may occur. Iterative design however, focuses on identifying and fixing less serious problems/bugs that occur more frequently e.g. users selecting incorrect commands, or users not being able to perform various functions.

As we mentioned in section 2.3.1, FCs obtain information about “low-level” features of an interface such as menu commands. Before producing re-designs (or prototypes) of systems, designers need estimates on the frequency of usage of features within that system. In practice however, there is little evidence of this being taken onboard. A possible reason why this is the case could be that no practical method exists for obtaining this information in “real-life” settings; as we have already described FCs could be a solution to this problem.

An important aspect of HCI design that has emerged recently is the increasing number of features contained in systems, sometimes termed “creeping featurism”. Whilst the intention behind these new features is to help to increase the effectiveness of a user’s interaction with a system, there is a danger that too many features might result in negative effects such as distracting users. In relation to FCs we have called this problem “information flood” i.e. users know that the command exists and what it does, yet express little or no need for using it. As a result of the data obtained on users knowledge of commands’ existence and function, and their need to use them, FCs are an ideal instrument for detecting this problem.

HCI Design methodology examples

It is useful to look at some alternative design methodologies to see how FCs might relate to them. Whiteside et. al. (1987) have suggested the following method:

- (1) Ask users to perform a specific task.
- (2) Monitor users during free use (logging and/or observation).
- (3) Give users a questionnaire.
- (4) Interview users.
- (5) Survey users.
- (6) Ask users for critical incidents revealing successes or failures.

Although this example does not define what information is obtained at stages 3 and 5 (this will obviously be dependent on the system being evaluated), it is highly likely that the FC could be used to replace at least one of these whilst at the same time obtaining information on users free use of the system (stage 2). After administering the FC users could be interviewed about issues arising from the FC data as well as any critical

incidents. Employing the FC in this example would therefore reduce the amount of instruments needed (with subsequent reductions in time and effort) and would also allow the evaluation to take place in “real-life” situations.

Tyldesley (1990), describes a hypothetical evaluation plan that involves a less formal methodology. In this example changes are to be made to an interface (system C) with a large installed base of users with a minimal disruption to their work. In addition “users must perceive the new version to be more usable than previous versions”. The following evaluation plan is proposed (table 3.1):

Table 3.1: Hypothetical evaluation plan (Tyldesley 1990)

Factor	Goal
Scope	The new “baselevel” of system C is to be installed and tested. System C is still being built.
Intended users	These are experienced users of system C, but an older version. They are customers. As large a sample size as possible is required.
Metrics	System usability, system consistency, customer comments.
Research or development	Direct input to development team. Time is crucial-4 weeks only before documentation is “frozen”. The product is subjected to non-disclosure to customers.
Methods	Benchmark tests modified by a researcher “sitting-in” to answer questions and to give hints; videotaping, laboratory based-trials.

Rather than employ the instruments in this plan, we suggest that FCs would be a more useful and appropriate alternative. Information could be gained from a large sample of users about their normal everyday interaction with the system. This information could then be used to highlight possible bugs (already mentioned) which could then be verified by interviewing a representative sample of users. During this interview there would also be scope for asking users about any other comments they might have. The advantages of employing this approach over the one described are:

- Data from large sample of users (ensures representativeness of data).
- Data from “real-life” working environment.
- Inexpensive for both investigator and users in terms of time and effort.

- Evaluation can be done quickly and with little disruption of normal routine of users.

The disadvantages of employing this approach over the one described by Tyldesley are:

- No quantitative measure of users' errors.
- No quantitative measure of users' attitudes, etc.

A final aspect concerning FCs and HCI design methodology relates to the constraints of conducting "real-life" evaluation. As we have already seen these constraints limit the number of instruments that can be feasibly employed. The only possible alternative to FCs (i.e. an instrument that can be used with large samples of users), is questionnaires. However, questionnaires at present are notoriously feeble at bug detection since they rely on users attitudes and opinions which are susceptible to change.

3.2 Human memory literature

In conducting experimental research on FCs it is necessary to examine the research carried out in the field of human memory in order to try to provide initial hypotheses, and estimates of how general the findings of FCs are likely to be. As we have already mentioned the success of FCs depends on subjects being able to recognise the features listed so that they can then be used as cues to elicit various types of information from the users memory. Since memory research has shown that recognition is usually an easier task than recall and that recognition performance is superior to that of recall, we would expect FCs to be more accurate than traditional questionnaires (study 1 addresses this issue). This is in line with the shift in both interface design and usability research in HCI over the last decade, from recall based designs to recognition based designs. However, as we shall see in later studies, the FC does not rely solely on simple recognition; the actual type of memory process involved depends on the type of information that the different columns are asking, and the way in which it is encoded.

There are a number of problems relating to memory research that will be discussed in later sections of the thesis, however, it may be useful to mention these briefly here. Firstly, there is the problem of defining terms such as recognition. Reber (1985) defines recognition as, "The awareness that an object or event is one that has been previously seen, experienced or learned"; this definition, however, could equally apply to the term recall. Another problem is that there is not necessarily a direct relationship

between the external task and the mental method used (this is discussed more fully in section 4.4.1).

A final problem, that has already been touched upon, is that it may be inappropriate to talk of tasks simply in terms of recognition; i.e. it may be that all tasks involve recall to an extent ($A \rightarrow B$ associative retrieval from memory). The important differences for FCs are likely to be:

- The amount of information in A that is in B.
- Whether the subject was using (or expecting to use) that link in the learning situation.

Since the aspects of human memory relating to this thesis are recall, recognition and encoding processes, much of the discussion will refer to the work of Endel Tulving.

At this point however, it is important to emphasise that the primary aim of this thesis is to develop FCs as a measurement instrument in HCI. As a result the subjects in these studies performed various everyday tasks on a computer and were unaware that they would be asked questions about the commands they used, etc. The learning was therefore to an extent “incidental learning”, i.e. subjects were not specifically intent on learning (memorising) as many commands as possible. This is in sharp contrast to experimental studies of recognition and recall in which subjects are told to try and learn/study words in a list so that they can be tested on them later.

The problem that arises, therefore, is that since the studies that follow were not specifically designed to test memory processes, it becomes difficult to say exactly what processes were involved. This problem is however, secondary to the primary aim of the thesis namely, the development of FCs as an HCI measurement instrument.

In addition it is worth pointing out that in the last decade arguments have been made by a number of psychologists, criticising the laboratory based approach to the study of human memory that has prevailed for the past 100 years (e.g. Neisser 1982; Linton 1975). In particular, strong arguments have been made for the naturalistic study of human memory, i.e. “conducting research on memory in naturally occurring conditions, whether that be in the home at work or wherever”, (Neisser 1982). As mentioned earlier, many of the studies described in this thesis involved users performing everyday tasks on computers in natural settings e.g. word-processing

classes. It may be that the findings of these studies could be as revealing about human memory as many of the laboratory ones were.

Chapter 4

A comparison of the feature checklist and the open response questionnaire in HCI evaluation

4.1 Introduction

The preceding chapters have described a measurement instrument known as the feature checklist (FC). These chapters have shown that FCs are at present unused in HCI, despite the fact that they may be a useful measurement method.

An earlier study conducted outside HCI showed that an appreciable difference in yield exists between the FC and an alternative method (the Open-Response Questionnaire - ORQ), even for behaviour as recent as the previous day's (Belson and Duncan 1962). These researchers concluded that, "the sharp differences in yields from the two methods makes it quite clear that at least one of them can be seriously in error when used to assess the previous day's behaviour", and that "in the circumstances validation of each of these methods becomes a pressing issue."

In the Belson and Duncan study, the behaviour in question was the publications (i.e. newspapers, magazines, etc.), that people had looked at the previous day and the television programmes that people had seen the previous day; the features were therefore, either the publications or television programmes. In the present study the behaviour in question was the selection of commands from menus in a word-processing package ("Microsoft WORD 5.0") in the previous day's study; the features were therefore, the menu commands.

With regard to human memory, the actual task in this study (i.e. selecting commands from menus), is in fact a recognition task. As a result, we would expect the FC to be a more accurate instrument than the ORQ for assessing what commands were used or unused, since generally speaking the former involves recognition and the latter involves unprompted recall, (Mayes et. al. 1988).

4.1.1 *The importance of assessing FC validity*

This issue of validation is central to the argument made for FCs in chapter 2. The usefulness and advantages of FCs can only be demonstrated once their validity (i.e. accuracy) has been established, i.e. before the data from FCs can be used for bug detection and measuring user performance, the accuracy of this data must be examined. This study is an attempt to establish the accuracy of FC data relating to subjects' usage of features.

As well as addressing the issue of validation, i.e. "the degree to which the instrument of measurement does indeed measure what it purports to measure", this study also applies the methods in an HCI setting.

In order to help us assess the accuracy of both instruments, "Tempo II Plus" was used as an electronic data log to record all the command selections that subjects made. However, since the accuracy of data logs may also be suspect, a detailed observation of each subject's command selections was also made by the experimenter throughout the study. An additional aim of this study was therefore to assess the usefulness of electronic logging techniques as an alternative method of measuring command usage in HCI.

The experimental hypothesis is: "that the FC will be a significantly more valid instrument than the ORQ for measuring subjects' memory of menu commands".

4.2 Method

Subjects: There were eighteen subjects in total; nine in group A and nine in group B. Of these eighteen subjects, eleven were female and seven were male; their ages ranged from 15 to 37 years of age with a mean of 22.8 years. All of the subjects were

recruited from the psychology summer school at the University of Glasgow and were informed that they had to participate in a study as part of their summer school course; subjects were told that they would be paid £4.00 for doing so. All of the subjects had little or no experience of using computers, this was assessed by asking subjects to complete a computer experience questionnaire that was administered in the recruiting phase of the study, (appendix 4.1).

Apparatus/Stimuli: The study was run on an “Apple Macintosh IIsi” and displayed on an “Apple Macintosh A4 Mono Monitor”. The word-processing application “Microsoft WORD 5.0” was used in this study and all the computer operations that subjects performed were recorded using the “Tempo II Plus” application. However, the normal “Microsoft WORD” interface was not used, but instead used as a platform for an artificial one. Since the subjects had not used a mouse before, the menus were keyboard operated in order to make errors much more clearly observable; the command and menu names were a specially designed subset.

Design/Procedure: The subjects were told that their task in this study was to select commands from the menus on the screen as the experimenter read them out. The experimenter demonstrated to each subject how to open a menu and choose a command using the appropriate keys in the correct order; the selection and sequence of the keys used were as follows:

- (i) <enter> to highlight the menu bar
- (ii) numbers 1 - 7 on the numeric keypad to open the relevant menu i.e. (1=File, 2=Edit, 3=Format, 4=Font, 5=Document, 6=Utilities, 7=Window)
- (iii) “up” and “down” arrow keys to highlight the relevant command
- (iv) <enter> to select that command once it has been highlighted
- (iv) <clear> to clear any dialogue boxes that might appear.

Subjects were allowed to practice these procedures for several minutes until they felt comfortable on a set of practice menus. When subjects were happy with this, the experimenter opened a new document that contained a different set of menus with new commands and initiated “Tempo II Plus”. The experimenter then proceeded to read out commands for subjects to select.

In total 112 commands were read out, this was split into two sessions of 56 with a 5 minute break in the middle to prevent boredom and fatigue of subjects. During this break subjects were given a questionnaire on using the keyboard to complete, (appendix 4.2).

Each menu contained 5 commands giving 35 commands in total. The commands were selected a different number of times ranging from 0 times, up to and including 7 times. The study was designed so that there were 16 selections made in each menu (giving a total of 112 selections). The actual layout of the menus, the number of times each command was selected and the required response by each subject is given in appendix 4.3. After this part of the study was completed, the subjects were asked to come back at the same time the following day to complete the second part of the study.

In this second part of the study, the subjects were given a couple of questionnaires to fill in; these questionnaires asked about the commands that they had selected in the previous day's study. The subjects in group A were given the ORQ first (appendix 4.4), and the FC second (appendix 4.5), and the subjects in group B were given the FC first and the ORQ second. After completing these questionnaires, all the subjects were given a payment of £4.00 and thanked for their participation. They were also debriefed as to the nature of the study.

4.3 Results

When the ORQ was issued first, subjects remembered on average 16 of the 35 commands; this increased to 19 commands when the ORQ was issued second, i.e. after the FC. Since some of these command names were only partially recalled, an assessment was made by the researcher and two independent assessors in order to determine their accuracy, i.e. whether or not these commands actually existed. Examples of partially recalled commands that were accepted by the researcher and the independent assessors as legitimate are shown below in table 4.1.

Table 4.1: Examples of incorrectly recalled and partially recalled commands by subjects using the O.R.Q.

Partially recalled commands	Actual commands
Repaginate	Full Repaginate Now
Remove heading	Demote Heading
Open New File	Open Any File
Go To Page	Go To
Compact Selection	Collapse Selection

Examples of incorrectly recalled commands by subjects include: “Page Set Up”; “Save”; “Finish” and “Copy”. None of the above commands were accepted as legitimate by the researcher or the independent assessors.

Table 4.2, shows the recall of commands by subjects using the ORQ and how accurate this recall was.

Table 4.2: Mean number of commands recalled by subjects using the ORQ

Instrument	Correctly Recalled	Incorrectly Recalled	Total Recalled
ORQ 1st	14.8	1.2	16
ORQ 2nd	16.9	2.1	19

We can now compare the ORQ with the FC on each subject’s recall of command usage (table 4.3). When the FC is issued first, subjects correctly recalled whether they had used an average of 30.4 out of 35 commands (i.e. 87%); however, when the FC was issued second (i.e. after the ORQ), this dropped to an average of 26.9 commands (i.e. 77%). When the ORQ is issued first subjects correctly recalled whether they had used an average of 14.8 commands (i.e. 43%) and when the ORQ was issued second (i.e. after the FC), this increased to 16.9 commands (i.e. 48%).

Table 4.3: Subjects' recall of command usage using the ORQ and the FC

Instrument	Correct Usage	Incorrect Usage	Don't Know
FC 1st	30.4 (87%)	2.3 (7%)	2.2 (6%)
FC 2nd	26.9 (77%)	4.4 (13%)	3.7 (10%)
ORQ 1st	14.8 (43%)	1.2 (3%)	0.0 (0%)
ORQ 2nd	16.9 (48%)	2.1 (6%)	0.0 (0%)

A two-factor analysis of variance was used to estimate the effect of instrument type and order on correct usage (i.e. the number of commands that subjects correctly remembered using). In the ANOVA table (table 4.4), Factor A is the Instrument (FC or ORQ) and Factor B is the Order of presentation (1st or 2nd).

Table 4.4: ANOVA table for effects of instrument type (factor A) and order (factor B) on correct usage

Source:	Sum of Squares:	df:	Variance:	F - test:	P value :
Factor A	1482.25	1	1482.25	121.137	.0001
Factor B	4.694	1	4.694	0.384	.54
Interaction AB	72.25	1	72.25	5.905	.0209
Within Groups	391.556	32	12.236		

The obtained F value for Factor A, 121.137 does exceed the F of 4.16 for 1 and 32 degrees of freedom at the 0.05 level. We can therefore conclude that the FC does produce significantly more accurate correct usage scores than the ORQ.

The obtained F value for Factor B, 0.384 does not exceed the F of 4.16 for 1 and 32 degrees of freedom at the 0.05 level. We can therefore conclude that the order of presentation of the instruments does not result in significant differences in the accuracy of correct usage scores.

Finally, the obtained F value for interaction (A x B), 5.905 exceeds the F of 4.16 for 1 and 32 degrees of freedom at the 0.05 level. We can therefore conclude that the combined effects of instrument type and order significantly effect the accuracy of subjects' correct usage scores.

After conducting a Tukey test to compare the interaction means (appendix 4.6) it was found that the FC was significantly more accurate than the ORQ on correct usage scores in all cases, regardless of the order of presentation. There were no significant differences in the accuracy of correct usage scores between the same instruments when their order was varied (FC 1st vs. FC 2nd) or (ORQ 1st vs. ORQ 2nd), i.e. there were no significant order effects.

4.4 Discussion

The results support the hypothesis that the FC is a more valid (i.e. accurate) instrument than the ORQ for measuring subjects' memory of commands that they selected. Furthermore, the FC also appears to possess high validity in itself, i.e. when the FC is presented first subjects could remember whether they had used on average 87% of all commands. This suggests that the FC could be a useful instrument for measuring command usage in HCI.

The most obvious reason why the FC was apparently so successful is that it requires only recognition, while the ORQ requires unprompted recall. This notion can best be explained by looking at "Generation-Recognition" (G-R) models of human memory.

4.4.1 "Generation-Recognition" models of memory

One of the most obvious facts about human memory is that research has shown that it is generally a much easier task to remember previous events or experiences when memory is tested by recognition rather than recall. However, before relating this issue to research on FCs it is important to try and define the terms recall and recognition; this is not as easy as first seems, as Brown (1976) states, "care is necessary when classifying tasks as recall or recognition".

The terms recognition and recall apply independently both to test situations and to memory processes, and there is no necessary correlation between the formal characteristics of a test situation and the processes it evokes in a subject. That is, the external task given to a subject may be recognition or recall, but the internal mental process they use may be either one. For example, a subject could do recall by generating many candidates and recognising the right one, or do recognition by

recalling the item and checking it against the cues (e.g. in police identity parades). This may be the case with FCs, i.e. what at first appears to be simple recognition may in fact involve recall processes to greater or lesser degrees.

Brown (1976), makes a distinction between recall and recognition tests: “the essence of a recall test is that the subject has to generate the target/s meeting the definition of the target in the recall instruction”, whereas “the essence of a recognition test is that one or more potential targets are presented to the subject, there is no requirement for overt generation of the target and the response may consist in accepting/rejecting a given choice, rating it, etc.”.

This definition relates to a number of theories of memory which can generally be grouped under the title of G-R models of memory and which are useful for explaining the phenomenon of recognition being superior to recall.

A particularly influential view that has been put forward is the the “two-process theory” (Watkins and Gardiner 1979). Although different versions of the theory have been proposed (e.g. Anderson and Bower 1972; Kintsch 1970), they all have the following in common:

- Recall involves a search or retrieval process, which is followed by a decision or recognition process based on the apparent appropriateness of the retrieved information.
- Recognition involves only the second of these two processes.

It can be seen from this that recall involves two fallible stages whereas recognition involves only one fallible stage. Recall will only occur when an item is both retrieved and then recognised. A study conducted by Bahrick (1970) showed that “the level of cued recall was predicted reasonably well by multiplying together the probability of retrieval by the probability of recognition”. Other evidence has shown that people can recall information by making extensive use of the retrieval process and then deciding which of the items produced by the retrieval process are appropriate (Rabinowitz et. al. 1977).

The G-R theories described here would seem to be useful for explaining the results obtained in study 1. The reason why the FC was significantly more successful than the ORQ is because the FC involved only one fallible stage (a decision-making stage)

whilst the ORQ involved two fallible stages (a search/retrieval stage and a decision-making stage).

However, a number of criticisms have been made of the G-R theories. In particular it has been shown that there are occasions in which recall is superior to recognition memory, e.g. (Watkins 1973; Tulving and Thompson 1973). One way in which these findings were accounted for was that people not only store “to-be-remembered” information in long term memory but also contextual information which was presented at the same time. Both recall and recognition tend to be best when the contextual information present at the time of learning is also present at the time of the memory test.

Another problem with the two-process theory (or G-R models in general) concerns the assumption that there is no retrieval problem in recognition memory, because the retrieval process is not involved at all. However, the fact that recognition memory is susceptible to context effects (Tulving and Thompson 1971) suggests that there can be a retrieval problem in recognition memory.

As we shall see in chapters 5 and 6, perhaps a more appropriate theory for explaining the findings of the FC studies conducted here is the “encoding-specificity hypothesis” (Tulving and Thompson 1973).

4.4.2 FCs: order of presentation and accuracy

An interesting feature that emerged from this study is the difference in correct usage scores between the FC presented first (87%) and the FC presented second (77%). Although this difference was found to be insignificant, it does seem to suggest that subjects become somewhat unsure of their responses when the FC is preceded by the ORQ. This view receives further support when we look at the scores for incorrect usage and don't know. The FC presented second scores higher on both of these indices than the FC presented first, (13% and 10% compared with 7% and 6%, respectively). The implication of this would seem to be that when FCs are used in HCI evaluation to measure command usage, they should be used alone or as the first of a series of measurements.

We can now look in more detail at the incorrect usage scores. From table 4.3, it was seen that when the FC was issued first, subjects incorrectly identified using 7% of the

commands; however, when the FC was issued second this increased to 13%. These incorrect answers could belong to one of the following categories:

- False Negatives - subjects using the command in the study but answering that they hadn't on the FC
- False Positives - subjects not using the command in the study but answering that they had on the FC

Table 4.5, shows the breakdown of the incorrect usage answers into these categories.

Table 4.5: Incorrect usage scores on the FC

Order of Presentation	False Negative	False Positive
1st	2 (9.5%)	19 (90.5%)
2nd	7 (17.5%)	33 (82.5%)
Total	9 (13.5%)	42 (86.5%)

From table 4.5, it can be seen that of the total number of incorrect usage scores to the FC presented first, 9.5% were false negatives whilst 90.5% were false positives. After conducting a t-test (appendix 4.7), it was found that these differences were highly significant. This indicates that when subjects were given the FC first, they were much more likely to falsely assume that they had used a command, than they were to forget using one. To put this another way, it is easier to remember using a command than it is to remember not using one. Of the total number of incorrect usage scores to the FC presented second, 17.5 % were false negatives and 82.5% were false positives; after conducting a t-test this was found to be insignificant. Therefore, although subjects were more accurate when the FC was presented first, any incorrect answers they did give were significantly more likely to be false positives than false negatives.

At this point however, it is worth noting a methodological concern that may limit the applicability of the findings of this study. In the study all the subjects were naive word-processor users; this was because we did not want each subject's previous knowledge of word-processing applications to interfere with their recall in the second part of the study. With expert word-processor users there was a possibility that they may have carried out other word-processing tasks between parts 1 and 2 of this study and this could have affected their recall. Since all of the subjects in this study were naive word-

processor users we can be certain that their recall of menu commands was for part 1 of this study only.

Since it is the intention however, that FCs should be used with expert computer users (when used for bug detection) in order to obtain information in real life situations, the range of usefulness of this study may be limited. Future studies should look at FCs with expert users performing realistic tasks, in order to see if results differ. Although there is no reason to expect so, it would still be useful and logical to look at whether such things as users goals might affect memory.

The results of this study suggest other important issues for future research on the use of FCs in HCI evaluation involving more complex tasks.

Firstly, it may be the case that the validity of FCs could be improved through better design. If the FC was designed to be more visually accurate (i.e. a closer match to the layout on the actual computer screen), it is possible that this may provide more cues to subjects and hence improve recall. This would be consistent with the “encoding-specificity hypothesis” (Tulving and Thompson 1973).

Secondly, it may be possible to extend the usefulness of FCs by asking more questions i.e. column headings. This is a simple addition to FCs that would result in more detailed data at little extra cost to the researcher or the subject. For example, if we are interested in command usage, the additional column headings could ask such things as “did you know this command existed?”, “do you know what this command is for?”, etc. It is important however, that the questions require only a tick or a cross for an answer so that the respondent’s workload is kept to a minimum. The actual questions that should be included will obviously be specific to the features under investigation. Future research could look at the validity of these additional column headings.

Future studies could also look at the reliability of FCs as an evaluation instrument in HCI settings, i.e. “the extent to which they yield the same approximate results when utilised repeatedly under similar conditions.”

4.4.3 FCs and “state-dependent memory”

One final aspect of feature checklists and human memory related to the design of this study concerns the issue of experimental setting and “state-dependent memory” i.e.

“recall of memories of previous events is enhanced if one returns to the original setting of those events”, (Reber 1985). A number of studies have looked at whether changing rooms could exert state-dependent effects on recall and recognition and in one case it was found that a state-dependent effect occurred only when subjects did not know that recognition would be tested later (Smith 1986). Parkin (1993), summarises the research on state-dependent effects: “state-dependent effects thus seem to occur under a variety of different circumstances, but are found regularly only when memory is measured using free recall”. Given that this is the case it was important in the present study that measuring subjects’ memory through the ORQ (free recall) was conducted in the same environment as the learning took place. In this study, both the learning and testing stages for all subjects were conducted in the same room, using the same arrangement and at approximately the same time of day.

4.4.4 *Electronic data logging*

As was mentioned earlier, the “Tempo II Plus” application was used to record the command selections that each subject made. However, one problem with such electronic data logs is that they record all selections made by subjects regardless of intention. As a result, some command selections that appear on the data log may not appear on the feature checklist since subjects were unaware of making them. In order to assess how much of a problem this was in the present study, a careful note was taken of any accidental command selections that the subjects made.

In this study three different types of errors that subjects could make were classified; these were as follows:

- Error Type 1 - the subject selects the wrong command but is unaware of doing so, e.g. subject is told to select “Cut” but instead selects “Copy” although he/she believes that they selected “Cut”.
- Error Type 2 - the subject selects the wrong command, realises this and corrects it; however, they don’t know what command they accidentally selected.
- Error Type 3 - the subject selects the wrong command and realises this and corrects it; they also know which command they selected by accident.

In this study we used the electronic data log (along with the researcher's record of observations) to assess the validity of the FC. At the same time, however, we can also use this record of observations to assess the accuracy of the data log. Before doing so it is worth pointing out that Error Type 3 will not affect the validity of the FC or the data log, i.e. if subjects were aware of the command that they selected by accident they can record this on the feature checklist. However, the problem arises with Error Types 1 and 2. These errors may affect the accuracy of each subject's answers to the question "did you use this command?", but only if the command incorrectly selected was never intended to be selected or intended to be selected only once throughout the whole study. In reality this occurred on three occasions throughout the whole study and the data log was amended accordingly. With regard to the question "how many times did you use this command?", a total of sixteen accidental command selections were made across all subjects and again the data log was amended accordingly. Table 4.6, lists all the selection errors that subjects made in this study; the highlighted cases are the errors that would have affected the question "did you use this command?" on the FC, i.e. commands intended to be selected once or not at all.

Table 4.6: Number and type of selection errors made by subjects

Subject No.	Error Type	Error Description
1	Error Type 1	selected "24 Point" instead of "12 Point"
2	Error Type 1	selected "Help" instead of "All Caps"
3	Error Type 2	selected "Italic Cursor" instead of "Sort"
4	Error Type 1	selected "Find" instead of "Word Count"
6	Error Type 1	selected "Delete Forward" instead of "Print"
7	Error Type 1	selected "Helvetica" instead of "New York"
7	Error Type 2	selected "Glossary" instead of "Word Count"
8	Error Type 1	selected "Word Count" instead of "Go To"
8	Error Type 1	selected "New York" instead of "Helvetica"
9	Error Type 1	selected "Plain Text" instead of "Bold"
11	Error Type 1	selected "All Caps" instead of "Help"
11	Error Type 1	selected "All Caps" instead of "Help"
12	Error Type 1	selected "Open Footer" instead of "Find"
13	Error Type 1	selected "Print" instead of "Close"
13	Error Type 1	selected "Delete Forward" instead of "Print"
13	Error Type 1	selected "Italic" instead of "Underline"

Since all the command selections made by subjects in this study were recorded by the experimenter (including accidental selections), it was possible to identify and eliminate any discrepancies from the data log thus ensuring a completely accurate assessment of the FC's validity. This was possible due to the rigidly controlled (and somewhat artificial) experimental design employed in this study.

In future studies using FCs this control may not be possible (or desirable) and as a result care should be taken in interpreting any results involving data logs. However, since it is envisaged that the main purpose of FCs will be to discover which commands are meaningfully employed by users, careful attention to the wording of accompanying instructions might help to reduce the problem of data log inaccuracy. For example, instead of instructions such as "did you use this command?" we could ask ones like "did you use this command to carry out tasks? Please omit those commands invoked by accident".

An interesting area of the study that was not central to the primary purpose is that concerning the relationships between command usage (i.e. how often each command was selected) and subjects' recall and recognition of commands.

4.4.5 *Unprompted recall (ORQ) and command usage frequency*

With respect to the ORQ, it would be interesting to discover whether or not the commands that subjects recalled were the ones that were most frequently selected. Given the findings of earlier research that experienced word-processor users could only recall those commands that they used often (Mayes et. al. 1990), it seems reasonable to expect so. Table 4.7 shows the usage frequency for each command (i.e. number of times it was selected) along with its respective recall score (i.e. the number of subjects that recalled seeing this command).

Table 4.7: Usage frequency and recall score for each command

Command	Usage Frequency	Recall Score
12 Point	2	17 (94%)
24 Point	1	17 (94%)
Helvetica	6	17 (94%)
Italic	2	16 (89%)
Bold	6	15 (83%)
Print	7	15 (83%)
Open Any File	6	15 (83%)
New York	3	14 (78%)
Times	4	14 (78%)
Help	5	13 (72%)
Underline	4	13 (72%)
Find	7	12 (67%)
Glossary	7	11 (61%)
Go To	5	11 (61%)
Full Repaginate Now	3	11 (61%)
Open	1	10 (56%)
Close	2	10 (56%)

Table 4.7: (continued)

Untitled 1	7	7 (39%)
All Caps	4	7 (39%)
Word Count	3	6 (33%)
Demote Heading	5	6 (33%)
Sort	5	6 (33%)
Commands	3	5 (28%)
Italic Cursor	1	5 (28%)
Plain Text	4	4 (22%)
Quit	0	4 (22%)
Expand Subtext	2	3 (17%)
Collapse Selection	6	3 (17%)
Show Ruler	0	2 (11%)
Spelling	0	2 (11%)
Change	1	1 (6%)
New Window	0	1 (6%)
Open Footer	0	0 (0%)
Show Clipboard	0	0 (0%)
Delete Forward	0	0 (0%)

From table 4.7 it can be seen that 17 subjects or 94% could recall seeing the commands “12 Point”, “24 Point” and “Helvetica”, whilst no subjects could recall seeing the commands “New Window”, “Open Footer”, “Show Clipboard” and “Delete Forward”; (maximum number of subjects was 18). In order to test our prediction of a significant, positive correlation between usage frequency and recall score a Spearman’s Rank Correlation Coefficient was calculated (appendix 4.8). A correlation coefficient of 0.576 was obtained; this is significant at the 0.01 level (one-tailed). We can therefore accept our hypothesis that recall score increases as usage frequency increases, i.e. subjects are more likely to recall commands that are more frequently selected.

N.B. It should be pointed out that although the measures for both variables were at the ratio level, it was noticed from the scattergram that the data did not conform to a normal distribution and as a result the Spearman’s Rank Correlation Coefficient was used in preference to a Pearson Product-moment Correlation Coefficient.

It is possible that less frequently used commands achieved high recall scores because these actual names (words) are used very frequently in the English language e.g. words such as “open”, “close”, etc. In order to see what effects (if any), frequency of occurrence of commands names in the English language had on the correlation coefficient between recall score and usage frequency, a Kendall’s rank correlation coefficient was calculated (appendix 4.9). This calculation showed that even with the effects of language frequency partialled out, there is still a significant correlation between recall score and usage frequency. However, there is no significant effect of language frequency alone on recall score, i.e. recall score and language frequency are unrelated (appendix 4.10).

4.4.6 *Recognition (FC) and command usage frequency*

As far as subjects’ answers on the FC were concerned it was also of interest to see what effect (if any) usage frequency had on subjects’ recognition score (i.e. the number of subjects that correctly identified whether they had selected this command). Again we would expect higher usage frequency to result in higher recognition scores. However, in addition it might also be reasonably expected that commands which were never selected would also have a high recognition score (this is really saying that it is easy to recognise not having selected a command if it was in fact never selected). Given this prediction we would expect results similar to those displayed in figure 4.1, i.e. commands never selected and those selected frequently should have high recognition scores.

Figure 4.1: Hypothesised scattergram showing the relationship between usage frequency and recognition score

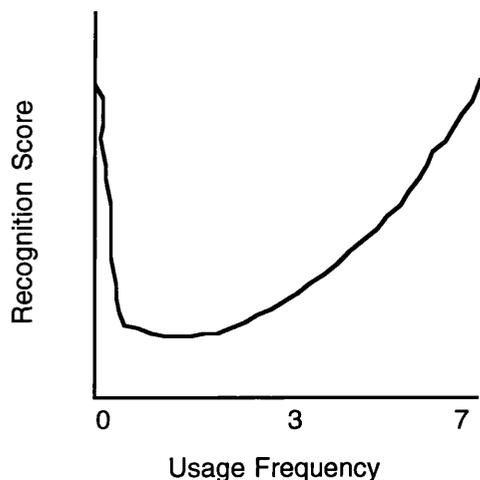


Table 4.8, shows the actual usage frequency and recognition scores for each command.

Table 4.8: Usage frequency and recognition score for each command

Command	Recognition Score	Usage Frequency
Print	18 (100%)	7
Glossary	18 (100%)	7
Helvetica	18 (100%)	6
Bold	18 (100%)	6
Full Repaginate Now	18 (100%)	3
Help	18 (100%)	5
Untitled 1	18 (100%)	7
Open Any File	17 (94%)	6
Italic	17 (94%)	2
Demote Heading	17 (94%)	5
12 Point	17 (94%)	2
Times	17 (94%)	4
Find	17 (94%)	7
Go To	17 (94%)	5
Collapse Selection	16 (89%)	6

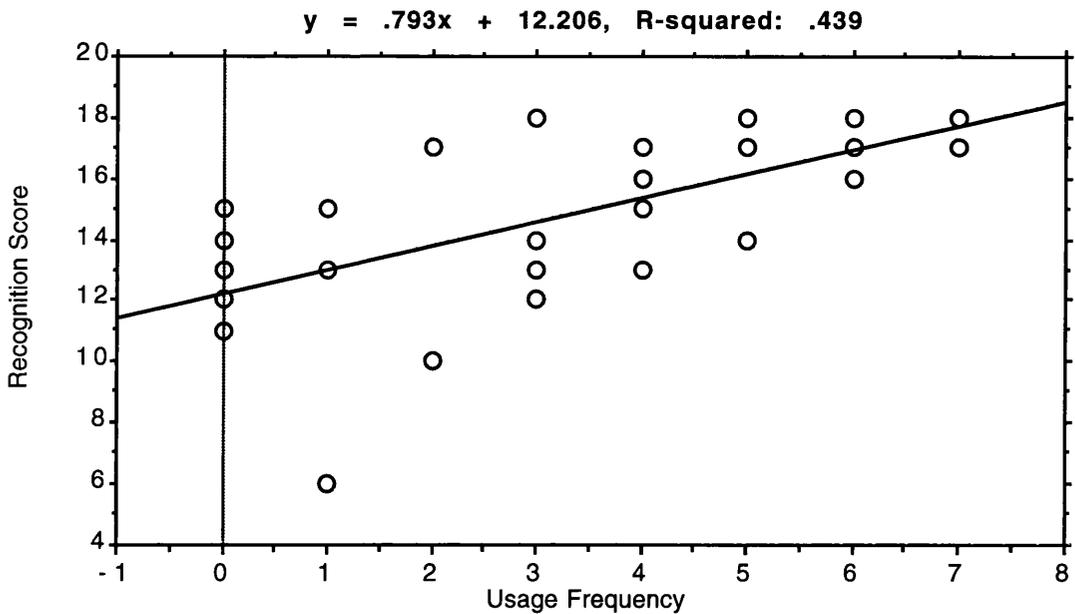
Table 4.8: (continued)

Underline	16 (89%)	4
24 Point	15 (83%)	1
Show Clipboard	15 (83%)	0
All Caps	15 (83%)	4
Sort	14 (78%)	5
Commands	14 (78%)	3
Show Ruler	14 (78%)	0
Open Footer	14 (78%)	0
Open	13 (72%)	1
Delete Forward	13 (72%)	0
Italic Cursor	13 (72%)	1
Plain Text	13 (72%)	4
New York	13 (72%)	3
Spelling	12 (67%)	0
Word Count	12 (67%)	3
New Window	12 (67%)	0
Quit	11 (61%)	0
Close	10 (56%)	2
Expand Subtext	10 (56%)	2
Change	6 (33%)	1

From table 4.8 it can be seen that all eighteen subjects correctly identified whether or not they had selected 7 commands (“Print”, “Glossary”, “Helvetica” “Bold”, “Full Repaginate Now”, “Help” and “Untitled 1”), whilst only 6 subjects could correctly identify whether they had selected the “Change” command

Using the data from table 4.8, a scattergram was plotted, (figure 4.2) and contrary to our prediction this appeared to show a positive linear relationship rather than our hypothesised non-linear, u-shaped curve.

Figure 4.2: Scattergram: recognition score by usage frequency



A second order polynomial regression was carried to test for non-linearity. However, since the F ratio actually decreased from the simple regression (25.8) to the second order polynomial regression (12.943), there was no need to calculate the incremental F as the difference was obviously not significant i.e. the relationship was linear.

In order to test whether this positive linear relationship was significant a Spearman's Rank Correlation Coefficient was calculated (appendix 4.11) for the same reasons as before. A correlation coefficient of 0.715 was obtained; this is significant at the 0.01 level (one-tailed). We can therefore conclude that recognition score increases as command usage frequency increases, i.e. subjects are more likely to recognise whether they had selected a command the more frequently that command was actually selected. To put this another way, never having used a command is quite likely to be confused with occasional use and vice versa. In order to examine the relationship between recognition score and language frequency, a Kendall's rank correlation coefficient was calculated (appendix 4.12). This calculation showed that with the effects of usage frequency partialled out, there is no significant correlation between recognition score and language frequency, i.e. they are unrelated.

4.4.7 “How often?” estimate and command usage frequency

Another interesting issue relating to subjects’ answers on the FC was to assess the accuracy of subjects’ estimate of how often they selected each command; this is really a case of validating the FC column heading “How often?” From a basic understanding of psycho-physics it is reasonable to expect that the discrepancy between subjects’ estimates of how often they selected a command and actual usage should increase as the frequency of actual command selection increases. What this is really saying is that, the more often a command is selected the harder it becomes to estimate exactly how often it actually was selected. The actual usage rate (number of times it was selected) for each command and subjects estimate of this is shown in table 4.9.

Table 4.9: Actual command usage, mean estimate of usage and difference between both for each command

Command	Usage frequency	Mean estimate	Mean difference
Open	1	1.69	0.69
Open Any File	6	2.47	3.53
Close	2	1.38	0.62
Print	7	3.41	3.59
Quit	0	0.47	0.47
Delete Forward	0	0.14	0.14
Glossary	7	3.56	3.44
Commands	3	1.25	1.75
Italic Cursor	1	1.06	0.06
Sort	5	2.82	2.18
Show Ruler	0	0.14	0.14
Plain Text	4	1.94	2.06
Bold	6	2.89	3.11
Italic	2	1.17	0.83
Underline	4	1.78	2.22
12 Point	2	1.00	1.00
24 Point	1	1.06	0.06
Helvetica	6	2.18	3.82
New York	3	1.44	1.56
Times	4	1.35	2.65

Table 4.9: (continued)

Open Footer	0	0.00	0.00
Full Repaginate Now	3	0.78	2.22
Demote Heading	5	2.53	2.47
Expand Subtext	2	1.00	1.00
Collapse Selection	6	3.59	2.41
Find	7	4.00	3.00
Change	1	1.00	0.00
Go To	5	2.50	2.50
Spelling	0	0.21	0.21
Word Count	3	1.67	1.33
Help	5	2.22	2.78
Show Clipboard	0	0.00	0.00
New Window	0	0.40	0.40
Untitled 1	7	3.94	3.06
All Caps	4	1.88	2.12

N.B. **Mean Estimate** = Subjects' mean estimate of the number of times that this command was used. **Usage Frequency** = Number of times this command was selected by each subject (range = 0 > 7). **Mean Difference** = Distance between Mean Estimate and Usage Frequency.

Using the scores shown in table 4.9, it is possible to plot a scattergram figure 4.3. This scattergram shows a clear, positive, linear relationship between command usage frequency and subjects' estimate of this.

Figure 4.3: Scattergram: "How often?" column answers by usage frequency

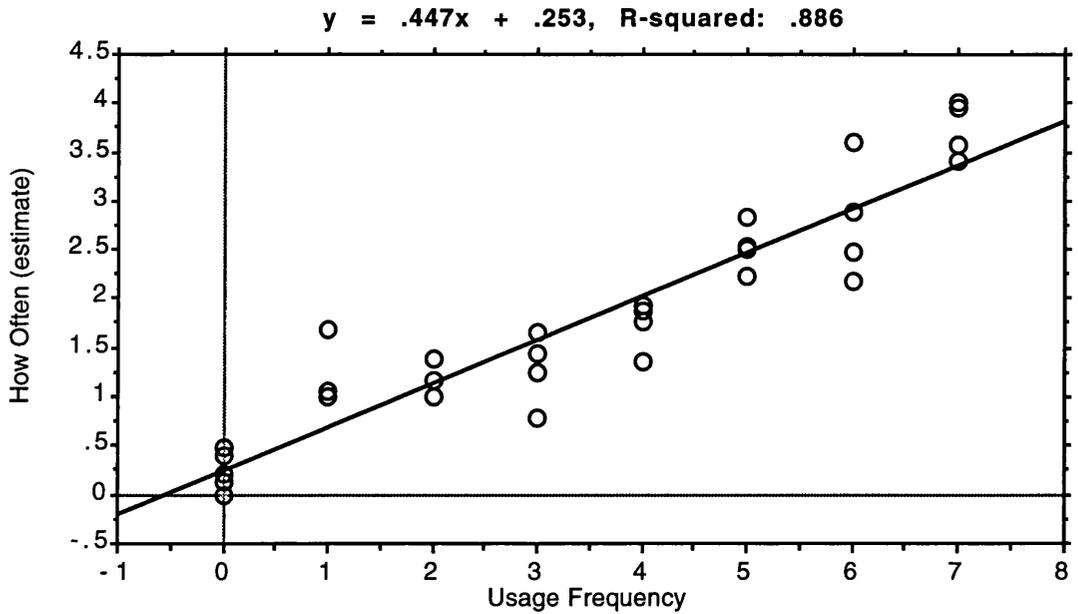


Table 4.10, summarises the results displayed in figure 4.3 by showing subjects mean estimate of usage for each usage category (0 to 7 selections). What becomes apparent from this table is that higher levels of command usage led to higher estimates of command usage, i.e. the accuracy of subjects' estimates is a function of frequency of usage.

Table 4.10: Subjects' estimate (mean) for each command usage category and difference between both

Actual usage (mean)	Subjects' estimate (mean)	Difference (mean)
0	0.19	0.19
1	1.20	0.20
2	1.14	0.86
3	1.29	1.71
4	1.74	2.26
5	2.52	2.48
6	2.78	3.22
7	3.73	3.27

Since it is proposed that in “real-life” studies, FCs should be used with experienced users over a set period of time, such as a week, the question arises as to how accurate and therefore useful the information from the column heading “How often?” will be.

In “real-life” studies it can be expected that some commands will be used much more frequently than has been the case in the present study, as a result, it may be more appropriate to ask subjects to estimate their command usage frequency by offering a choice of categories rather than asking for an actual numerical estimate. Of course this in itself raises its own problems, such as devising category descriptions that will convey the same notion to different subjects, i.e. what is the difference between often and frequently or seldom and rarely? However, it should be pointed out that FCs ask only about relative frequency as opposed to exact estimates.

4.5 Summary

To conclude, therefore, the results of this study suggests that FCs may be a useful and valid (i.e. accurate) instrument in HCI evaluation. This accuracy may be further improved by making design changes to the visual layout of the FC. The next study (study 5) is an attempt to a test this idea.

Chapter 5

Visual realism in feature checklist design: implications for validity (“Brickles”).

5.1 Introduction

The previous study suggested that FCs may be a useful and valid instrument for measuring subject usage of commands in computer applications such as word-processing packages. It was found that by using FCs, subjects could accurately identify whether they had used between 77% to 87% of all commands. In comparison, subjects using an open-response questionnaire (ORQ) could accurately identify whether they had used between 43% to 48% of all commands.

The FC used in study 1, simply listed by name all the commands contained in the word-processing application in the form of a list, together with menu titles preserving order; subjects then marked a tick or a cross against each command to indicate whether they had used it. In addition, subjects had to indicate approximately how often they used each command by writing the appropriate number beside each command.

The most obvious reason given as to why FCs were apparently so successful was that they require only recognition, while the ORQ required recall. However, a further reason cited was that when subjects were using FCs to identify which commands they actually used, they were in fact performing a similar task to that encountered in the “real-life” situation, i.e. selecting a command from a list of the same possible alternatives. In contrast the task involved in using ORQs to identify command usage is clearly not analogous to the “real-life” task. The point is that the FC and the “real-life”

situation both have the same cues, i.e. menu commands and titles, sequence, groupings, etc., and both involve similar tasks, namely:

- Recognition of command names.
- Task→action mappings, (i.e. giving command name or task description and remembering usage).

This view stems from previous findings which suggested that “the way in which information is encoded determines what retrieval cues are effective in providing access to what is stored” (Tulving and Thompson 1973).

This study leads on from these previous findings in that it takes the view that “more accurate remembering may result from retrieval processes that match or simulate the “real-life” encoding process more closely”. By making changes to the visual layout of the FC, it may be possible to increase the similarity between the recognition task on the FC and the computer task that subjects performed. As a consequence of this, it was hoped that FCs with increased visual realism (i.e. closer visual match to the actual computer screen) would produce more accurate results than the standard FC used in study 1.

The idea behind this study is consistent with the theory of cue-dependent forgetting (Tulving 1974); namely, that subjects are more accurate at recalling features they have seen, used, etc. when the features present at encoding are also present at recall. It could be the case that some of the features (cues) present at encoding relate to the visual layout of the computer screen. This is a possibility we set out to test.

Another aspect that this study examined was the different ways of listing the commands to be selected, and the possible effects this might have on subject’s memory of command usage. It may be the case that giving subjects descriptions of commands’ functions, may be a closer match to the way in which subjects use and encode commands in “real-life” situations. What this is really saying is that when subjects want to perform a task on the computer they usually think of the task that they want to do and then select the command that performs this action; i.e. task→action mappings. If this is the case, then subjects that received these functional descriptions should be more accurate in remembering which commands they used, than the subjects that received the command names. Again, this is an assumption that we set out to test.

The hypotheses of this study were twofold:

- (1) The FC that is a closer visual match to the actual computer screen (visual FC) will produce significantly more accurate reports of command usage than the standard FC.
- (2) Subjects that received descriptions of commands' functions would produce more accurate reports of command usage than those that received command names.

5.2 Method

Subjects: The study contained twenty-four subjects, split into four groups of six. Of these twenty-four subjects, eleven were female and thirteen were male; their ages ranged from 17 to 44 years of age with a mean of 20.8 years. Subjects were recruited from the psychology summer school at the University of Glasgow and were informed that they had to participate in a study as part of their summer school course; subjects were told that they would be paid £2.00 for doing so. All of the subjects had little or no experience of using computers, this was assessed by asking subjects to complete a computer experience questionnaire that was administered in the recruiting phase of the study, (appendix 4.1).

Apparatus/Stimuli: The study was run on an "Apple Macintosh IIsi" and displayed on an "Apple Macintosh A4 Mono Monitor". The computer game "Brickles-7.0" was used in this study and all the computer operations that subjects performed were recorded using the "Tempo II Plus" application.

Design/Procedure: The purpose of the game was explained to all subjects and they were then shown how to select commands and play the game using the mouse. In this game the menu commands which were to be the features of the FC are not normally used during a game, but are instead used to arrange different game settings before commencing play. Subjects were allowed to practice playing the game for a couple of minutes until they felt comfortable doing so.

Subjects were then issued with a task sheet which contained a list of commands that subjects had to select in order to play different versions of the game. Subjects in groups A and B received a command task sheet (CTS), i.e. one that listed the actual command names for them to select (appendix 5.1), and subjects in groups C and D received a descriptive task sheet (DTS), i.e. one that listed descriptions of commands'

functions without actually using that name (appendix 5.2). After completing each version of the game, subjects had to record their score on the task sheet.

When subjects had completed all the versions of the game on the task sheet, they were then issued with a questionnaire that obtained information for a later study on their use of public transport (appendix 5.3). After completing this, subjects in groups A and C were issued with a standard FC (appendix 5.4) followed by a visual FC (appendix 5.5); subjects in groups B and D received these FCs the other way around. Table 5.1 explains these experimental conditions:

Table 5.1 - Task sheet type and FC order in experimental groups

FC Order	Task Sheet	
	Command	Descriptive
Standard FC / Visual FC	A	C
Visual FC / Standard FC	B	D

5.3 Results

Since study 1 has shown that FCs are most useful when issued alone (or as the first of a series of instruments), this study only looked at the results of the FCs that were issued first.

From table 5.2, it can be seen that when the standard FC was issued first, subjects correctly recalled whether they had used 31.2 commands (82%) with the CTS and 32.0 (84.2%) commands with the DTS. When the visual FC was issued first, subjects correctly recalled whether they had used 30.5 commands (80.3%) with the CTS and 32.2 commands (84.7%) with the DTS.

Table 5.2: Subjects' mean recall of commands used with the standard FC and the visual FC

Experimental Condition	Correct Usage	Incorrect Usage	Don't Know
Standard FC 1st (CTS)	31.2 (82.0%)	6.8 (18.0%)	0.0 (0.0%)
Standard FC 1st (DTS)	32.0 (84.2%)	4.5 (11.8%)	1.5 (4.0%)
Visual FC 1st (CTS)	30.5 (80.3%)	4.2 (11.0%)	3.3 (8.7%)
Visual FC 1st (DTS)	32.2 (84.7%)	4.8 (12.7%)	1.0 (2.6%)

N.B. (CTS) = command task sheet, and (DTS) = descriptive task sheet).

A two-factor analysis of variance for independent groups was used to estimate the effect of FC type and task sheet type on correct usage (i.e. the number of commands that subjects correctly remembered using). Table 5.3, shows the ANOVA table where Factor A = FC type (standard or visual) and Factor B = task sheet type (command or descriptive).

Table 5.3: ANOVA table for effects of FC type (factor A) and task sheet type (factor B) on correct usage

Source:	Sum of Squares:	df:	Variance:	F - test:	P value :
Factor A	0.375	1	0.375	0.044	0.8363
Factor B	9.375	1	9.375	1.095	0.3078
Interaction AB	1.042	1	1.042	0.122	0.7308
Within Groups	171.167	20	8.558		

The obtained F value for Factor A, 0.8363 does not exceed the F of 4.3512 for 1 and 20 degrees of freedom at the 0.05 level. We can therefore conclude that the visual FC does not produce significantly more accurate results of command usage than the standard FC

Similarly, the obtained F value for Factor B, 0.3078 does not exceed the F of 4.3512 for 1 and 20 degrees of freedom at the 0.05 level. We can therefore conclude that the subjects that received the descriptive task sheet did not have a significantly better memory for command usage than the subjects that received the command task sheet.

Finally, the obtained F value for interaction (A x B), 0.7308 does not exceed the F of 4.3512 for 1 and 20 degrees of freedom at the 0.05 level. We can therefore conclude

that the combined effects of FC type and task sheet type do not significantly affect the accuracy of command usage scores.

5.4 Discussion

The results reported in this study do not support the first experimental hypothesis that, “the FC that is a closer visual match to the actual computer screen (visual FC) will produce significantly more accurate results of command usage than the standard FC”. Furthermore, it was also found that the results did not support the second experimental hypothesis that, “subjects that received descriptions of each commands’ function had a better memory for command usage than those that received command names”.

However, before moving on to discuss possible reasons for the above results, it is worth pointing out that all the FCs scored highly on correct usage of commands, i.e. a minimum score of 80% and a maximum of 84%. This offers further support for previous findings that FCs are a highly valid and useful instrument for measuring command usage in HCI.

Rather than take the view that closer visual realism in FC design is unimportant, it is possible that the visual FC used in this study was perhaps not as good (close) a match to the computer screen as it could have been. Although the visual FC was designed by using actual screen dumps of the “Brickles 7.0” menus, these menus were displayed on the FC vertically and on separate pages (two menus on each page). In addition other visual features of the screen were missing from the visual FC, e.g. spatial layout, the game itself, etc. It may be the case that FCs which included these aspects in their design could produce more accurate results than the standard FC. It is therefore proposed that future research could look into this possibility, perhaps by letting subjects look at the actual computer screen when filling in the FC (this is likely to be the closest possible attempt at visual realism).

There are two other possibilities however, for the results reported. Firstly, in view of the uniformly high accuracy scores achieved, it could be the case that the “ceiling” of FC performance has already been reached, i.e. around 80 - 85% accuracy may be the maximum they could achieve regardless of any design changes.

Secondly, it could also have been the case that visual realism is an inappropriate way to improve FC validity. Perhaps listing commands on FCs by descriptions of their

functions rather than their name could increase accuracy. Although these functional descriptions were used on the task sheets in the above study, they were not used as cues on the FCs. The reasoning behind this view is that if these descriptions are analogous to the encoding process that subjects go through, then they should also be used at the retrieval process (i.e. FCs). Thus future research could also look at the accuracy of FCs that list commands by descriptions of their function; a problem here however, will be whether the description matches the internal representations that subjects have.

5.4.1 “How often?” estimate and command usage frequency

From study 1, it was found that the discrepancy between subjects’ estimate of how often they selected a command and the actual usage (i.e. relative frequency), increased as the frequency of command selection increased; to put this another way, “the more often a command is selected the harder it becomes to estimate accurately how often it actually was selected”. However, in spite of this it was found that subjects’ answers to the column headed “How often?”, were generally a good and accurate indicator of the frequency of usage of that command, i.e. as command usage increased so did subjects’ estimates of command usage. We can now see if this is still the case for study 2.

The actual usage of each command and subjects’ estimate of this (across all subjects) for both the standard FC and the visual FC is shown in appendices 5.6 and 5.7 respectively. The correlation between the actual usage scores and subjects’ estimates using the standard FC was 0.849, whilst the correlation between the actual usage scores and subjects’ estimates using the visual FC was 0.669. Since both of these correlation scores are positive and high we can again say that, generally speaking subjects’ estimates of how often commands were used, increases as the actual frequency of usage increases. It is interesting to note that subjects using the standard FC were more accurate in their estimates of command usage frequency than subjects using the visual FC. After conducting a t-test between the two FCs on the discrepancy between subjects’ estimates and actual usage frequency (estimate error), a t-value of -2.477 was obtained; this indicates that the difference in means between the two FCs for estimate error, were not significant.

5.4.2 *FC accuracy: recognition vs. recall*

In section 4.4.1, a distinction was made between recall and recognition, i.e. “the essence of a recall test is that the subject has to generate the target/s meeting the definition of the target in the recall instruction”; whereas “the essence of a recognition test is that one or more potential targets are presented to the subject, there is no requirement for overt generation of the target and the response may consist in accepting/rejecting a given choice, rating it, etc.” (Brown 1976).

However, using this definition it becomes clear that classification of a test as either recall or recognition is not a simple matter, e.g. if part of a task is presented and the task is to identify it, then there is a tendency to classify this as a recognition task; however, using the above definition it is a recall task unless one or more possibilities are provided for the missing part.

In order to use Brown’s definition for understanding the processes at work in the present study, it is helpful to summarise the crucial distinction he makes, that is “the generation of the target word”. In recall, the subjects must generate all or part of the target word and then decide whether or not it occurred. In recognition, subjects are presented with the whole target word either alone or with distractors (either false or genuine) and have to retrieve one piece of information about whether it occurred, i.e. “yes” or “no”. FCs give subjects the whole target word (command name) and ask people to retrieve, not whether the command existed (i.e. is part of the interface) but whether or not they used it; this is a crucial distinction between the study described here and traditional studies on recall and recognition. Thus subjects are not generating the target but they are having to remember (recall?) whether they used that command. What is apparent however, is that recognition may sometimes be mediated by processes characteristic of recall i.e. processes not dependent on the presence of potential choices in the recognition task.

Brown (1976) lists a number of different aspects of recognition tasks that are likely to encourage the use of recall; of these, the need for inference is of particular importance for FCs. The need for inference (Bartlett 1932), emphasises the reconstructive nature of memory, e.g. the question “did you use this command?”, appears to be a recognition task requiring a yes/no answer. However, the user may only be able to answer this question by trying to recall what the command does, what commands he/she didn’t use, etc.

When discussing whether recognition or recall may be more relevant for explaining FCs, it is important to note that there is a major difference between studies 1 and 2. In study 1, the subject's task is to recognise the appropriate command from the list on the menus and then select it via the keyboard; the subject received no feedback about the effect of their action and simply continued to select commands in this rote manner. In this study it is likely that the subject learned discriminators to aid recognition.

In study 2, the subject's task was again to recognise and select the appropriate command from the menus, however, in this study the subject could actually see the effect/s of their actions and were therefore more likely to learn task→action mappings, i.e. going from an internal memory description of the effect you want to the name of the command on the menu. This is especially true for those subjects that received descriptions of commands to choose rather than command names. Given that this is the case, it seems likely that when subjects were completing the FC in study 2, recognition was merely the first stage of a reconstructive process, i.e. recognition presupposes recall. Having said this however, it should be noted that the task for subjects in this study was still to an extent artificial, i.e. they were still following instructions as to what tasks to perform rather than initiating these tasks themselves.

If FCs do indeed require subjects to perform a reconstructive memory process based on such things as internal task→ action mappings, it may be that including semantic descriptions of command functions as memory prompts, would be an appropriate way of increasing FC accuracy. Study 5 (Chapter 8), looks at the possibility of including these descriptions on FCs to act as a cue for asking questions about.

5.5 Summary

To conclude, therefore, the results of this study do not support the hypothesis that closer visual realism in FCs design will increase the accuracy of subjects' memory for command usage beyond that of standard FCs. However, rather than abandon this issue, the view taken is that more research needs to be carried out in order to explore the issue more fully. The next study will therefore be a further attempt at exploring the issue of visual realism in FC design.

Chapter 6

Visual realism in feature checklist design: implications for validity (“MacPaint”)

6.1 Introduction

Earlier studies conducted on FCs (studies 1 and 2), showed that FCs may be an accurate and useful instrument for measuring subject usage of commands in different types of computer software packages, e.g. word-processing applications and computer games.

In chapter 5 (study 2), an attempt was made to try and increase the accuracy of FCs even further, by improving their visual realism, i.e. making FCs more visually similar to the actual computer screen. The idea behind this study was consistent with the theory of “cue-dependent forgetting” proposed by Tulving i.e. “subjects are more accurate at recalling features they have seen, used, etc. when the features present at encoding are also present at recall”, (Tulving 1974). It was hypothesised that some of the features (cues) present at encoding (i.e. selecting command names), pertain to the visual layout of the screen. As a result, more visually realistic FCs would make it easier for subjects to remember which commands they had used.

In spite of this however, the results of Chapter 5 (study 2) did not support the experimental hypothesis. This was in part explained by the fact that the visual FC used in this study still did not contain many of the visual details that were present on the computer screen. This study is therefore a further attempt at improving FC accuracy through improved visual realism.

In this study it was the intention to use an application that would allow subjects to explore and use features in a more natural way than the forced selection procedure used in the “Brickles 7.0” study (chapter 5). It was also the intention to use an application that contained a wide variety of visual features as well as textual ones (the logic behind this is that visual realism might be a more important consideration for FC design when the actual features in question are pictorial rather than textual e.g. drawing tools, line icons, etc.). As a result of these considerations, the graphics package “MacPaint” was used in the study. It is hoped that FCs will be as accurate in obtaining information about pictorial features of interfaces as they have been in obtaining information about textual features. If this is the case, then FCs may be a useful addition to research on the design of icons, etc. in HCI (e.g. Chessari and Lindegaard 1988; Blackenberger and Hahn 1991).

Earlier research (Draper and Barton 1993), has shown that new users of “MacPaint” learn quite quickly and effectively by exploration; these users also used many different features without prompting by the experimenter, although there was a tendency to use the pictorial features to a much greater extent than the textual ones on the pull down menus.

As was mentioned earlier, an explanation offered for the results of Chapter 5 (study 2), was that the visual FC was still not a very close match to the visual layout of the actual computer screen. It was suggested that a possible improvement would be to let subjects look at the application on the computer screen (i.e. open menus etc.), whilst completing the standard FC; this therefore became one of the conditions in the following study along with the standard FC on its own. A third condition was to use a visual FC that very closely matched the visual layout of the screen (particular attention was paid to spatial characteristics of the screen). It was hoped that the visual FC that paid close attention to such things as horizontal vs. vertical arrangements, neighbouring icons and tools, etc. would be more accurate than the standard FC on its own and possibly equal to the standard FC with the computer screen. Since it was expected that the textual features of the interface in this study would not be used very frequently, differences between the three groups in correct usage scores were only anticipated for pictorial features.

The experimental hypotheses in this study were twofold:

- (1) The standard FC + screen and the visual FC alone, will have significantly higher correct usage scores for visual features than the standard FC alone.

- (2) There will be no significant differences between the three groups in correct usage scores for textual features.

6.2 Method

Subjects: The study contained eighteen subjects in total, split into three groups of six. Of these eighteen subjects, 14 were female and 4 were male; the ages ranged from 19 to 51 years of age with a mean of 21.9 years. Subjects were recruited from the campus at the University of Glasgow and were asked if they were willing to participate in a psychology study for a 3rd year student project. All of the subjects had little or no experience of using computers, this was assessed by asking subjects to complete a computer experience questionnaire that was administered in the recruiting phase of the study, (appendix 4.1).

Apparatus/Stimuli: The study was run on an “Apple Macintosh IIsi” and displayed on an “Apple Macintosh A4 Mono Monitor”. The graphics application package “MacPaint” was used and all the computer operations that subjects performed were recorded using the “Tempo II Plus” application. Each subject’s actions on the computer screen were also recorded by video camera.

Design/Procedure: The purpose of the “MacPaint” package was explained to all subjects and they were then asked if they were familiar with using a mouse; subjects who were not, were shown how to use one using the “MacDraw” application, i.e. how to select commands, use tools, etc. Once subjects felt comfortable using the mouse, the “MacPaint” application was opened and subjects were told to “play” with the application for about half an hour whilst trying to use as many features as possible.

After this half hour session was over, subjects were given either a standard FC (appendix 6.1) or a visual FC (appendix 6.2) to complete. Half of the subjects that received the standard FC were allowed to look at the “MacPaint” package on the computer screen and open menus etc. in order to help them fill in the FC. The three experimental conditions are summarised in table 6.1.

Table 6.1: FC type in the three experimental conditions

Group	FC Type
A	standard FC
B	standard FC + computer screen
C	visual FC

6.3 Results

For the purposes of this study, the features of the “MacPaint” interface were categorised as either textual (i.e. pull-down menu command names), or pictorial (i.e. drawing tools, line icons and pattern icons). For each group there was a total of 63 textual features and 63 pictorial features.

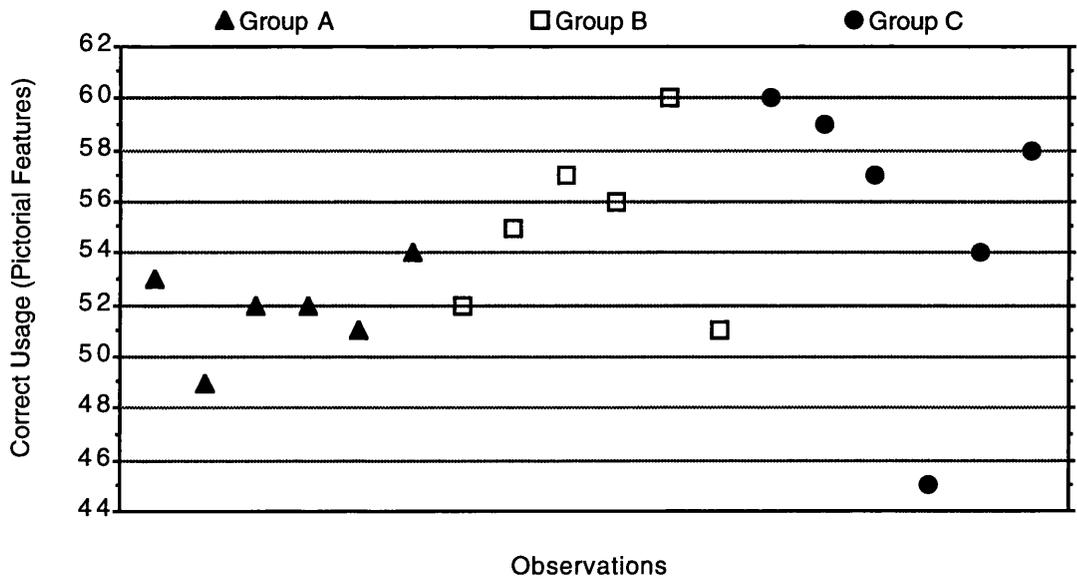
From table 6.2, it can be seen that using the standard FC alone (group A), subjects correctly recalled whether they had used an average of 59.7 textual features (94.7%) and 51.8 pictorial features (82.2%). Using the standard FC along with the aid of the computer screen (group B), subjects correctly recalled whether they had used on average 60.3 textual features (95.7%) and 55.2 pictorial features (87.6%). Finally, using the visual FC (group C), subjects correctly recalled whether they had used on average 61.2 textual features (97.1%) and 55.4 pictorial features (87.9%).

Table 6.2: Subjects’ scores for correct usage of “MacPaint” features in the three experimental conditions

Feature Type	Standard FC	Standard FC + screen	Visual FC
Textual Features	59.7 (94.7%)	60.3 (95.7%)	61.2 (97.1%)
Pictorial Features	51.8 (82.2%)	55.2 (87.6%)	55.4 (87.9%)

It can be seen that there was a greater variation in scores between the groups for pictorial features, i.e. groups B and C scored better than group A (a difference of 5.4% and 5.7% respectively); any significant differences between group scores are likely to come from these means. However, before conducting an analysis of this data it is necessary to look at the distribution of subjects’ scores; this is done in figure 6.1, below:

Figure 6.1: Scattergram of subjects' correct usage scores for pictorial features in "MacPaint"



From this scatterplot it can be seen that subject number 16 (i.e. Group C) scored much lower on correct usage scores for pictorial features than any other subject, especially amongst the other subjects in their group (subject 16 is 9.2 points away from this mean of 54.2). Since the data does not correspond to a normal distribution we cannot conduct a parametric statistical analysis and must instead use a non-parametric test. The most appropriate non-parametric test in this case is an “extension of the median test”.

After conducting an extension of the median test (appendix 6.3) we achieved an $X^2 = 7.333$. When we compare this against a table of critical values of Chi Square we find that an X^2 equal to or greater than 7.333 for $df = 2$ has a probability of occurrence between 0.05 and 0.02. Since this p is smaller than our previously set level of significance ($p = 0.05$), we can accept our main experimental hypothesis that: “the standard FC + screen and the visual FC alone, will have significantly higher correct usage scores for visual features than the standard FC alone”.

6.4 Discussion

The results obtained in this study show that visually realistic FCs are significantly more accurate than the traditional standard FC. We can now discuss these results.

As far as textual features are concerned there was very little difference between the three groups. This was not too surprising since it was hypothesised that closer visual realism might be more important for pictorial features in “MacPaint” than textual ones.

However, a further point that helps to explain these results is that subjects in all three experimental conditions, used very few textual features at all (i.e. menu commands). This low usage of textual features resulted in very high validity scores (98.1%) across all three experimental conditions and reduced the chances of finding any significant differences. What this is really saying is that subjects in this experiment found it very easy to remember whether they had used the textual commands, simply because they never (or very rarely) used them.

It was seen from table 6.2 that the correct usage scores for the visual FC were slightly higher for pictorial features than those for the standard FC + the computer screen. One reason that would help to explain these findings is that subjects in group B (standard FC + computer screen) did not look at the computer screen that often when filling in the FC even although they were told that they were allowed to do so. As a result they may not have received as much of a visual benefit (i.e. memory prompt) as was anticipated. It would have been interesting to discover just how much of a nuisance subjects felt looking at the computer screen whilst filling in the FC actually was.

6.4.1 FCs and the “encoding-specificity hypothesis”

Earlier discussions based on the literature on recall vs. recognition have shown that it is inappropriate to talk of FCs as pure recognition. Although research on human memory has demonstrated cases where one may be easier than the other, it should be emphasised that these studies were not in fact directly relevant since the experimental tasks are not comparable to the ones described in this thesis. Nevertheless, the basic argument that FCs probably involve both recall and recognition to varying degrees is still valid, although not proven. What is proven however, is that:

- FCs ask little from the respondent and yet yield a lot of accurate information for the researcher.
- FCs give a big part of the cue actually present during learning, i.e. normal use.

In addition to this it also seems likely that the task required of subjects when completing FCs is very similar to that encountered in the “real-life” situation, i.e. selecting a

command (whether textual or pictorial) from a list of the same possible alternatives. This view relates well to the “encoding-specificity hypothesis” (Tulving and Thompson 1973), which states that “what is stored is determined by what is perceived and how it is encoded, and what is stored determines what retrieval cues are effective in providing access to what is stored”. Two aspects to this are that firstly, you may need the same modality, and secondly, you may need the whole cue. With respect to both of these, FCs give a good approximation to the perceptual cues that users were exposed to, i.e. they give a lot, and probably cover a lot of the cues actually used. Thus even if the cues used to support behaviour shift as learning progresses (e.g. with experienced users), there is still a fair chance of them being presented in FCs.

This issue of giving a large part of the cue that is present during normal use is particularly relevant in the present study, i.e. subjects were asked to use the “MacPaint” package in a normal, explorative manner, characteristic of the learning stage. Given that the subjects in this study only used the package for a short period of time (about half an hour), it is likely that the cues used to support user behaviour were primarily visual ones. Later on with more experienced use it is likely that these cues may become internalised and be more related to less obvious features of the interface; this view receives support from Kaptelinin (1993) who states that “the acquisition of menu selection skills is associated with a transformation of the internal representation used. While novices rely on word recognition, expert users extract global visual features from the display”.

6.4.2 *An evaluation of the “encoding-specificity hypothesis” and “generation-recognition” models of human memory*

A crucial distinction between the “encoding-specificity hypothesis” (E-SH) and G-R models of human memory is that E-SH claims that recall and recognition are different forms of a single retrieval process, and that only information in the retrieval environment can facilitate trace retrieval, (i.e. a simple, direct comparison between information available at retrieval and information stored in memory). G-R theories on the other hand, specify that recall and recognition are separate stages in retrieval and that the retrieval environment may influence retrieval directly, or indirectly by causing the generative mediation stage to produce information that overlaps with the trace (although not specifically specified by the retrieval environment).

A number of studies have shown that recall is more complex than that proposed by E-SH and that recall does involve a generative stage (Parkin 1981; Jones 1982). It was shown that instructions to subjects at the encoding stage could allow access at retrieval to an indirect route (Jones 1982). This led Jones to claim that there are two different ways in which recall can occur:

- The direct route, in which the cue permits direct accessing of the to-be-remembered information.
- The indirect route, in which the cue leads to recall via the making of inferences and the generation of possible responses.

It should be noted that this relates to the need for inference (discussed in section 5.4.2) as an aspect of recognition tasks that is likely to encourage the use of recall (Brown 1976).

Despite the criticisms of both G-R models and E-SH, it should be emphasised that both involve a reconstructive approach to retrieval. Recent research on recall and recognition, such as that of Jones (1982), suggests that there are several different strategies which can be used in order to produce recall and recognition.

6.4.3 *Current views on recall and recognition*

Of the current views put forward to help understand recall and recognition, an attempt made by Mandler (1980) to identify strategies used in recognition memory, may be particularly relevant to studies 2 and 3. In his theory, Mandler argues that recognition is not a single process but rather it involves two components/stages. In the first of these, recognition is based on familiarity of the stimulus and in the second recognition is based on identification following a context retrieval process. An important assumption of Mandler's theory of recognition memory is that familiarity decays more rapidly over time than identification based on context retrieval; this is due to the fact that identification involves more detailed organisation and structuring of information at learning. A number of studies have found results supporting this view (Mandler 1967; Mandler et. al. 1969; Mandler and Boeck 1974). It is possible that this rapid decay of familiarity over time could explain the depression of FC accuracy when it is presented after the ORQ (section 4.4.2).

It is possible to use Mandler's work to discuss the results found in the present study. However, one problem with this is that in all the research referred to, the task that subjects had to perform was to either recall or recognise the existence of a word from a previously presented list. In the present study however, the task is not simply identifying whether a feature (command name or icon) existed, but also to remember whether or not the feature was used. It is therefore likely the additional information required by FCs would involve subjects conducting an explicit reconstructive search process more similar to identification based on context retrieval rather than implicit context free familiarity.

Given the results that improved visual design of FCs significantly improve the accuracy of subjects' responses, it seems likely that subjects organised the information in the "MacPaint" interface by means of perceptual/visual cues to a greater or lesser extent. This context information was to a large extent supplied by the visual FC and so facilitated identification following context retrieval.

It should be noted that the preceding discussion served only to demonstrate how Mandler's work could be used to understand the results of this study. At present there is still much debate about Mandler's theory, especially concerning the roles of the two stages and whether they operate at the same time; as Eysenck et. al (1990) state, "there is considerable uncertainty about the relative roles played by the two processes in most situations".

From the discussion of recall and recognition and retrieval based models of memory, it can be seen that despite the differences it is clear that context plays an important role in human memory. The research discussed has shown that both recall and recognition are complex processes. In addition, recent research relying on behavioural measures of memory not dependent on conscious awareness of past events (Gardiner 1988), has led to a move away from traditional recall-recognition studies; some researchers have made the claim that, "the implication is that tests of recall and recognition are less revealing about human memory than cognitive psychologists used to think" (Eysenck et. al. 1990).

6.4.4 *The “transfer-appropriate processing” account of memory*

So far the different memory theories discussed have concentrated on different memory tasks and have tried to interpret the findings as evidence for different underlying memory systems. However, these “system” approaches have been criticised because “the various systems postulated do not meet the criteria for separable memory systems” (Sherry and Schacter 1987). As a result, the “transfer-appropriate processing” account of memory has been proposed.

This account focuses less on the memory tasks themselves than on the match in mental operations performed at study and test. The basic premise is that memory for a prior occurrence results from the overlap between the retrieval processes induced by a memory test and the encoding operations undertaken during learning; as a result, “memory performance will be improved to the degree that the types of operations performed at study overlap with those required at test” (Morris et al. 1977). It should be seen that this theory has much in common with the “encoding-specificity hypothesis” (Tulving and Thompson 1973). Within this framework, memory tests are either conceptually driven, in that performance depends on some recapitulation of the elaborative process that took place during encoding, or data-driven, in that performance depends on the overlap between perceptual processing at encoding and retrieval.

Support for the “transfer-appropriate processing” view has come from a number of studies which have investigated dissociations among memory measures (e.g. Jacoby 1983; Blaxton 1989). In one series of studies, performance on multiple episodic tasks was compared with semantic tasks using the same study manipulations and target items (Blaxton 1989). From this it was observed that dissociations existed between two tasks tapping the same system. In explaining this, the “transfer-appropriate processing” view claimed that, “performance on data driven tests relying on the analysis of physical features was improved when those features were similar at study and test. On the other hand, conceptually driven tasks benefited most from elaborative processing of meaning during study” (Blaxton 1989).

It should be seen that the “transfer-appropriate processing” framework is a useful account for discussing the present study. Using this framework, the important question for FCs becomes, “Do the conditions/features under which subjects learned to use “MacPaint”, overlap with the conditions/features present on the FCs?” In line with other research that suggests that novices rely more on word recognition in menu selection (Kaptelinin 1993), it seems highly likely that the important features concerned

physical aspects of the command names, icons, etc. It was these same conditions/features that the visual FC attempted to replicate, i.e. there was considerable overlap between the conditions at encoding and the conditions at retrieval.

Yet again it is important to point out that unlike the research referred to, the present study was not designed to investigate system or processing theories of memory but rather to validate the usefulness of FCs. The discussion of memory research serves only to try and explain/understand the success of FCs in the present study.

6.5 Summary

To conclude, the results of this study support the view that closer visual realism in FC design results in a significant increase in FC accuracy for pictorial features. These results indicate that FCs should be designed with closer visual realism. This is especially true when FCs are intended to be used with applications that contain many pictorial features.

Chapter 7

Using feature checklists for discovering user knowledge of the Glasgow underground railway system

7.1 Introduction

Previous studies concerned with assessing the accuracy of FCs, have demonstrated that FCs are an accurate and valid way of measuring what features (commands) of an interface subjects actually use, and how often these are used. Having established this, it would also be of benefit to the evaluation process to find out additional information about users knowledge of features (commands), i.e. such things as:

- “Do users know what the command actually does?”
- “Do users think they will ever have a need to use that command?”
- “Do users know if the command actually exists?”

By adding extra question columns that require only a tick or cross, it may be possible to obtain this extra information from FCs, whilst at the same time minimising the cost to the experimenter and the user.

This study is an attempt to see:

- Whether the idea of obtaining extra information through FCs is feasible.
- How accurate subjects’ answers are to these new question columns.

In order for subjects to answer the questions outlined above, it is important that the subjects are experienced in using the particular domain under investigation. In addition it is also important that any assessment of these answers can be accurately validated against some true measure, i.e. there must be a check on users knowledge. However, the difficulty for this study is in finding a system that contains a small number of features and many users with comparable experience. Since the only subjects available for this study were naive computer users it was also necessary that comparisons could be made between domain under investigation and computer applications, i.e. the same question columns could be used.

As a result, it was decided to use the Glasgow Underground Railway System (GURS) as the domain of investigation in this study, since this satisfied both requirements, and based on an earlier survey (appendix 5.3) that was administered in study 2, it was found that it was possible to obtain a large number of experienced users.

Questions such as, “did you know that this command existed?”, “do you know what this command can be used for?”, and “do you think you would ever have a need to use this command?”, would still be appropriate if the feature under investigation was the underground station as opposed to the computer command. In this case the additional information about the features would come from the following questions:

- Did you know this station (command) existed?
- Do you know what this station (command) can be used for?
- Do you think you would ever have a need to use this station (command)?

Furthermore, people in general (and the subjects in this study) typically use only a personal subset of train stations, just as users typically use only a personal subset of commands in a program.

An additional aspect that this study examined was whether differences in the organisation of the FC would result in differences in the accuracy of subjects’ answers. The view taken here, is that when people use the GURS they are likely to encode the train stations on the basis of geography and the order in which they appear. If this is the case, then a FC that lists stations by their geography and order of appearance will match the encoding process more closely and will subsequently improve the accuracy of subjects’ memory. In order to test this we compared a geographical or “route” FC with a FC that simply listed the train stations alphabetically

The following study is an attempt to discover whether it was feasible to use FCs to ask these questions, and if so whether these new questions are valid. There were two experimental hypotheses in this study, these were:

- (1) Subjects' answers to these new column headings will be accurate.
- (2) Subjects that received the route FC will be more accurate in their answers than subjects that received the alphabetical FC.

7.2 Method

Subjects: The study contained twenty subjects in total, split into two groups of ten. Of these twenty subjects, 12 were female and 8 were male; the ages ranged from 17 to 44 years of age with a mean of 23.8 years. Subjects were recruited from the psychology summer school at the University of Glasgow and were asked to participate in a study as part of their summer school course; subjects were told that they would be paid £4.00 for doing so. All of the subjects were experienced users of the GURS, i.e. they used it either frequently or occasionally (but regularly and had done so for a long time); this was assessed through a question on the FC.

Apparatus/Stimuli: The subjects were issued with either a route FC which listed the stations on the GURS by geographical sequence (appendix 7.1), or an alphabetical FC which listed the stations on the GURS alphabetically (appendix 7.2). Both FCs contained fifteen stations plus four false stations (distractor features) that varied in degree of plausibility. The FCs therefore contained 19 stations in total.

In the second part of the study all subjects were interviewed using the same semi-structured interview (appendix 7.3).

Design/Procedure: Subjects in the psychology summer school were issued with either a route FC or an alphabetical FC on the GURS at the beginning of one of their computer laboratories. Subjects were instructed that all FCs had to be completed and returned by the end of the laboratory. At the bottom of each FC, subjects were asked to indicate how often they used the GURS and if they would be willing to be interviewed about certain aspects of the system. A total of twenty students were selected to be interviewed; of these ten had completed a route FC and ten had completed an alphabetical FC.

All twenty subjects were interviewed the following day by the researcher using the semi-structured interview. This asked them about their knowledge of the GURS in more detail. The actual questions on the interview were as follows:

- How sure are you that this underground station exists? Please indicate this using the following scale (appendix 7.4).
- Do you know what this underground station might be used for generally?
- Do you think you would ever have a need to use this underground station? Please estimate this need using the following scale (appendix 7.5).

The semi-structured interview contained various prompts to aid subjects in their answers. The purpose of this interview was to validate subjects' answers on the FC to the columns headed "What for?" and "Need?". Subjects' answers to the distractor features were used to validate their answers on the FC to the column headed "Existed?".

7.3 Results

The following results are split into three sections each of which deals with one of the three new question columns.

7.3.1 *"Did you know this underground station existed?"*

The first new column on the FC that we wished to validate was the column headed "Existed?" i.e. "did you know this underground station existed?". As we mentioned earlier the method of carrying out this validation was to compare subjects' answers with the actual answers (remember both FCs contained false stations). It can be seen from table 7.1, that overall, subjects gave 332 correct answers (90.9%) and 33 incorrect answers (9.1%) out of a total of 365. N.B. Don't Know answers were excluded from the analysis here.

Table 7.1: Agreement between subjects' answers on the FCs and the actual answers to the FC question "did you know this station existed?"

		Actual Answer	
		Yes	No
FC Answer	Yes	264 (72.3%)	6 (1.6%)
	No	27 (7.4%)	68 (18.6%)

In order to summarise the agreement between subjects' answers on the FC and the actual answers, Cohen's Kappa statistic was calculated (appendix 7.6). Using the data from table 7.1, a Kappa value of 0.76 was obtained; this signifies that subjects' answers to the column heading "Existed?" were accurate 76% of the time.

It was also hypothesised that subjects' answers on the route FC to the column headed "Existed?" would be more accurate than subjects' answers on the alphabetical FC. Table 7.2 shows the mean for subjects' answers to the question "did you know this underground station existed?" using the two different types of FC. From this table it can be seen that subjects using the route FC were correct for an average of 16.9 out of a maximum of 19 stations. Subjects using the alphabetical FC were correct for an average of 16.3 out of 19 stations. It is important to note that this is significantly better than chance i.e. if subjects simply answered "yes" to all stations, (appendix 7.7).

Table 7.2: Mean group answers to FC question "did you know this station existed?"

FC type	Correct	Incorrect	Don't Know
Route FC	16.9 (88.9%)	1.5 (7.9%)	0.6 (3.2%)
Alphabetical FC	16.3 (85.8%)	1.7 (8.9%)	1.0 (5.3%)

A one-factor analysis of variance for independent samples was used to estimate the effect of FC type on the mean number of correct responses; the resulting ANOVA table is shown in table 7.3.

Table 7.3: One-factor ANOVA table (independent samples) for effects of FC type on correct answers (mean) to the FC question “did you know this station existed?”

Source	Sum of Squares	df.	Variance	F - test
Between Groups	1.8	1	1.8	0.364
Within groups	89	18	4.944	
Total	90.8	19		

The obtained F value of 0.364 is smaller than the F value of 4.4139 for df’s = 1 and 18 at the 0.05 level. We therefore conclude that there is no significant difference between the group means for correct answers to the column heading “Existed?”.

7.3.2 “Do you know what this underground station might be used for?”

In order to assess the validity of the second new column on the FC, i.e. “What for?”, the results of the semi-structured interview were used to cross-reference the answers that subjects gave on the FC, i.e. subjects had to expand their FC answers and say what they thought the station could be used for. These answers were then checked against actual knowledge of what the stations could be used for (this was assessed by checking the amenities etc., within a ten minute walk of the station).

For each subject, six different stations were randomly selected for analysis (a different random selection was done for each subject in order to obtain information about all stations). From table 7.4, it can be seen that by using this validation method, subjects were correct on the FC for 101 answers out of a total of 107 (94.4%) and incorrect for 6 answers (5.6%). N.B. “Don’t Know” answers were excluded from the analysis here.

Table 7.4: Agreement between subjects’ FC answers and actual answers to the FC question “do you know what this station is used for?”

		Actual Answers	
		Yes	No
FC Answer	Yes	62 (57.9%)	3 (2.8%)
	No	3 (2.8%)	39 (36.5%)

In order to summarise subjects' answer agreement between the FC (using the semi-structured interview) and the actual answers, Cohen's Kappa statistic was calculated (appendix 7.8). Using the data from table 7.4, a Kappa value of 0.87 was obtained; this signifies that subjects' answers to the column heading "What for?" were accurate 87% of the time.

It was also hypothesised that subjects' answers on the route FC to the column headed "What for?" would be more accurate than subjects' answers on the alphabetical FC. Table 7.5 shows the mean for subjects' answers to the question "do you know what this underground station is used for?" using the two different types of FC. From this table it can be seen that subjects using the route FC were correct for an average of 5.5 out of a maximum of 6 stations. Subjects using the alphabetical FC were correct for an average of 4.6 out of 6 stations.

Table 7.5: Mean group answers to the FC question "do you know what this station is used for?"

FC type	Correct	Incorrect	Don't Know
Route FC	5.5 (91.7%)	0.3 (5.0%)	0.2 (3.3%)
Alphabetical FC	4.6 (76.7%)	0.3 (5.0%)	1.1 (18.3%)

A one-factor analysis of variance for independent samples was used to estimate the effect of FC type on the mean number of correct responses; the resulting ANOVA table is shown in table 7.6.

Table 7.6: One-factor ANOVA table (independent samples) for effects of FC type on correct answers (mean), to the FC question "do you know what this station is used for?"

Source	Sum of Squares	df.	Variance	F - test
Between Groups	4.05	1	4.05	2.216
Within groups	32.9	18	1.828	
Total	36.95	19		

The obtained F value of 2.216 is smaller than the F value of 4.4139 for $df's = 1$ and 18 at the 0.05 level. We therefore conclude that there is no significant difference between the group means for correct answers to the column heading “What for?”.

7.3.3 ***“Do you think you would ever have a need to use this underground station?”***

In order to assess the validity of the third new column on the FC, i.e. “Need?”, the results of the semi-structured interview were again used to cross-reference the answers that subjects gave on the FC, i.e. subjects had to expand their FC answers and say what they thought they would need to use the station for and also estimate this need (using a 5-point scale). These answers were again checked against the location of that station.

For each different subject, six different stations were randomly selected for analysis (these were the same stations that had been selected for asking about users’ knowledge). However, it was only possible to obtain data from the stations that subjects had answered “yes” to in the column “What for?”, i.e. we can’t ask subjects about their need if they don’t know what the station can be used for. From table 7.7, it can be seen that by using this validation method, subjects were correct on the FC for 79 answers out of a total of 90 (87.7%) and incorrect on the FC for 11 answers out of a total of 90 (12.2%). N.B. “Don’t Know” answers were excluded from the analysis here.

Table 7.7: Agreement between subjects’ FC answers and actual answers to the FC question “do you think you would ever have a need to use this underground station ?”

		Actual Answers	
		Yes	No
FC Answer	Yes	48 (53.3%)	1 (1.1%)
	No	10 (11.1%)	31 (34.4%)

In order to summarise subjects’ answer agreement between the FC (using the semi-structured interview) and the actual answers, Cohen’s Kappa statistic was calculated (appendix 7.9). Using the data from table 7.4, a Kappa value of 0.74 was obtained; this signifies that subjects’ answers to the column heading “Need?” were accurate 74% of the time.

It was also hypothesised that subjects' answers on the route FC to the column headed "Need?" would be more accurate than subjects' answers on the alphabetical FC. Table 7.8 shows the mean for subjects' answers to the question "do you know what this underground station is used for?" using the two different types of FC. From this table it can be seen that subjects using the route FC were correct for an average of 4.4 out of a maximum of 4.8 stations. Subjects using the alphabetical FC were correct for an average of 3.5 out of 4.6 stations.

Table 7.8: Mean group answers to the FC question "do you think you would ever have a need to use this underground station ?"

FC type	Correct	Incorrect	Don't Know
Route FC	4.4 (91.7%)	0.3 (6.2%)	0.1 (2.1%)
Alphabetical FC	3.5 (76.1%)	0.8 (17.4%)	0.3 (6.5%)

A one-factor analysis of variance for independent samples was used to estimate the effect of FC type on the mean number of correct responses; the resulting ANOVA table is shown in table 7.9.

Table 7.9: One-factor ANOVA table (independent samples) for effects of FC type on correct answers (mean), to the FC question "do you think you would ever have a need to use this underground station ?"

Source	Sum of Squares	df.	Variance	F - test
Between Groups	4.05	1	4.05	3.857
Within groups	18.9	18	1.05	
Total	22.95	19		

The obtained F value of 3.857 is smaller than the F value of 4.4139 for df's = 1 and 18 at the 0.05 level. We therefore conclude that there is no significant difference between the group means for correct answers to the column heading "Need?".

7.4 Discussion

The experimental hypotheses of this study were that:

- (1) Subjects' answers to the three new column headings would be accurate.
- (2) Subjects that received the route FC would be more accurate in their answers than subjects that received the alphabetical FC.

We can now discuss the results obtained with respect to each of the three column headings.

7.4.1 *“Existed?” column accuracy*

From the analysis of the results displayed in table 7.1, it was seen that a Kappa value of 0.76 was obtained. This value indicates that subjects' answers to this column on the FC agreed highly with the actual answers, i.e. subjects' answers on the FC to the question “did you know this underground station existed?” were very accurate. This new FC column therefore possesses high validity.

An obvious criticism of these findings could be that the distractor items (false stations) that were selected to assess the accuracy of subjects' answers, were in reality very obviously false. However, this view receives no support from the results shown in table 7.10, which displays the ranking of all the stations on incorrect answers. The scores were obtained by giving an incorrect answer a score of three and a “don't know” answer a score of one); the false stations are highlighted.

Table 7.10: Rank order of stations on incorrect answers to the FC question “did you know this station existed?”

Station	Score	Rank
Kinning Park	26	1
West Street	18	2
Charing Cross	13	3
Shields Road	10	4
Cessnock	8	5
Bridge Street	6	8
Ibrox	6	8
St. George’s Cross	6	8
Hyndland	5	9
Bellahouston	4	10
Buchanan Street	3	13
Kelvinhall	3	13
Partick	3	13
Woodlands	2	14
Kelvinbridge	1	15
Cowcaddens	0	18
Govan	0	18
Hillhead	0	18
St. Enoch	0	18

N.B. rank 1 = most incorrect answers and rank 18 = least incorrect answers

From table 7.10, it can be seen that the false stations (Charing Cross, Hyndland, Bellahouston and Woodlands), were ranked 3, 9, 10 and 14 respectively; these stations were therefore, not obviously false. It is interesting to note that the stations ranked first and second, i.e. those with most incorrect answers, actually do exist.

Another interesting aspect of these results is that before administering the FCs, the researcher hypothesised that of the false stations, Charing Cross would receive most incorrect answers (since it has a surface train station and its name could be confused with St. George’s Cross); Hyndland would receive the second most incorrect answers (since it has a train station); Bellahouston would receive the third most incorrect answers (since it has no train station but was located accurately on the route FC,

geographically) and lastly that Woodlands would receive the fewest incorrect answers (since it has no train station and was located inaccurately on the route FC, geographically). This hypothesised order of the false stations was the actual order obtained (table 7.10). Thus although no direct validity check was possible we created a situation where false items generated similar numbers of errors to true items, and it seems reasonable to estimate that errors for the “Existed?” column in a FC without distractors might be at a comparable level.

The second analysis carried out on subjects’ answers to the column heading “Existed?” was concerned with the differences (if any) between subjects’ answers using the alphabetical FC and those using the route FC, (table 7.2).

It was hypothesised here, that subjects using the route FC would be more accurate in their answers to the column heading “Existed?” than those using the alphabetical FC, the reason for this being that the route FC matched subjects’ conceptual model of the underground system more closely than the alphabetical FC, i.e. subjects probably encoded the stations geographically and by neighbouring stations.

The results reported did not support this hypothesis, i.e. there were no significant differences between subjects’ correct responses on the route FC compared with those on the alphabetical FC; both FCs scored highly on correct responses, 88.9% and 85.8% respectively (table 7.2). There are two different ways of interpreting these results. Firstly, it may be the case that subjects’ memory of the underground stations was not based on an internal representation of the actual geographical route, or secondly, that the number of features on the FCs (i.e. 19 stations) was too small to produce any significant differences between the FCs.

Having established that subjects’ answers to the column heading “Existed?” are accurate in the context of this study, we can now look at the possible usefulness of the information that this column yields.

Data from this column may be a good indicator of a user’s knowledge of the system; this is sometimes referred to as “experienced user performance” or EUP (Jordan 1992). Information from this new column may help to identify areas in which users could increase their EUP, e.g. a quicker way of performing a particular task. Of course it may be the case that users are not aware of a command’s existence simply because they never use this command; this can be easily verified by looking at either the automatic data log (if available) or by FC answers to the column heading “Used?”. As we shall

also see, the information gathered from this column may be valuable when used in conjunction with information gathered from other columns on the FC.

7.4.2 “What for?” column accuracy

From the analysis of the results displayed in table 7.4, it was seen that a Kappa value of 0.87 was obtained. This value indicates that subjects’ answers to this column on the FC (checked through the semi-structured interview), agreed highly with the actual answers, i.e. subjects’ answers on the FC to the question “do you know what this underground station might be used for generally?” were again very accurate. This second new FC column therefore possesses high validity.

Of the fifteen real stations listed, subjects on average knew what 10.6 could be used for; the variation between subjects ranged from one subject that knew what all fifteen stations could be used for to one that only knew what five of the stations could be used for. Subjects in general, therefore, knew what a large subset of the stations could be used for. The complete data for subjects’ answers to this column is shown in table 7.11.

Table 7.11: Subjects' answers to the FC question "do you know what this station could be used for?"

Subject Answers			
Subject No.	Yes	No	Don't Know
1	11	4	0
2	8	7	0
3	14	1	0
4	9	6	0
5	8	7	0
6	14	1	0
7	15	0	0
8	9	6	0
9	9	3	3
10	11	4	0
11	9	0	6
12	10	3	2
13	7	8	0
14	14	1	0
15	5	10	0
16	8	6	1
17	13	2	0
18	11	4	0
19	11	4	0
20	10	5	0
Total	206	82	12
Average	10.3	4.1	0.6

Although the validation of this column is based on whether subjects could in fact answer what an underground station is used for generally (i.e. what amenities, places of interest, etc. were located close by), there is no reason to believe that this column would not be as useful for asking subjects about their knowledge of computer command functions.

The second analysis carried out on subjects' answers to the column heading "What for?", was concerned with the differences (if any) between subjects' answers using the alphabetical FC and those using the route FC, (table 7.4).

Again it was hypothesised that subjects using the route FC would be more accurate in their answers to the column heading "What for?" than those using the alphabetical FC. The results reported however, did not support this hypothesis, i.e. there were no significant differences between subjects' correct responses on the route FC compared with those on the alphabetical FC; both FCs scored highly on correct responses, 91.7% and 76.7% respectively (table 7.4).

As in the previous section, there are different ways of interpreting these results. It could be the case that subjects' memory of the underground stations was not based on an internal representation of the actual geographical route, or secondly, that the number of stations sampled (i.e. 6), was too small to produce any noticeable differences between the FCs.

Having established that subjects' answers to the column heading "What for?" are accurate in the context of this study, we can now look at the possible usefulness of the information that this column yields. It may be the case that commands which receive many "no" responses in this column have a "guessability" problem, i.e. the command name does not obviously denote its function to the user or, if the command has also been used then there may be a "memorability" problem, i.e. subjects used this command at some point but can't remember what it is for later on. Again, the information gathered from this column may be valuable when used in conjunction with information gathered from other columns on the FC, such as "Need?".

7.4.3 "Need?" column accuracy

From the analysis of the results displayed in table 7.7, it was seen that a Kappa value of 0.74 was obtained. This value indicates that subjects' answers to this column on the FC (checked through the semi-structured interview) agreed highly with the actual answers, i.e. subjects' answers on the FC to the question "do you think you would ever have a need to use this underground station?" were again accurate (although less so than the other columns). This third new FC column therefore possesses high validity,

provided the subject knew what the item was for, and so could go on to answer this question.

Yet again there is good reason to believe that the high validity exhibited by this new column heading would extend to research in an HCI setting, i.e. asking subjects to estimate their need for a particular command. Similarly, the information that this column heading might yield, could be extremely valuable in software evaluation. For instance it may be possible to identify certain commands that are very rarely used even although subjects indicate that they know what they are for and need to use them; if so there may be a “reminding” problem. In order to answer these questions it may be useful to back up FCs with semi-structured interviews wherever necessary.

Again it was hypothesised that subjects using the route FC would be more accurate in their answers to the column heading “Need?” than those using the alphabetical FC. The results reported however, did not support this hypothesis, i.e. there were no significant differences between subjects’ correct responses on the route FC compared with those on the alphabetical FC; both FCs scored highly on correct responses, 91.7% and 76.1% respectively (table 7.8).

These results could be interpreted in the same way as before, i.e. either subjects’ memory of the underground stations was not based on an internal representation of the actual geographical route, or the number of stations sampled (on average 4.6 for each subject), was too small to produce any noticeable differences between the FCs.

It may be possible to increase the validity of this column by listing the command on the FC by its function as opposed to its name. By giving a brief description of each command’s function it would then be possible to obtain information about subjects’ need for this command even if they had previously indicated on the FC that they did not know that it existed or what it is for (this was not possible in the present study i.e. subjects’ answers to column 3 were omitted from the analysis if they did not know what the command was for).

Again this may highlight a “guessability” problem, i.e. not knowing what a command is for (when given its name), but stating a strong need to use it (when given a description of its function). One problem with this idea of giving subjects a brief description of the command function is in designing descriptions that accurately signify the function to subjects. Future research should investigate this issue.

The two issues of this study that relate to memory processes are:

- Whether the recall vs. recognition debate discussed earlier is still applicable here.
- Designing FCs to match users internal memory representations.

We will now look at the latter of these two first.

7.4.4 *Memory processes in different FC columns*

In this study it was found that attempts at matching internal mental representations more closely (route FC), did produce slightly more accurate results for the “Need?” and “What for?” columns, however these differences were not significant. In explaining this, the view taken is that subjects probably do to an extent, form internal representations of the train stations based on geographical layout (particularly sequence); the problem in this study was that since the domain contained a small number of features, it is likely that subjects had a good knowledge of the system in general and that this masked any potential differences between the FCs. Perhaps with a domain that contained many features, e.g. a word-processing system, the idea of simulating subjects’ encoding process more closely through FC design would be more appropriate. The question that arises is “how do subjects form internal memory representations (encode) menu commands?” It is reasonable to expect that this is done by task → action associations (or goal directed behaviour). Study 8 will explore this issue more fully.

As we have seen from earlier chapters, the idea of explaining FC success on the fact that they rely on recognition as opposed to recall is inappropriate; this is particularly true in the present study. It seems unlikely that subjects’ answers on the FC were based purely on recognition of the station names. What seems more likely is that recognising the station name was the first stage in subjects reconstructing (recalling) information about the station (this was probably based on things like: recalling where the station is located, what it looks like, why they might have used it, etc.). One other possibility is that when subjects were completing the FCs they did so by first of all recalling stations that they used often and then looked for these (recognition) on the FC, i.e. recall first then recognition.

So far it can be argued that FC columns seem to operate by subjects' firstly recognising the feature and then using this to reconstruct (recall) extra information about it; the actual amount of reconstruction (mental elaboration) done, probably varies depending on the column e.g. it can be reasonably assumed that more processing is required to remember how often you use a station than is required to remember whether that station actually existed. At this stage however, it is useful to look at different views of long-term memory and memory tasks.

7.4.5 *FCs and memory: an explicit reconstructive process*

Tulving (1983; 1985), has put forward a view of long term memory in which he proposes a tripartite system where a distinction is drawn between memories associated with personal recollection and others that are not. The three distinct components are:

- Episodic memory - information is stored with mental tags about where, when and how the information was picked up, (i.e. the material in memory concerns fairly sharp, circumscribed episodes).
- Semantic memory - "memory that allows the individual to construct mental models of the world...It makes possible the cognitive representation of objects, situations, facts and events" (Tulving 1985).
- Procedural memory - memory about how to do something, (i.e. memory that is not consciously accessible).

Although all 3 are held to be functionally distinct, they are also interactive.

Since episodic and semantic memory can both be consciously accessed, it is these two types that we are interested in discussing FCs. The attraction of this view is that knowledge about the world can come to be represented independently of the events that give rise to that event in the first place, e.g. "Remembering the meaning of a new word, may not at first be possible without some episodic record confirming why the word means what it does. With time, however, the word's meaning is assimilated into semantic memory, and there is no need for the episodic record to be maintained", (Parkin 1993). The central point about the theory is the idea that the record of experience, (episodic memory), resides separately from the knowledge gained from experience, (semantic memory).

How do these descriptions relate to the memory processes involved in FCs in the present study? With respect to FC questions such as “have you ever used this station?” and “how often do you use this station?” it is highly likely that these relate more to Tulving’s episodic memory, i.e. in answering these questions subjects are most likely to think of occasions (episodes) in which they themselves actually used that station. However, the FC questions “do you know what this might be used for?” and “did you know that this station existed?” may involve semantic memory to a greater extent than episodic memory, e.g. some answers that subjects give to these questions are likely to be based on self-knowledge of actual episodes when they have used these stations, however, answers are more likely to be based on world knowledge of the area in which the stations are located, (this is perhaps more true of answers to the FC question “do you know what this might be used for?”). N.B. this world knowledge is not the same as that proposed by Norman (1988), in which he talks about behaviour being determined by a combination of information residing in the world (world knowledge) and information in the head. In our definition, world knowledge relates to information in the head, about the world.

Attempts at finding evidence to support the semantic-episodic distinction is problematic and relies heavily on research with patients suffering from amnesia. However, the overall finding from this research is summed up by Parkin (1993), “the data do not support a functional separation between episodic and semantic systems. Instead, it seems more correct to see these forms of memory as differing expressions of a single system”.

One suggestion is that the terms episodic and semantic should be replaced by a single term “declarative memory”, i.e. any memory that is consciously inspectable (Squire 1987). However, we should not reject the terms episodic and semantic altogether, because although they may not map on to different memory systems, they have important descriptive value in that they define different types of memory tasks.

A similar dichotomy to the episodic-semantic distinction has been proposed in situations where memory tasks are classified as either explicit or implicit (Graf and Schacter 1985). This theory attempts to understand long-term memory by examining how it responds to different forms of memory tests. These tests can be either explicit (i.e. one which requires the subject to recollect a previous learning event), or implicit (i.e. any memory task in which a subject’s memory for a learning event is tested without specific reference to that event). A well known characteristic of explicit memory tasks is that

they depend on the degree of conscious effort expended during learning. From the above definitions it should be clear that the memory test involved in FCs is explicit, regardless of which column is being answered, i.e. in answering all columns on the FC, the subject must make a conscious attempt to recollect stored information about the train station (whether that information relates to an episode in which they have used or seen the train station, or whether it relates to their world knowledge).

Although, both of these system classification theories (episodic-semantic and explicit-implicit), have been criticised (see section 4.4.4), it should be pointed out that the transfer-appropriate processing account of memory described earlier is inappropriate for explaining the current study. There was very little overlap between the types of operations performed at learning (i.e. using the underground) and those required at test (i.e. the FC). The FCs used in this study simply listed stations by their name. Although we wanted to ask subjects about whether they knew the stations existed or what they could be used for, we were not able to supply them with additional information such as pictures of the stations, maps showing locations, etc. (i.e. we could not give subjects much information at retrieval that would overlap with the processing done at learning).

Given that this is the case, it is highly likely that the memory prompt on the FC, i.e. the station name, enabled/required subjects to perform an explicit reconstructive process based on physical attributes of the station as well as personal experience; this process has more in common with G-R models of memory described earlier. The high accuracy of subjects' responses in this study, suggest that given time, subjects were able to perform this detailed reconstruction; as mentioned earlier however, this was probably due to the fact that the subjects were all experienced users of the underground system and the system itself contained a small number of features (fifteen).

7.5 Summary

To conclude, therefore, the results of this study suggest that subjects' answers to these new FC columns are accurate, and could produce valuable information for software evaluation. However, it is worth noting that the domain in which the study was conducted, contained only 15 features (stations) or 19 including false stations. It is likely that "real-life" use of FCs in an HCI setting would involve computer applications containing many more features e.g. the menus in "Microsoft WORD 5.0" contain 95

textual commands alone (before customisation). The validation of these new columns in an HCI setting therefore becomes a pressing issue.

Chapter 8

Employing feature checklists for measuring users' knowledge and need for menu commands

8.1 Introduction

The previous study (study 4), suggested that additional columns on FCs that asked users about their knowledge of the existence and function of features, as well as their need to use these features, may provide accurate and useful information. It was suggested that the data from these new columns could provide software designers with valuable information relating to the menu commands contained in a particular piece of software. However, since this research was conducted outwith the field of HCI, it became apparent that validation of these new columns within an HCI context was a pressing concern.

In addition to this, a discussion of the relationship between memory research and FCs has led to the suggestion that listing menu commands by semantic descriptions of their function may be a more appropriate memory prompt than simply listing menu commands by their names; if so this may have important implications for FC validity. The idea behind this view is that listing commands by descriptions of their function will match the “real-life” task of users much more closely and thereby their encoding procedure; i.e. when users want to do something on the computer they usually think of the task they want to do and then select the appropriate command that will enable them to perform this task (task → action mappings or goal-directed behaviour).

It is the intention of this study to try and address both of these issues, i.e. to replicate the findings of study 4 in an HCI setting (that the additional column headings possess

high validity), and to compare the validity of the standard FC with that of a semantic FC (assess the usefulness of semantic descriptions as a memory prompt).

In order to do this however, it was important to:

- Obtain subjects that were competent in using a particular computer package.
- To allow subjects to use the package in a natural or “real-life” situation.

In addition to this it was also important that the researcher was able to record all command selections that users made. Given these requirements, it was decided to recruit psychology summer students to participate in a word-processing course. Although these students had little or no experience in word-processing it would be possible to train them to reach a reasonable level over a six hour course (split into two sessions). In the last hour of this second session the students would then word-process their own curriculum vitae (CV), using the skills they had recently acquired (natural usage). During this course it would also be possible for the researcher to use an automatic data-logging application to record subjects’ command selections.

In addition to these primary aims, this study will also look at reliability of the standard FC when it is administered to the same subjects at two different times, and try and gain an insight into the tasks that subjects are actually performing on the computer through the semantic FC. The aims of this study are summarised below:

- (1) To assess the accuracy of the standard FC, i.e. how accurate are subjects’ answers about whether they used commands?
- (2) To assess the accuracy of the extra columns on the standard FC, i.e. how accurate are subjects’ answers about commands in the columns “What for?” and “Need?”.
- (3) To assess the reliability of the standard FC, i.e. when subjects are given the same FC at two different times (with a gap of one week in between), are their answers the same?
- (4) To compare the validity of the standard FC with that of a combined FC which lists both the command name and its semantic description as memory prompts, i.e. are there any differences between the number of commands correctly identified as being used on these FCs?

- (5) To compare the accuracy of the standard FC with that of the semantic FC, i.e. are there any differences between the number of commands correctly identified as being used on these FCs?

8.2 Method

Subjects: The study contained 16 subjects in total, of these 12 were female and 8 were male. Their ages ranged from 16 to 43 years of age with a mean of 25.9 years.

Subjects were recruited from the psychology summer school at the University of Glasgow and were invited to enrol for a word-processing course as an alternative to a couple of the laboratory classes that made up part of the summer school course. All of the subjects had little or no experience of word-processing but were eager to participate in the course.

Apparatus/Stimuli: The word-processing course took place in the “Apple Macintosh” teaching suite in the University Library using “Apple Macintosh IIx’s”. The word-processing package used was “Microsoft WORD 5.0”.

Procedure: All subjects attended two, 3 hour classes over the period of one week; these classes took them from a basic introduction to computing to a competent level of word-processing. There were three separate word-processing groups, each containing seven, five and four students respectively. During all of these classes the application “Tempo II Plus” was running on the computers that the subjects were using; this recorded all the command selections that subjects made. The course took the format of the lecturer demonstrating progressively harder tasks on the computer using a “LCD view frame”. After each demonstration the students then performed some prepared exercises that enabled them to practice these tasks, (appendices 8.1, 8.2 and 8.3).

In the second half of the final session students word-processed their own CV and were told to use any of the techniques and commands that they had learned over the course. The students were also encouraged to try and explore any other commands that they wished (i.e. ones that they had not been specifically taught during the course) in order to learn more about the word-processing package.

At the end of this second session, half of the subjects were given a standard FC to fill in (appendix 8.4), then a computer experience questionnaire (appendix 4.1) and finally a semantic FC (appendix 8.5). The remaining half of the subjects received these FCs in

the reverse order, i.e. the semantic FC first, followed by the standard FC, with the computer experience questionnaire in the middle.

After completing all questionnaires, subjects were invited to come back a week later to complete some more questionnaires and to participate in a short ten minute interview relating to the course (appendix 8.6); they were informed that they would be paid £5.00 for doing so.

When subjects came back for their semi-structured interview a week later, half were given an abbreviated standard FC to complete (appendix 8.7), followed by a combined standard/semantic FC (appendix 8.8), the other half were given these FCs the other way round. The semi-structured interview was used to validate the additional columns on the standard FC. The experimental conditions are explained in table 8.1.

Table 8.1: Experimental conditions used in study 5

Session	No. of Subjects	FCs + Order	Interview
A	8	Standard/Semantic	No
	7*1	Semantic/Standard	No
B*2	7*3	Combined/Standard	Yes
	6*4	Standard/Combined	Yes

*1 One subject left without completing any FCs.

*2 Session B took place one week later.

*3 One student in this condition failed to turn up for the interview.

*4 Two students in this condition failed to turn up for the interview.

Design: With respect to the aims of the study listed in the introduction, the following experimental designs were employed:

- (1) To assess the accuracy of the standard FC, all sixteen subjects were issued with a standard FC at the end of the second session of the course (one subject never completed any FCs); half of the subjects received this FC first in the series of two different FCs, the other half received it second. Subjects' answers on the standard FC as to whether or not they had used the commands, were validated against the data log. It is therefore a two group design where the independent variable is the

order of presentation of the standard FC (first or second) and the dependent variable is the number of commands that subjects could correctly remember using.

- (2) To assess the accuracy of subjects' answers on the standard FC to the columns headed "What for?" and "Need?"; a semi-structured interview was conducted with all subjects one week later. In this interview, eight commands were selected at random (for purposes of time) and subjects were again asked "do you know what this command actually does?", and "have you ever had a need to use this command?".
 - (a) For the column headed "What for?", subjects had to say what each command actually did (rather than answer yes or no). Subjects' answers on the FC were then marked as either correct (i.e. saying yes when they did know what the command was for or no when they didn't know), or incorrect (i.e. saying yes when they did not know what the command was for or vice versa). The design was therefore a single group with repeated observations on the same subjects under two conditions. The dependent variable is the accuracy of subjects' answers on the standard FC.
 - (b) For the column headed "Need?", subjects had to estimate their need for each command using a need scale (appendix 8.9); an answer of 1, 2 or 3 on the need scale was taken as a "yes" and an answer of 4 or 5 was taken as a "no". Using this data all subjects' answers on the standard FC were marked as being either correct (i.e. saying "yes" when they did have a need for that command or "no" when they didn't have a need), or incorrect (i.e. saying "yes" when they did not have a need for that command or "no" when they did have a need). Yet again it is a single group design with repeated observations on the same subjects under two conditions. The dependent variable is the accuracy of subjects' answers on the standard FC.
- (3) In order to assess the reliability of the standard FC, all subjects were issued with an abbreviated version of the FC one week later. Each subject's answers on the standard FC at these two times were then compared to see if there was still the same agreement. It is therefore a single group design with repeated observations on the same subjects under two conditions (time A and time B); the dependent variable is the agreement within subjects' answers on the standard FC at these two times.
- (4) In order to compare the accuracy of the standard FC with that of the combined FC, both of these instruments were administered to all subjects one week later; the

answers to both were then validated against the data log. It is therefore a two group design (standard FC vs. combined FC), and the dependent variable is the number of commands that subjects could correctly remember using.

- (5) In order to compare the accuracy of the standard FC with that of the semantic FC, both of these instruments were administered to all subjects immediately after the course finished; the answers to both were then validated against the data log. It is therefore a two group design (standard FC vs. semantic FC), and the dependent variable is the number of commands that subjects could correctly remember using.

8.3 Results

The results section is split into five sub-sections each of which relates to the experimental aims.

8.3.1 *Standard FC accuracy*

When the standard FC was issued first, immediately after the word-processing course finished (time A), subjects correctly recalled using 72.4 out of a total of 80 commands, i.e. 90.5%. However, when the FC was issued second at time A (i.e. after the semantic FC), this dropped to 66.1 commands, or 82.7%. This is shown in table 8.2.

Table 8.2: Subjects' correct recall of command usage (mean) using the standard FC at time A

Order of Presentation	No. of Subjects	Correct Usage Score	Std. Error
1st	8	72.4 (90.5%)	1.546
2nd	7	66.1 (82.7%)	5.492

A t-test for independent samples (separate variance) was used to test the difference in the means between the two groups (order) for correct usage scores. Our obtained t-value of 1.105 does not exceed either of the averaged t-values at the 5% level for 6 and 7 degrees of freedom. We therefore cannot reject the null hypothesis and conclude that there is no significant difference between the mean correct usage scores for the FC presented first and the FC presented second.

8.3.2 Standard FC accuracy of extra columns

Since the FCs in this study did not include false command names, it was not possible to validate the answers to the column headed “Existed?”. However, using the information from the semi-structured interview we can still validate the columns headed “What for?” and “Need?”.

In order to assess the accuracy of subjects’ answers to the column on the FC headed “What for?”, the semi-structured interview was used to measure subjects’ actual knowledge of what commands actually did. From table 8.3, it can be seen that subjects’ answers on the standard FC to this column, were correct for 82 of the commands (90.1%) and incorrect for 9 of the commands (9.9%).

Table 8.3: Subjects’ answers to the FC question “do you know what this command actually does?”

		Semi-Structured Interview	
		Yes	No
FC Answer	Yes	51 (56.0%)	6 (6.6%)
	No	3 (3.3%)	31 (34.1%)

In order to summarise how accurate subjects were in their FC answers, Cohen’s Kappa statistic was calculated (appendix 8.10). Using the data from table 8.3, a Kappa value of 0.79 was obtained; this signifies that subjects were accurate in their answers to the column heading “What for?”, 79% of the time.

In order to assess the accuracy of subjects’ answers to the column on the FC headed “Need?”, the semi-structured interview was again used to measure subjects’ actual need for commands. However, it was only possible to obtain data from the commands where subjects had answered “yes”, to the column “What for?”; i.e. we can’t ask subjects about their need if they don’t know what the command can be used for. From table 8.4, it can be seen that subjects’ answers on the FC to this column, were correct for 50 of the commands (81.9%) and incorrect for 11 of the commands (18%).

Table 8.4: Subjects' answers to the FC question "do you think you would ever have a need to use that command?"

		Semi-Structured Interview	
		Yes	No
FC Answer	Yes	24 (39.3%)	3 (13.1%)
	No	8 (4.9%)	26 (42.6%)

In order to summarise how accurate subjects were in their FC answers, Cohen's Kappa statistic was calculated (appendix 8.11). Using the data from table 8.4, a Kappa value of 0.64 was obtained; this signifies that subjects were accurate in their answers to the column heading "Need?" 64% of the time. It is worth noting that this is a considerably lower score than the "What for?" column.

8.3.3 Standard FC reliability

In order to assess the reliability of the standard FC, we administered the same FC to subjects one week later (time B), thus enabling us to compare the agreement between the same subject's answers on the same FC at two different times. Before looking at the results of this however, we can look at the validity of the FC (administered one week later), in isolation.

It can be seen from table 8.5, that when the standard FC was issued first, one week later (time B), subjects correctly recalled using 64 out of a total of 80 commands, i.e. 80.0%. However, when the FC was issued second at time B (i.e. after the combined FC), this dropped to 62.9 commands, or 78.6%. Clearly, when the standard FC is presented at time B, the order of presentation has even less of an effect on the correct usage score than at time A.

Table 8.5: Subjects' correct recall of command usage (mean) using the standard FC at time B

Order of Presentation	No. of Subjects	Correct Usage Score	Std. Error
1st	6	64.0 (80.0%)	10.296
2nd	7	62.9 (78.6%)	5.521

When comparing subjects' answers on the FC over the period of one week, it was decided to look at only the data in which the order of FC presentation was the same for each subject at both times, i.e. subjects that received the FC first at time A and time B, and subjects that received the FC second at time A and time B; in total, this reduced our analysis to the results of eight subjects. Table 8.6, shows the agreement in subjects' answers between the FC issued at time A (immediately after word-processing class) and the FC issued at time B (one week later).

Table 8.6: Agreement in subjects' answers using the standard FC at time A and time B

		TIME B	
		Yes	No
TIME A	Yes	180 (30.6%)	21 (3.6%)
	No	66 (11.2%)	321 (54.6%)

In order to summarise subjects' answer agreement on the standard FC at two different times over the period of one week, Cohen's Kappa statistic was calculated (appendix 8.12). Using the data from table 8.6, a Kappa value of 0.70 was obtained; this signifies that over the period of a week, subjects' answers on the FC as to whether they used a command were consistent 70% of the time.

8.3.4 Standard FC accuracy vs. combined FC accuracy

Another aspect of the study was to improve the validity of FCs by including a description of each command's function alongside that command. It was hoped that this semantic description might be an additional memory prompt and consequently increase the accuracy. In this study we called the FC that included both the command name and a brief description of its function the combined FC. It should be noted that in order to compare these FCs fairly, the combined FC was compared with the standard FC at time B only (this is because we could not administer the combined FC at time A due to time constraints).

Table 8.7, shows the mean correct usage scores for both the standard FC and the combined FC, when they were administered first at time B. From this table we can see that subjects using the standard FC could correctly identify using 64 commands out of a

total of 80 (i.e. 80%), whilst subjects using the combined FC could correctly identify using 61.9 commands (i.e. 77.4%).

Table 8.7: Correct usage scores (mean) for the standard FC and the combined FC when administered first, at time B.

FC Type	No. of Subjects	Correct Usage Score	Std. Error
Standard FC	6	64.0 (80.0%)	4.203
Combined FC	7	61.9 (77.4%)	1.831

It can be seen that contrary to our expectations, there is very little difference in correct usage scores between the standard FC and the combined FC; (overall subjects using the combined FC perform slightly worse than those using the standard FC, although this is clearly insignificant).

8.3.5 Standard FC accuracy vs. semantic FC accuracy

In order to assess the accuracy of the semantic FC for measuring what commands subjects used, we again used the information obtained from the data log. Each task described on the semantic FC had a corresponding command name in the pull down menus. Table 8.8, shows that when the semantic FC was issued first, immediately after the word-processing class finished (time A), subjects correctly recalled using 51.4 out of a total of 80 commands, i.e. 64.3%. However, when the semantic FC was issued second at time A (i.e. after the standard FC), this rose to 53.9 commands, or 67.3%.

Table 8.8: Subjects' correct recall of command usage (mean) using the semantic FC at time A

Order of Presentation	No. of Subjects	Correct Usage Score	Std. Error
1st	7	51.4 (64.3%)	6.321
2nd	8	53.9 (67.3%)	4.291

When we compare the accuracy of the semantic FC with that of the standard FC (table 8.9), we can see that using the semantic FC, subjects correctly recalled using 51.4

commands out of a total of 80 (i.e. 64.3%), whilst using the standard FC subjects could recall using 72.4 commands (i.e. 90.5 %).

Table 8.9: Correct usage scores (mean) for the standard FC and the semantic FC when administered first, at time A.

FC Type	No. of Subjects	Correct Usage Score	Std. Error
Standard FC	8	72.4 (90.5%)	4.373
Semantic FC	7	51.4 (64.3%)	6.321

A t-test for independent samples (separate variance) was used to test the difference in the means between the two groups (i.e. FC type) for correct usage scores. Using the appropriate formula we obtained a t value of 7.588. Our averaged t-value at the 5% level is 2.406 and 3.603 at the 1% level (two-tailed). Since our obtained t-value of 7.588 exceeds both of these we can reject the null hypothesis and conclude that there is a significant difference between the mean correct usage scores of the two FCs.

8.4 Discussion

The results reported in table 8.2, show that the standard FC produces highly accurate information about the commands that subjects were using. These results are consistent with those of earlier studies, in that the most accurate results are found when the FC is presented first (90.5% of commands were correctly identified as being used). Although no significant difference was found between presenting the FC first or second, it should be noted that subjects could remember on average 6 commands more when the FC was presented first (i.e. 8% of the total). Yet again this trend could be explained by the view proposed by Mandler that recognition based on familiarity decays rapidly over time (Mandler 1967).

8.4.1 “What for?” column accuracy

It was hypothesised that the additional column headings on the standard FC would possess high validity, i.e. produce accurate results. From the results shown in table 8.3, it can be seen that subjects’ answers on the FC to the question “do you know what

this command does?”, were accurate 79% of the time. This new column heading could provide software designers with useful information about subjects’ knowledge of command function in a particular application. For instance, commands that receive many “no” responses in this column may have a “guessability” problem, i.e. the command name does not obviously denote its function to the subject in this particular study.

Of the 80 commands contained in “Microsoft 5.0” used in this study, there were 2 commands for which none of the subjects knew their function (“Ribbon” and “Table Cells”); and there were 10 commands for which all fifteen subjects knew their function (“Open”, “Save”, “Print”, “Cut”, “Paste”, “Ruler”, “Italic”, “Spelling”, “Thesaurus” and “Word Count”). It is interesting to note that during the word-processing course, the demonstrator actually explained and showed (to all subjects), the function of 24 commands. These results are similar to those found by Draper (1985), in which users of the “UNIX” system were highly specialised in the commands that they used and knew about; the highest number of commands used by any individual was around 60%.

The actual number of subjects that knew what each command actually did, is shown in appendix 8.13. Table 8.10, shows a summarised version of this, i.e. subject knowledge of commands broken down into three categories (high, medium and low). The commands shown in bold are those that the demonstrator specifically taught and explained to the subjects.

N.B. In table 8.10 the maximum score is 15 (i.e. 15 subjects know what this command actually does)

Table 8.10: Summary of subjects' knowledge of command function

Knowledge level	Command name
Low ($> 0 \leq 5$)	Print Merge; Clear; Go To; Glossary; Normal; Outline; Ribbon; Hide/Show; Table; File; Picture; Character; Paragraph; Section; Document; Border; Table Cells; Table Layout; Style; Renumber; Sort; Repaginate Now; Preferences; Commands; Show Clipboard; New Window; Untitled 1.
Medium ($> 5 \leq 10$)	New; Save As; Page Setup; Find; Page Layout; Date; Courier; Helvetica; Monaco; Palatino; Symbol; Times; Venice; Zapf Dingbats; Grammar; Help.
High ($> 10 \leq 15$)	Open; Close; Save; Print Preview; Print; Quit; Undo/Redo; Cut; Copy; Paste; Select All; Replace; Ruler; Header; Footer; Page Break; Plain Text; Bold; Italic; Underline; 9 Point; 10 Point; 12 Point; 14 Point; 18 Point; 24 Point; Athens; Cairo; Chicago; Geneva; London; Los Angeles; New York; San Francisco; Spelling; Thesaurus; Word Count.

Table 8.10 clearly shows that subjects in the course could remember the function of the commands that they were specifically taught; however they did not learn the function of many other additional commands (except font sizes and styles). It is the view of the researcher that the information obtained from the “What for?” column on the FC would be most useful when it is looked at conjunction with information from other FC columns; e.g. situations in which commands are selected by subjects but their function is not recalled a short time later (poor “memorability” or “guessability”), or situations where subjects know what the command does but indicate through the “Need?” column that they have no need to use it (redundant commands).

8.4.2 “Need?” column accuracy

From the results shown in table 8.4, it can be seen that subjects' answers on the FC to the question “do you think you would ever have a need to use that command?”, were accurate 64% of the time. This accuracy level is much lower than we would have

hoped for, and indicates that subjects were not as consistent in their answers to whether or not they actually needed a command. As a result, the usefulness of the information from this column (at least in this study) may be somewhat limited.

There are a number of possible reasons why the accuracy of this column is not as great as anticipated. Firstly, the actual number of commands that we obtained information on was very low i.e. 61 commands across 13 subjects; this was because we could only ask subjects about their need for a command if they knew what it actually did, in addition we had to exclude “don’t know” answers for the purpose of our statistical analysis.

Secondly, we have no guarantee that subjects did not use the word-processing package between receiving the FC and then receiving the semi-structured interview a week later, and thereby discover a previously unrealised need for some commands.

A final possible reason may be due to the fact that on the FC, subjects answered “yes”, “no” or “don’t know” about their need for commands, however on the semi-structured interview subjects were asked to answer using a need scale (appendix 8.9). As a result it may have been the case that the two measures could have shown a subject to change their answer across time whereas in fact the subject was still (in their own mind) answering the same; i.e. a subject may have said “yes” on FC and then indicated a need of “4” on the need scale (weak need), this was then marked as a negative answer on the semi-structured interview. In future research it may therefore be a useful idea for subjects’ answers to the “Need?” column to be in the form of a numerical estimate anchored to time e.g. 1 = “every day”, 2 = “once a week”, 3 = “once a month”, etc.

Despite these problems we can still use the information from the “What for?” and the “Need?” columns to demonstrate a use for FCs. Table 8.11, shows a summarised version of the number of subjects that indicated they would have a need to use each command (the actual numbers are shown in appendix 8.14). Again, subjects’ need for each command is broken down into three categories (high, medium and low).

N.B. In table 8.11 the maximum score is = 15 (i.e. 15 subjects said they would have a need to use this command)

Table 8.11: Summary of subjects' need for commands

Knowledge level	Command name
Low ($> 0 \leq 5$)	New; Page Setup; Print Merge; Clear; Find; Go To; Glossary; Normal; Outline; Ribbon; Hide/Show; Table; Date; File; Picture; Character; Section; Document; Border; Table Cells; Table Layout; Style; 9 Point; 10 Point; 12 Point; 14 Point; 24 Point; Athens; Cairo; Chicago; Courier; Geneva; Helvetica; London; Los Angeles; Monaco; Palatino; San Francisco; Symbol; Times; Venice; Zapf Dingbats; Grammar; Renumber; Sort; Repaginate Now; Preferences; Commands; Help; Show Clipboard; New Window; Untitled 1.
Medium ($> 5 \leq 10$)	Close; Save As; Print; Undo/Redo; Copy; Select All; Replace; Page Layout; Paragraph; Plain Text; New York.
High ($> 10 \leq 15$)	Open; Save; Print Preview; Quit; Cut; Paste; Ruler; Header; Footer; Page break; Bold; Italic; Underline; 18 Point; Spelling; Thesaurus; Word Count.

By comparing the data from the “What for?” and “Need?” columns (table 8.10 and table 8.11), we can see that there are a number of cases in which subjects indicate that they know what a command does but that they don’t have a need to use it (or at least very little need); these are all either Font sizes or Font Styles. Rather than say that these commands may be unnecessary it may be the case that the subjects in this study may have more of a need for these commands if they word-processed more extensively.

A useful alteration to the “Need?” column that future research could look at is the possibility of including a description of each command’s function along with its name as a memory prompt; this would enable subjects to indicate their need for a command even if they had previously stated that they didn’t know what it did. As we shall see later, a problem with this is in creating a description that denotes each command’s function correctly.

8.4.3 *FC reliability*

From table 8.5 it can be seen that over the period of one week (Time A - Time B), the accuracy of subjects' answers on the standard FC when issued first falls from 90.5% to 80%, this indicates that subjects' answers one week later are still accurate and useful. However, to obtain a better picture of the standard FCs reliability it was necessary to look at the exact agreement within each subject's answers. Using the data from table 8.6, it was found that on average subjects' answers on the standard FC over the period of a week were the same 70% of the time. Again this is somewhat less than we would have hoped for, however it should be noted that the intention is to use FCs with experienced users that have been working with the relevant package continually over a period of time; therefore the problem of a long time lag between using the package and completing the FC should be eliminated.

It is interesting to note from table 8.6 that the most common inconsistency in subjects' answers, was where subjects indicated at Time A that they had not used a command and then at Time B they then indicated that they had used that command. This was not a case of subjects forgetting that they used a command over time but rather that they falsely recalled or assumed using commands, i.e. "errors of commission" (Bartlett 1932).

8.4.4 *Using semantic descriptions of command function as memory prompts*

As was mentioned earlier, it was hypothesised, that the accuracy of FCs may be improved by including a semantic description of each command's function alongside its name, i.e. a combined FC. We can now discuss the results relating to this issue.

Contrary to our expectations it can be seen from table 8.7 that when the two instruments were compared in the same conditions (i.e. administered first at Time B), subjects' answers on the standard FC as to what commands they had used were accurate 80% of the time whilst subjects' answers on the combined FC were accurate 77.4% of the time; i.e. the combined FC actually performed slightly worse (although not significantly). We can now look at possible reasons for these results.

Firstly, it may have been the case that subjects using the combined FC were confused as to whether they were answering about using the command name or performing its description. Although it was hoped that each description would relate solely and

accurately to each command name, it became apparent that this was not the case; (this is despite the fact that the descriptions were piloted with three independent users).

Through the semi-structured interview it was found that some descriptions actually denoted different commands to some subjects, e.g. the description “change format/measurements of part of file” for subject 8 meant using the font sizes and font names in the “Font” menu and not the “Section” command as intended.

Another important explanation is that some subjects had actually performed the task that the semantic description referred to but they did not necessarily do it by using the relevant command; therefore instances in which subjects answered about the semantic description might reduce the accuracy of the combined FC. For example, on the combined FC a subject that had performed the task “finish editing a file” might answer “yes” to the command name “Close” even although they actually performed the task by clicking the “close box” with the mouse. However, when this was compared with the electronic data log there would be no record of the command “Close” being invoked and the answer on the combined FC would be marked as incorrect. As we shall see this is even more of a problem when we look at the results for the semantic FC.

Clearly however, the inclusion of the semantic description did not improve the accuracy of subjects’ answers and if anything it may have even confused subjects more and lowered the overall accuracy of their answers. It is possible that clearer instructions to subjects which specified that the combined FC is asking only about the command names and that the descriptions are merely an extra aid, might reduce the confusion; however, this would seem to defeat the purpose of the combined FC somewhat. It may also be the case that having the semantic descriptions on their own would avoid this confusion and be a better prompt (i.e. a semantic FC). We can now see if this is the case by comparing the standard FC with the semantic FC.

Table 8.9 shows that when the two different FCs are compared under the same conditions (i.e. administered first at Time A), subjects using the standard FC could remember using 90.5 % of commands whilst those using the semantic FC could remember using 64.3%. After conducting a t-test it was found that these differences were significant, i.e. the standard FC was significantly more accurate than the semantic FC; the t-test statistics are shown in appendix 8.15.

The most obvious reason for these results which was mentioned earlier, is that in many cases subjects were performing the tasks described on the semantic FC but not necessarily by selecting commands from the menus; in fact many of the tasks could be

done in this way, e.g. the same result of invoking the “Replace” command could be achieved by simply highlighting the word that you wanted to change with the mouse and then typing in the new word. Likewise the same effect as invoking the “Save” command could be achieved by clicking the close box and then the “yes” button in the save dialog box when prompted. In both of these cases the answers on the semantic FC would have been marked as incorrect using the data log (which records only command selection) even although they were in fact correct.

The problem, therefore, is in not having a proper validity check (i.e. measure) for the tasks that subjects are performing; it is difficult to see how this could be achieved without the researcher observing each subject individually and noting down every task performed. As a result the semantic FC will almost certainly have obtained a lower accuracy score than it otherwise deserved.

Another explanation for the above results that was touched on earlier, is that the semantic descriptions varied markedly in how good they were at denoting commands to subjects. This may have affected subjects’ answers in two ways. Firstly, based on the description subjects may have answered that they did not perform a task when they in fact had. Secondly, based on the description subjects may have answered that they did perform a task when in fact they had not. From the design of the word-processing course and the semi-structured interview we know that both of these to be true in at least some cases, e.g. many subjects indicated that they had performed the task “change format/measurements of part of file” even though the data log showed that the relevant command (i.e. “Section”) had not been invoked; from the semi-structured interview however, it was found that some subjects indicated that they thought this description referred to the font size and font style commands in the “Font” menu.

Despite the problems with the semantic FC of a validity check and designing appropriate descriptions, it may still be useful in other ways. The semantic FC could be useful for discovering tasks that it is known that subjects are performing without using a menu command i.e. frequently performed tasks like opening, closing and saving documents. Also, through an additional column that asked subjects how they actually performed tasks, it would be possible to discover frequently used short cuts or to identify ways in which users performance could be improved, i.e. a measure of “Experienced User Performance” (Jordan 1992). However, this additional column would involve considerably more effort on behalf of the subjects and would therefore run counter to one of the major benefits of FCs i.e. the speed and ease with which they

can be filled in. It is interesting to note that in this study the vast majority of subjects indicated that they much preferred completing the standard FC to the semantic FC.

8.5 Summary

To summarise, therefore, the results of this study indicate yet again that standard FCs provide highly accurate information about:

- The commands that subjects use.
- Subjects' knowledge of what these commands are for.
- Their need for these commands (to a lesser extent).

In addition it was found that semantic descriptions of command function do not improve on the already high accuracy of FCs.

It now seems logical to conduct a “real-life” study, that asks experienced computer users to complete FCs in order to demonstrate the usefulness of FCs in software evaluation.

Chapter 9

Using feature checklists to evaluate menu commands in a word-processing package

9.1 Introduction

Previous research on the use of FCs in HCI, has demonstrated that FCs can provide reliable and valuable information concerning users' knowledge, usage and need for features in different software packages. However, the majority of research conducted and described so far, has been laboratory based; as such the ecological validity of this research may be suspect.

It therefore, seems desirable (if not essential) to actually apply FCs to a "real-life" study with experienced, frequent users, rather than a laboratory study. By doing so we will be able to demonstrate the kind of information that FCs can yield and how this can be applied to the evaluation process in HCI.

In order to maximise the amount of information obtainable, it was decided to incorporate all five question columns on the FC, namely:

- "Existed?" - "did you know this command existed?"
- "Used?" - "have you ever used this command?"
- "How often?" - "how often do you use this command?"
- "What for?" - "do you know what this command does?"

- “Need?” - “how often do you have any need to use this command?”

The FC used in this study, listed commands by their name only (rather than a combination of command name and a semantic description of their function). Although this would not allow us to obtain information about users’ need for commands where they had not previously known the command function, it would avoid confusion as to what users were actually being asked about (study 5). In addition it also reduced the actual size of the FC; this is an important consideration in an intensive evaluation process where various instruments and measures are being used. The size (length) of the FC is also an important consideration if we bear in mind one of the main advantages of FCs proposed in chapter 2, i.e. their “cheapness” (in both time and effort) for the respondent and the investigator.

Although there was no practical method of validating subjects’ answers on the FC (due to the ecological validity of this study), the findings from earlier research suggest that it is reasonable to assume that subjects’ answers are in fact accurate and valid. Given this finding, we can use FCs to obtain information on:

- Possible bugs in the “overall” system.
- User performance and how this can be improved.

With respect to the former of these information types, it was hypothesised that FCs may be capable of identifying five different types of bugs that could exist in the word-processing package used in this study (i.e. “Microsoft WORD 5.0”). These were:

- Information Delivery - users know what the command does and judge it to be useful, however they haven’t yet discovered its existence.
- Reminding - users know that the command exists and what it does, yet express that their need to use it is greater than their actual usage (possibly because they don’t remember at the right moment).
- Guessability - users know that the command exists but don’t know what the command actually does (possibly because the command name does not clearly indicate its purpose/function).

- Memorability - users indicate that they have used the command at some point but can't remember what it actually does. It is possible to determine if this is due to infrequent use by looking at the information from the "How often?" column.
- Information Flood - users know that the command exists and know what it does, however they don't judge it to be useful. If this is the case there may be situations where users are distracted by too many features that they do not want.

Table 9.1, summarises these bugs and the FC answers that would indicate their existence:

Table 9.1: Summary of different bug types that FCs may detect and corresponding column answers

Bug Type	FC Column Answers
Information Delivery	("What for?"✓ + "Need?"✓) - "Existed?"✗
Reminding	("Existed?"✓ + "What for?"✓ + "Need?"✓) > "How often?"✗
Guessability	"Existed?"✓ - "What for?"✗
Memorability	"Used?"✓ - "What for?"✗
Information Flood	("Existed?"✓ + "What for?"✓) - "Need?"✗

In addition to providing information about the different types of bugs that may exist in the system, FCs may also be useful on an individual level with each user to improve their performance. For example it should be possible to identify useful commands of which individual users are unaware of the function. The following study was conducted in order to try and demonstrate how FCs could be used as a measurement instrument in a "real-life" HCI setting.

Since this study is similar to how we envisage FCs will be used in "real-life" HCI research, we are in a better position to examine the user (respondent) aspect of FCs. An additional aim of this study therefore, will be to look at the cost (in time and effort) to the user of completing a FC (chapter 2).

9.2 Method

Subjects: Eight subjects in total took part in the study (7 female and 1 male); their ages ranged from 23 to 44 years of age with a mean of 29.4 years. Although this may seem a small sample it should be remembered that this is a survey study and not a comparison between experimental groups; we also have no reason to believe that these expert users would differ from other expert users. Subjects were postgraduate students recruited from the psychology department at the University of Glasgow. All of the subjects were experienced and frequent users of the word-processing package “Microsoft WORD 5.0” on “Apple Macintosh” computers; this was assessed by asking subjects to complete a computer use questionnaire, (appendix 9.1).

Design/Procedure: All of the subjects were asked to complete a computer use questionnaire (appendix 9.1), which asked them about the type of computer and word-processing package that they used most frequently, and how often they used it. In addition the questionnaire asked subjects if they would be willing to complete a more detailed questionnaire about word-processing for which they would be paid three pounds (£3). Subjects were selected on the basis that the word-processing package that they used most frequently was “Microsoft WORD 5.0”, and that they used this package at least every day.

The subjects were then issued with a sealed envelope which contained a “Microsoft WORD 5.0” FC (appendix 9.2), and were told that it contained a questionnaire about word-processing along with instructions. Their task was to complete the questionnaire at home that same evening (i.e. away from the computer). Subjects were informed not to look at the questionnaire beforehand and that it was important that they completed the questionnaire in one session (i.e. not to do half one day and the other half later). Finally, subjects were told that it was important that they answered the questions as truthfully as possible. All FCs were returned to the experimenter the following morning.

Once all FCs had been returned, the experimenter checked the “Microsoft WORD 5.0” menus on the computers that each subject used. If any of the commands on the FC did not appear on the menus used by subjects then these were deleted from the FC (along with any answers).

Finally, all subjects were posted a FC assessment form (appendix 9.3), that asked various questions about completing a FC. After completing this form, subjects returned it to the experimenter in the enclosed, self-addressed envelope.

9.3 Results

The following section is split into three parts. The first part deals with a general analysis of subjects' answers grouped together in order to try and identify the possible bugs mentioned earlier. The second part looks at each subject's individual answers in order to give feedback to subjects about possible ways in which their performance may be improved. Finally, the third part looks at the cost to the user of completing a FC (using the FC assessment form).

9.3.1 *System bug detection*

Appendix 9.4 shows the number of subjects that:

- Knew whether each command existed ("Existed?").
- Had ever used each command ("Used?").
- Knew what each command does ("What for?").

The maximum score for each command was 8 for each answer (i.e. 8 subjects). In addition, appendix 9.4 shows the average frequency of usage for each command across subjects ("How often?") and the average need for each command across subjects ("Need?").

From an analysis of the results shown in appendix 9.4, it is possible to identify which of the bugs (hypothesised earlier), actually exist; these are summarised in table 9.2.

Table 9.2: Summary of overall system bugs detected

Bug type	Number of cases detected
Information delivery	0
Reminding	4
Guessability	11
Memorability	0
Information Flood	24

The actual cases detected were as follows:

- Reminding - 4 cases detected, i.e. there were four cases in which users knew that the command existed, and what it did, yet expressed their need to use it was greater than their actual usage. These commands were: “Page Setup”; “Normal”; “Hide/Show”; and “Style”.
- Guessability - 11 cases detected, i.e. there were eleven cases in which users knew that the command existed but didn’t know what it did. These commands were: “Print Merge”; “Go To”; “Outline”; “Paragraph”; “Section”; “Border”; “Renumber”; “Sort”; “Repaginate Now”; “Preferences”; and “Commands”.
- Information Flood - 24 cases were detected, i.e. there were twenty-four cases in which users knew that the command existed and what it was for, yet expressed little need for using it. These were: “Clear”; “Replace”; “Header”; “Table”; “File”; “Character”; “Document”; “9 Point”; “14 Point”; “18 Point”; “24 Point”; “Chicago”; “Courier”; “Helvetica”; “Monaco”; “New York”; “Symbol”; “Grammar”; “Thesaurus”; “Word Count”; “Help”; “Show Clipboard”; “New Window”; and “Untitled 1”.

9.3.2 Individual user performance

The individual answers of each subject on the FC for each command is shown in appendix 9.5. However, tables 9.2 - 9.9, show the numbers for each possible bug type that actually occurred for each user.

Table 9.2: Subject 1 - bug detection summary

Bug Type	No .	Commands
Information Delivery	0	
Reminding	0	
Guessability	7	Print Merge; Date; Document; Sort: Repaginate Now; Preferences; Commands
Memorability	0	
Information Flood	20	Clear; Outline; Ruler; Hide/Show; Table; File; 14 Point; 18 Point; 24 Point; Chicago; Helvetica; Monaco; New York; Symbol; Times; Grammar; Renumber; Help; Show Clipboard; New Window

Table 9.3: Subject 2 - bug detection summary

Bug Type	No .	Commands
Information Delivery	0	
Reminding	1	Clear
Guessability	7	Print Merge; Replace; Hide/Show; Table; Section; Renumber; Sort
Memorability	0	
Information Flood	30	Page Setup; Find; Outline; Page Layout; Ruler; Header; Footer; Page Break; Paragraph; Section; Border; Style; Italic; Underline; 9 Point; 14 Point; 18 Point; 24 Point; Courier; Helvetica; New York; Palatino; Symbol; Word Count; Repaginate Now; Preferences, Commands; Help; Show Clipboard; New Window

Table 9.4: Subject 3 - bug detection summary

Bug Type	No.	Commands
Information Delivery	0	
Reminding	0	
Guessability	15	Print Merge; Clear; Replace; Go To; Glossary; Normal; Outline; Page layout; Hide/Show; Border; Style; Renumber; Sort; Repaginate Now; Commands
Memorability	0	
Information Flood	16	Date; File; Chicago; Courier; Geneva; Helvetica; Monaco; New York; Palatino; Symbol; Grammar; Thesaurus; Preferences; Help; New Window; Untitled 1

Table 9.5: Subject 4 - bug detection summary

Bug Type	No.	Commands
Information Delivery	1	Select All
Reminding	1	Find
Guessability	8	Go To; Glossary; Normal; Ribbon; File; Border; Thesaurus; Commands
Memorability	1	Border
Information Flood	26	Print Preview; Print Merge; Undo/Redo; Replace; Hide/Show; Header; Footer; Character; Paragraph; Table Cells; 9 Point; 18 Point; 24 Point; Chicago; Courier; Geneva; Helvetica; Monaco; New York; Symbol; Times; Grammar; Word Count; Help; Show Clipboard; Untitled 1

Table 9.6: Subject 5 - bug detection summary

Bug Type	No .	Commands
Information Delivery	0	
Reminding	1	Repaginate Now
Guessability	3	Sort; Preferences; Commands
Memorability	0	
Information Flood	40	Undo/Redo; Clear; Replace; Go To; Glossary; Normal; Outline; Page Layout; Ribbon; Ruler; Page Break; Table; Date; Picture; Character; Paragraph; Section; Document; Border; Table Cells; Table Layout; 9 Point; 10 Point; 18 Point; 24 Point; Chicago; Courier; Geneva; Helvetica; Monaco; New York; Symbol; Spelling; Grammar; Thesaurus; Word Count; Renumber; Help; Show Clipboard; Untitled 1

Table 9.7: Subject 6 - bug detection summary

Bug Type	No .	Commands
Information Delivery	1	Glossary
Reminding	3	Undo/Redo; Document; Times;
Guessability	5	Print Merge; Outline; Table; File; Table Layout;
Memorability	0	
Information Flood	11	Character; Border; 14 Point; 18 Point; 24 Point; Chicago; Palatino; Symbol; Thesaurus; Sort; Preferences;

Table 9.8: Subject 7 - bug detection summary

Bug Type	No .	Commands
Information Delivery	0	
Reminding	0	
Guessability	2	Print Merge; Repaginate Now
Memorability	0	
Information Flood	20	Clear; Go To; Glossary; Date; File; Character; Paragraph; Section; Border; Style; Monaco; Symbol; Grammar; Thesaurus; Renumber; Preferences; Commands; Help; Show Clipboard; New Window

Table 9.9: Subject 8 - bug detection summary

Bug Type	No .	Commands
Information Delivery	0	
Reminding	0	
Guessability	5	Print Merge; Page Break; Paragraph; Section; Repaginate Now
Memorability	0	
Information Flood	16	File; Picture; Document; Border; Style; 9 Point; 14 Point; 18 Point; 24 Point; Helvetica; Monaco; Palatino; Symbol; Times; Grammar; Renumber

From tables 9.2 - 9.9, it can be seen that the most common bug type that was identified for each individual user was information flood; i.e. cases in which users were aware of a command's existence and what it was for, yet felt that they had very little or no need to ever use it.

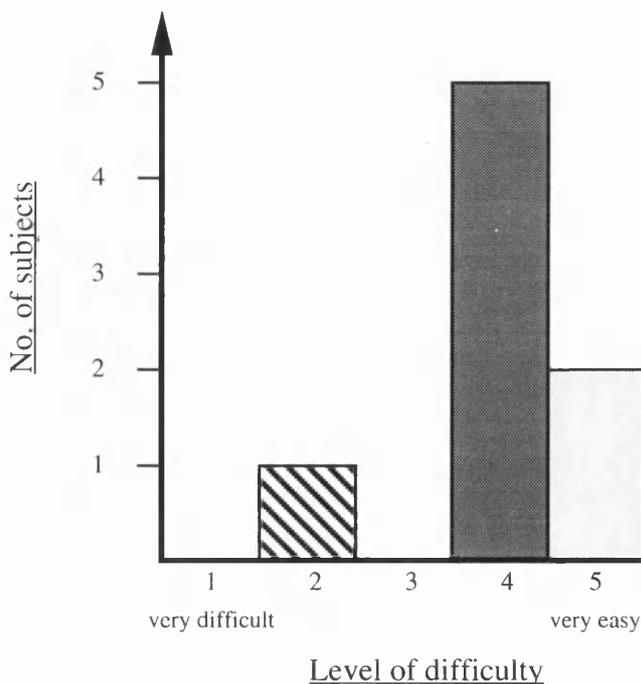
Guessability of commands was also a common problem for many users, i.e. knowing that the command existed but not knowing what it actually does. From the results reported it can be seen that some cases of this type of bug, existed for every user; this ranged from 2 commands (subject 7) to 15 commands (subject 3).

The other types of bugs, were much less common, in particular memorability in which only one user (Subject 4), indicated that they had used a command before but yet could not remember what it did. There were only 2 cases (subjects 4 and 6), in which subjects' answers indicated the possible problem of information delivery, i.e. users know what the command does and judge it to be useful, but haven't yet discovered its existence; these were "Select All" (subject 4) and "Glossary" (subject 6). Finally, the bug type "reminding", was identified as a possible problem for 4 subjects, i.e. knowing that the command existed and what it does and judging it to be very useful, but yet very rarely (or never) using it.

9.3.3 *Feature checklist user cost*

The user assessment form asked three questions about completing a FC as well as a space for any additional comments that users wished to make. The first question asked subjects to rate how easy/difficult it was to fill in a FC on a 5-point scale. From figure 9.1 it can be seen that the majority of subjects rated the FC as level 4 on the scale i.e. between neutral and very easy to fill in. The actual mean score of all subjects answers was 4.00.

Figure 9.1: Level of difficulty experienced by subjects completing a feature checklist



Question 2 asked subjects to arrange the FC column headings (questions) in order of difficulty; 1 = “easiest to answer”, and 5 = “hardest to answer”. The mean score for each FC column heading is shown below in table 9.10. From this it can be seen that the “Existed?” column was the easiest to answer followed by “Used?”, “What for?” and “How often?” columns respectively. The “Need?” column was rated the hardest column to answer.

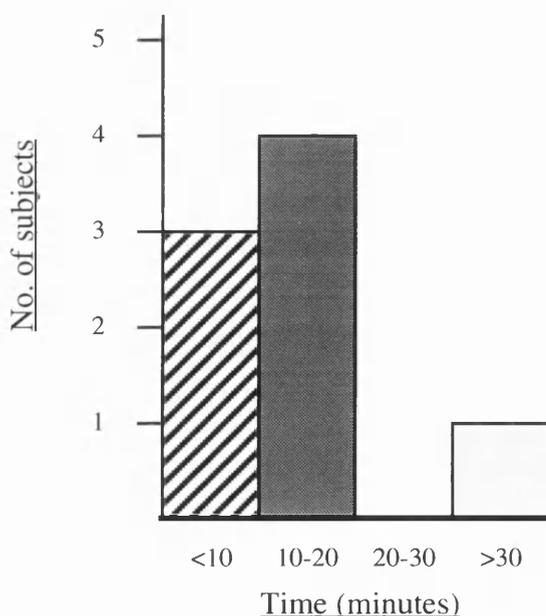
Table 9.10: Subject rating (mean score) of FC columns on ease/difficulty to answer

FC column	Rating	Mean score (level of difficulty to answer)
“Existed?”	1	1.38
“Used?”	2	1.75
“How often?”	4	3.13
“What for?”	3	2.75
“Need?”	5	3.60

N.B. 1 = easiest to answer; 5 = hardest to answer

Question 3 asked subjects how long it took them to fill in a FC. From figure 9.2 it can be seen that the majority of subjects (7 out of 8) took less than 20 minutes to fill in the FC used in this study (3 of these took less than 10 minutes).

Figure 9.2: Length of time taken by subjects to fill in the FC



9.4 Discussion

Yet again this section is split into three parts; the first of these discusses the role of FCs in system bug detection, the second discusses the role of FCs in identifying possible ways in which user performance can be improved, and the third discusses the cost to the user of completing a FC.

9.4.1 *Using FCs for system bug detection*

When FCs are used to evaluate the menu commands in a system across a group of experienced, frequent users, they can provide useful information that highlights suggestions for future designs. From the results reported, it can be seen that some command names may not be appropriate for conveying their function/purpose to users (i.e. they may have a guessability problem). Although this is often related to the fact that the commands have been very rarely used, it still highlights a possible area for design improvement; this is particularly true for the command “Print Merge” (all users knew it existed but only 2 claimed they knew what it actually did).

FCs also seem to be useful in identifying redundant commands in a system, i.e. ones which users express little or no need for using, even although they know what they actually do. If there are a large number of commands that come into this category then this could lead to an information flood problem, i.e. too many commands distracting users. This is quite often the case with font styles that users feel they will never (or very rarely) use, e.g. “Symbol”, in this study. Of course this problem can be alleviated by allowing users to customise menus to their own preference; ironically however, many users are unable to do this because the command needed to perform this task (i.e. “Command”) has very low guessability (few users know what it does). If, after conducting a large scale study using many users, it is discovered that there are still commands that few users express a need to use, then it may be better to omit these commands from the system altogether.

The only other type of system bug that was identified as existing in the present study was reminding, i.e. cases in which users knew that the command existed, and what it did, yet expressed their need to use it was greater than their actual usage. It may be that users actually over-estimated their need for using these commands on the FC; however,

it could also be the case that users fail to remember that these commands exist when they really need to use them. Follow-up interviews with users may reveal possible reasons for this e.g. commands being located on unexpected menus.

There were no overall cases of either of the other two types of bugs existing in this study, i.e. information delivery and memorability.

9.4.2 *Using FCs to improve user performance*

It was seen from tables 9.2 - 9.9, that the most common bug type that was identified for each individual user was information flood; i.e. cases in which users were aware of a command's existence and what it was for, yet felt that they had very little or no need to ever use it. Although many of these commands were font styles and sizes (reflecting each individual's own personal preference), it should be noted that there were still many other commands that users felt they would never need to use. Yet again it is ironic that the method for alleviating this problem (i.e. using the "Commands", command to remove commands from menus) had not been discovered by half of the users. An interesting point worth mentioning here is that 6 out of the 8 users felt that they did not have much need for the "Help" command. Yet again follow-up interviews may reveal whether this is because users feel the command is not very helpful or whether they feel they have a good knowledge of the system.

Guessability of commands was also a common problem for many users, i.e. knowing that the command existed but not knowing what it actually does. Although it was nearly always true that this was the case for commands that users had never used, it again indicates that the actual command name does not convey its function to many users; for example 6 out of the 8 users did not know what the "Print Merge" command did, even although most knew it existed.

The other types of bugs, were much less common, however it is still interesting that there was one case in which a user indicated that they had used a command before but yet could not remember what it did, (of course this could be a problem with understanding command function rather than poor memorability, i.e. although they had selected that command before they still might not know what it does).

In order to demonstrate how this information could be used to improve or aid individual user performance we can select an actual case as an example. If we look at the results

from subject 6 (table 9.7 and appendix 9.4), it can be seen that it would be useful to inform this user that:

- The “Glossary” command does exist in the system and that it is located in the Edit menu.
- The function of the “Print Merge”; “Outline”; “Table”; “File”; and “Table Layout” and demonstrate how to use them.
- It may be helpful for them to remove certain commands that they will probably never use, e.g. font commands such as “Chicago”, “Palatino”, “Symbol”, etc.

It is important to point out however that all these recommendations should be discussed with the user so that the problems identified are real, e.g. users may say they have no need to use a command (information flood) because they have a false understanding of its actual purpose.

9.4.3 *Completing a feature checklist: the cost to the user*

In chapter 2 of this thesis, we argued that one of the probable advantages of FCs over alternative methods such as questionnaires and semi-structured interviews was that FCs would be a simple and quick method for the user to complete. It was seen from figure 9.1 that only one user found the FC difficult to complete; the remaining seven users found the FC either easy (4 users) or very easy (3 users) to complete. This lends considerable support to our view that the FC is a simple method for users to answer.

Individual column difficulty

As far as the individual columns on the FC are concerned, there was general consensus that the two easiest columns to answer were “Existed?” and “Used?”, and the two hardest columns were “How often?” and “Need?” These results are not unexpected since it is reasonable to assume that the “How often?” and “Need?” column require additional information than that required to answer the “Existed?” and “Used?” columns, e.g. subjects must first of all try and recall any occasion when they might have used a command (“Used?”), only after this, can they then start recalling particular occasions and situations when that command was used (“How often?”). Similarly, it is

only after recalling what a command does (“What for?”), that subjects can then estimate their need for a command by thinking of examples where they might use it (“Need?”).

Although subjects found the “How Often?” and “Need?” commands to be the hardest to answer, it should be remembered that, overall subjects found the FC easy to complete; it is perhaps more appropriate to talk of these columns as being “less easy” rather than “harder” than the others.

“Need?” column data

The user assessment form also contained space at the end where subjects were asked to write any additional comments about the FC. Although only 2 subjects made additional comments it is worth noting that both of these referred to the difficulty of the “Need?” column. When this information is considered along with that from table 9.10 it can be seen that subjects have more difficulty answering this column than any other. If subsequent research confirms this finding, then the benefits of having this column need to be weighed against the difficulty involved in answering it before a decision can be made on whether or not to include this column on the FC.

The data from the “Need?” column is used to detect reminding, information delivery and information flood bugs (see table 9.1). The question that arises is: If we exclude the “Need?” column from The FC can we still detect these bugs? If subjects answer that they know that a command exists and what it is for yet never or very seldom use it, then this could mean either: (i) they have no need for this command (information flood) or, (ii) they don’t remember to use it at the right moment (reminding). It would still be possible to determine which of these was correct through another method (e.g. a semi-structured interview). As far as information delivery is concerned, it should still be possible to detect this bug without the data from the “Need?” column, i.e. if subjects answer that they know what the command does but didn’t know it existed; (N.B. it should also be noted that this type of bug is expected to be very rare and did not occur in this study).

As we have already mentioned the benefits of having the “Need?” column must be weighed against the difficulty involved in answering it before a decision can be made on whether or not to include this column on the FC.

Time taken to complete a FC

With regard to the cost in time to the user, it was found that the vast majority of subjects in this study took between 10 and 20 minutes to complete the FC (5 out of 8). It is reasonable to expect that this is considerably less than the time required to complete alternative instruments that could be used to collect the same information i.e. a questionnaire or a semi-structured interview that asked about users' usage, knowledge and need for each of the 80 features contained in a system.

These results would seem to confirm the view expressed in chapter 2, namely: that FCs are a cheap instrument in terms of time and effort for a user to complete.

9.5 Summary

This study has shown how FCs can be used in “real-life” evaluation in HCI. In particular, FCs seem to be a useful method for identifying different kinds of bugs in a system as well as indicating ways in which individual performance with that system could be improved. This information can be obtained at relatively little cost to the user, especially when compared to alternative instruments.

Chapter 10

Applying feature checklists in an industrial setting: monitoring usage of a “new” system.

10.1 Introduction

So far we have demonstrated that FCs can provide accurate information about the features in an interface that users have used, and their knowledge of these features. We have also demonstrated:

- How information obtained from FCs can be used.
- That FCs are a cheap instrument to employ (both for the user and the investigator).

Having said this however, it would be useful to investigate FCs when they are applied in “real-life” industrial settings. This study is concerned with such an investigation.

10.1.1 *EasyReader: an electronic documentation system*

The interface that this study used FCs to investigate was an electronic documentation system introduced by the Dutch Oil Company (NAM). The system is known as *EasyReader*.

EasyReader is basically a collection of engineering documents that have been produced in an electronic format and issued on a CD; the system offers a number of advantages over traditional paper systems in the control and management of documentation such as ensuring the recency of documents and accessibility of information.

Three months after the system was introduced into the work environment, it was decided to conduct a monitoring programme to assess a number of aspects of *EasyReader* including users' usage of the system and the features it contains.

This study describes the monitoring programme that was implemented, in particular, the use of FCs in this programme. The major aim of this study is to obtain feedback from users about their opinions of completing the FC (including a comparison with a "traditional" questionnaire).

10.2 Method

The entire monitoring programme is described in this section to provide context, however, our main interest is with the stage of the programme that employs FCs (stage 4).

Design/Procedure: The monitoring programme employed a variety of methods over a period of 9 weeks. The programme itself was split into 5 distinct stages:

- (1) To ensure that the monitoring programme produced relevant results it was important that it measured performance on a number of requirements specified by NAM. After examining these requirements it was possible to specify two overall objectives namely:

- To assess the control of documentation.
- To assess the accessibility of information.

In order to assess performance against these high-level objectives it was necessary to identify relevant aspects of *EasyReader* that could be usefully measured. In total 8 aspects of the system were identified; these were:

- System usability.
- Overall system usage.
- System bugs/error rates.
- Training.

- User performance/feature usage.
- User attitudes.
- Interface aspects.
- Document management process.

Each of these aspects can be measured by a variety of methods, some of which are more useful than others. We decided to employ 3 methods, namely: questionnaires, a semi-structured interview and a FC. The FC was used to measure system bugs/error rates and user performance/feature usage, whilst the questionnaires and the semi-structured interview were used to measure the remaining aspects.

- (2) The second stage of the programme was to send a letter to all potential *EasyReader* users informing them that a monitoring programme was about to begin and that this would obtain feedback from them about *EasyReader*. This letter outlined the monitoring plan and explained that the information obtained would be used to improve the *EasyReader* system.
- (3) The third stage of the programme was based on a questionnaire that was sent out to all prospective *EasyReader* users (appendix 10.1); a total of 175 questionnaires were issued with a return deadline of 2 weeks. Since the aim of this questionnaire was to obtain information on a wide variety of topics relating to *EasyReader*, it contained a total of 40 questions.
- (4) The next stage of the monitoring programme involved visiting a representative sample of *EasyReader* users at both well site and office locations. During this visit the monitor conducted a semi-structured interview with these users (appendix 10.2); this interview explored in more detail, important aspects identified in the questionnaire. In addition, an *EasyReader* feature checklist was issued to all 24 users that were interviewed (appendix 10.3). This feature checklist obtained information about users usage, knowledge and need for features contained in the *EasyReader* system (i.e. menu commands and icons). Users were asked to complete this FC in their own time and return it in the enclosed, addressed envelope by internal mail. The FC also included an “information sheet” (appendix 10.4) that asked users some questions about the FC.

(5) In the final stage of the monitoring programme, a second questionnaire (appendix 10.5) was issued to all prospective *EasyReader* users; a total of 136 of these second questionnaires were issued (N.B. this number is lower than the first questionnaire because it was discovered that one business unit did not have access to *EasyReader* and also some office personnel previously thought to use *EasyReader*, did not). This second questionnaire was significantly shorter than the original questionnaire since it was intended to focus on specific aspects of *EasyReader* identified from the previous monitoring methods. When this second questionnaire was sent out it had an “feedback sheet” attached to it which gave a brief description to users of some preliminary findings from the first questionnaire.

10.3 Results

This section will only discuss the results from the monitoring programme that are relevant to the FC.

Of the 24 FCs issued, 15 were returned; this represents a return rate of 62.5%. This return rate is encouraging especially considering:

- That users had already completed and returned a questionnaire.
- The diverse location of users and the working environment (users on the well site would have to have completed the FC in their own leisure time).

The “information sheet” asked 3 questions about completing the FC. Questions 1 and 2 asked users to rate how difficult it was to complete both the FC and the initial questionnaire respectively, using a 5-point scale. The results are shown below in table 10.1.

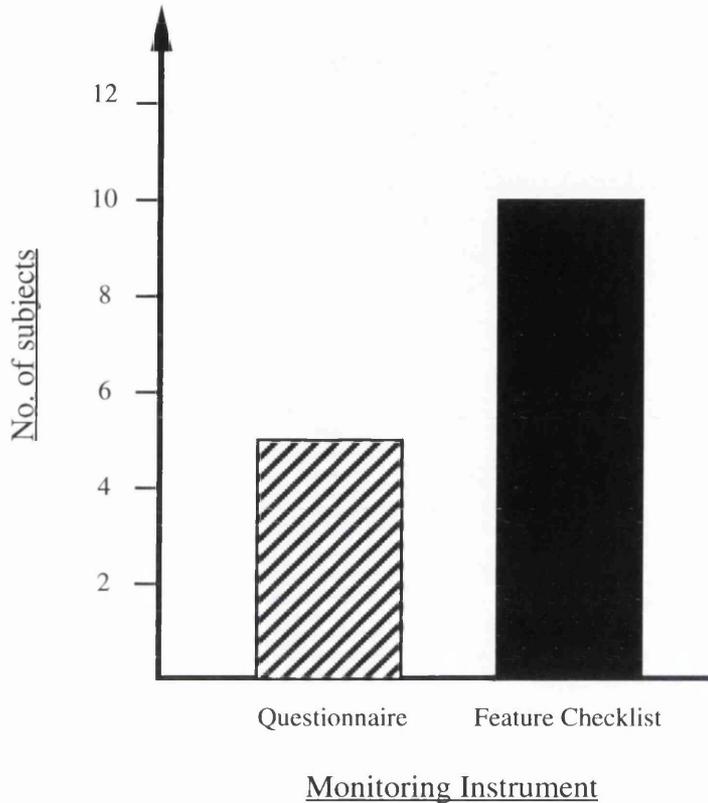
Table 10.1: Users’ rating of difficulty for completing both monitoring instruments

Monitoring Instrument	Difficulty Rating (mean)
Feature Checklist	3.1
Initial Questionnaire	3.5

N.B. 1 = very difficult; 5 = not at all difficult

Question 3 asked users to indicate which of the two instruments took longer to complete. The results are shown below in figure 10.1.

Figure 10.1: Users' answers to the question "Which instrument took longer to complete"



From figure 10.1 it can be seen that 5 users answered that the initial questionnaire took longer to complete whilst 10 users thought that the FC took longer to complete.

10.4 Discussion

It can be seen that contrary to our expectations, the FC was rated as slightly harder to complete than the questionnaire, although not significantly. In addition, the majority of users claimed that the FC took longer to complete than the questionnaire.

In discussing these results however, it is essential to mention an important aspect of the monitoring programme that emerged. Contrary to NAM's expectations it was found that *EasyReader* was in general poorly used; many employees had used the system only

rarely and had not “explored” or used many of the features that *EasyReader* offered. The monitoring programme identified a number of reasons why this was the case:

- Due to unexpected IT problems many prospective users did not have *EasyReader* installed until very recently; there were 2 major implications of this:
 - Users had only recently started to use the system and have therefore not had time to explore it’s functionality.
 - There was a long delay between users attending an *EasyReader* training course and being able to practice their newly acquired skills.
- A significant number of users did not attend an *EasyReader* training course (approximately 24%).

It was clear from the monitoring programme that the majority of respondents were not experienced *EasyReader* users. Since it has been stated on a number of occasions throughout this thesis that FCs should be employed with experienced users (when used for bug detection), the results found in section 10.4 may not be as surprising as first thought. It should also be mentioned that even when FCs are used to measure user performance it is important that users have used the system on at least several occasions (see section 2.4.1); it was clear in this study that this was not always the case. In this study, the FC was used for both bug detection and measuring user performance.

In order to examine the results more fairly it was necessary to categorise users by their level of experience in using *EasyReader*. This was done by looking at users’ responses to the initial questionnaire. The 15 users that completed a FC were categorised using their answers to questions 5 and 8 on the initial questionnaire:

5. How often do you normally use *EasyReader*?

- Every day.
- 2-3 times per week.
- Once a week.
- At least once a month.
- Less than once a month.
- Never.

8. Have you ever attempted to perform any of the following tasks?

- Open *EasyReader* .
- Use the table of contents.
- Use the go back button.
- Use the table icons.
- Use the link icons.
- View drawings/illustrations.
- Print from *EasyReader* .
- Exit from *EasyReader* .
- Create and use annotations.
- Use keywords to find documents.

For the purposes of this study, subjects were classified as experienced users if they had used *EasyReader* at least once a week and had performed the majority of the tasks listed (i.e. a minimum of 6 out of 10). Using this classification system we classified 7 of the users that completed a FC as “experienced” and 8 as “inexperienced”.

Having categorised users on their level of experience with *EasyReader* we can now re-examine the results to the “information sheet”.

Table 10.2: Mean difficulty rating for both instruments by level of users' experience

Users' experience level	Feature Checklist	Questionnaire
Experienced	4.1	4.0
Not experienced	2.1	3.0

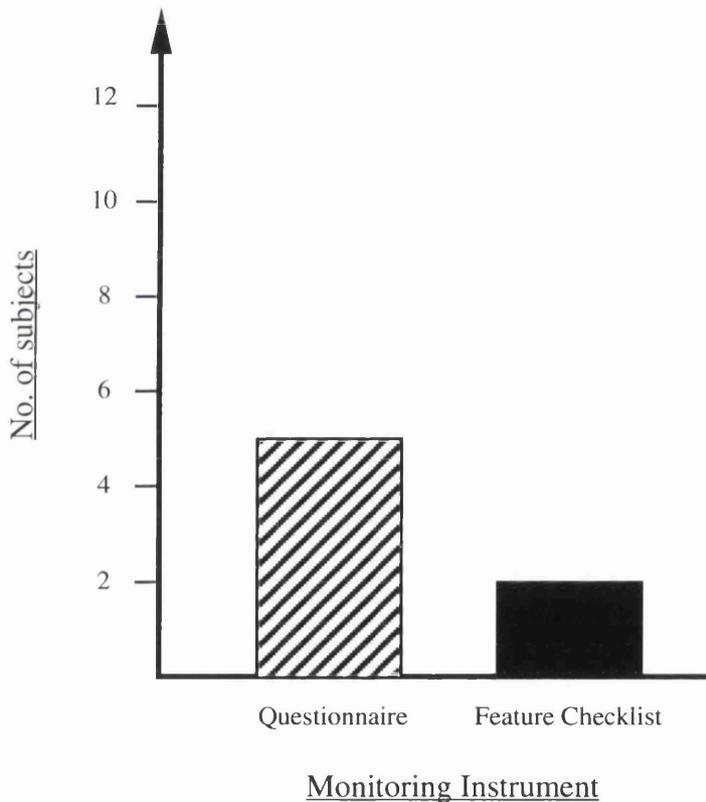
N.B. 1 = very difficult; 5 = not at all difficult

When we look at the FC results with experienced users (i.e. intended users) it can be seen that it is rated slightly easier to complete than the questionnaire, although not significantly. It should be noted that both the FC and the questionnaire were rated as relatively easy to complete. More importantly however, it can be seen that there is a greater difference between inexperienced users on difficulty rating (i.e. the FC is rated more difficult to complete). After conducting a t-test for independent samples (appendix 10.6) it was found that our obtained t-value of 2.198 exceeds the t-value of

2.145 at the 5% level for 14 degrees of freedom (one-tailed). We can therefore reject the null hypothesis and conclude that when both instruments are employed with inexperienced users, the FC is rated significantly harder to complete than the questionnaire.

The second aspect of the information sheet that we can look at, concerns the time taken by subjects to complete the FC. When we look at the data from experienced users only (figure 10.2), we can see that 5 users thought that the questionnaire took longer to complete whilst only 2 thought the FC took longer; (N.B. all 8 inexperienced users thought that the FC took longer to complete).

Figure 10.2: Experienced users' answers to the question "Which instrument took longer to complete?"



When we categorise users by their experience with using *EasyReader*, then the data clearly supports the view that FCs should be used with experienced users (or users that have used the system on at least several occasions). Inexperienced users found that the FC was harder and took longer to complete than the questionnaire, i.e. the proposed

advantages of FCs are not supported with this type of user. (N.B. the only situation where we have suggested FCs can be used with inexperienced users, is in evaluating the effectiveness of training courses i.e. what commands have users learned to use? In this situation it is likely that alternative methods such as interviews and tests would also be time consuming for both the trainer and trainee).

An interesting aspect that is suggested by these results is that inexperienced users did not simply work through the FC placing crosses to indicate that they hadn't used commands, otherwise the FC would have been completed very quickly. It seems likely that both experienced and inexperienced users completed the FC in a thoughtful and accurate manner.

Since only 7 FCs were completed by experienced users, we were not in a position to make any general recommendations about *EasyReader* based on the FC data. One of the most important findings from the monitoring programme as far as the oil company were concerned, was that at present the uptake of *EasyReader* is disappointing; the FC helped to demonstrate this point. A number of reasons for the poor uptake were identified and recommendations made on how to improve this.

10.4.1 *Cost to the investigator*

An important advantage of FCs that was proposed in chapter 2 is that they can be designed cheaply and quickly (i.e. at little cost to the investigator), so far this has not been discussed. Although no detailed, objective data exists on this topic, it is useful to report the evidence (although anecdotal) in this study.

The features of *EasyReader* (i.e. icons, menu commands, etc.) had previously been used to produce training material and they therefore existed in a convenient electronic format. Consequently, it took the investigator a maximum of 2 hours to completely design the FC employed in this study, using a word-processing application ("Microsoft WORD 5.0"). In contrast, the questionnaires took considerably longer and went through several iterations, this is especially true for the initial questionnaire. Since the client in this study was based in the The Netherlands, each design of the questionnaire had to be faxed to the client where it was then analysed and faxed back with comments. The whole process of designing the questionnaire took place over the period of 10 days;

it should be noted however, that other activities took place during this time (i.e. designing the initial questionnaire did not take 10 whole days of constant effort).

The situation in which this study took place is often typical of “real-life” industrial research in HCI; it therefore indicates that FCs are a very cost-effective method for the researcher to employ, especially when compared to other instruments.

10.4.2 *Implications for future research*

As a result of this study a number of possibilities for future research have emerged:

- The importance of instructions on how to complete a FC should be investigated. There are 2 main ways that subjects can complete a FC:
 - by moving down each column one at a time (i.e. answer for each command in the “Existed?” column before moving on to “Used?” column).
 - by moving across rows one at a time (i.e. answer each column for one command before moving to next command).

The effect of these different instructions on user difficulty and time taken to complete the FC should be examined.

- We have stated that FCs should be employed with experienced users of a system (for bug detection) however, the actual level of experience necessary should be investigated more fully.
- The return rate for FCs in this study was 62.5%; however, because we did not anticipate the lack of subject inexperience we were not able to use the FC data in a constructive manner. Future research should investigate the return rates for FCs in different environments and with a variety of systems.
- A quantitative investigation should be conducted which compares the costs involved to the investigator in designing a FC compared with the costs involved in designing an alternative method e.g. a questionnaire.

10.5 Summary

To summarise therefore, this study has indicated that when FCs are used for bug detection they should be employed with experienced users of the system under investigation. Where this is the case, FCs are a simple and quick method for obtaining valuable data in “real-life” settings.

Chapter 11

Using feature checklists in HCI

11.1 Feature checklists: a critical review

The research that has been conducted in this thesis has demonstrated that feature checklists are an accurate and useful measurement/evaluation instrument for HCI. The studies described have followed a logical progression in the development and assessment of the instrument itself, leading to actual examples of practical applications of the method. We are now in a position to what this thesis has and has not achieved with respect to FCs.

11.1.1 Accuracy of feature usage data

Since the primary aim of FCs is to measure feature usage within a system, it is important to repeat the findings of the first three studies with respect to the “Used?” column. When the FC is presented first, the accuracy of subjects’ responses varied between 80.3% and 97.1% for textual features such as menu commands (mean score across all 3 studies was 87.9%), and varied between 82.2% and 87.9% for pictorial features such as icons (mean score across all 3 studies = 85.9%). FCs are therefore an accurate measurement instrument that reliably score over 80% when measuring what features subjects have used. It is reasonable to believe that if subjects had been presented with commands, icons etc. individually, their accuracy in identifying which features they had used would have been around 60%, however, since FCs give important contextual cues this accuracy is increased to around 86%.

Although these results are encouraging, the important question that arises is “how accurate do FCs need to be, to be useful to design engineers?” This question however, applies equally to any of the measurement methods that were discussed in chapter 2. In our opinion the most appropriate way of answering this question is to make design changes based on the data obtained from these methods and then assess the impact of these changes on users’ interaction i.e. errors made, level of performance, speed of learning, satisfaction, etc.

An important advantage that FCs have is that they can be used conveniently with large numbers of users, as a result the information that they obtain is likely to be an accurate and reliable assessment of the system under investigation (especially compared with methods that obtain detailed data from a small group of users). Throughout this thesis we have proposed that the data obtained from FCs should be backed-up with additional information e.g. from semi-structured interviews; this is true regardless of whether FCs are used for bug detection or improving user performance. Implementation of this procedure should also serve to ensure the accuracy of FC data and the usefulness of resulting suggestions for improved design.

An aspect of FC accuracy that was mentioned in section 5.4, was that there may be a “ceiling effect” with FCs, beyond which their accuracy cannot be improved. Although this may be the case, the results from the studies in this thesis suggest that this “ceiling” is likely to vary depending on the system being investigated and the type of user interaction involved i.e. FC accuracy ranged from 80.2% to 97.1% across the first 3 studies.

Although this thesis has not explored the problem of “ceiling effects” in detail, it is our opinion based on the results reported, that when every effort is made to ensure visual realism in FC design, e.g. by on-screen FCs (discussed later), then ceiling effects will not be a major problem, i.e. this “ceiling” is likely to be “high”.

In addition to showing that FCs are an accurate instrument for measuring what features subjects have used, it was also shown in studies 1 and 2 that they are an accurate instrument for measuring relative frequency of usage. The results to the column headed “How often?”, show that subjects’ estimates on the FC of how often they used a feature, increase as actual usage increases.

11.1.2 Accuracy of feature knowledge data

As far as the other aim of FCs is concerned, i.e. measuring users' knowledge of commands, studies 4 and 5 indicated that subjects' responses to these columns were also accurate; this is especially true for the "Existed?" and "What for?" columns (the accuracy of subjects' responses to these columns varied from 74% to 87%). Again the data from these columns should be backed-up with data from other methods to ensure the accuracy and reliability of this information. In study 6 (chapter 9), the information resulting from these columns was discussed in terms of its usefulness in detecting system bugs and in measuring and improving levels of user performance.

Although the accuracy of subjects' estimates of how often they might need to use a command were slightly lower (74% and 64% respectively) it should be remembered that these estimates were based on opinions rather than subjects' memory. The possibility of excluding this column was discussed in chapter 10. It is possible that answering this question might require considerable effort from the user and thereby conflict with one of the main advantages of FCs. Since FCs should be used in conjunction with other instruments, it may be possible to obtain the same information from the "Need?" column elsewhere (if necessary). The issue of the costs involved in employing FCs is discussed more fully below.

11.1.3 The cost of employing FCs

The results from chapter 9 suggest that FCs can be completed easily and quickly by subjects whilst at the same time obtaining useful information. In this study 7 out of 8 subjects took less than 20 minutes to complete a "Microsoft WORD 5.0" FC. However, it was demonstrated in chapter 10 that when FCs are used for bug detection these advantages are only evident when FCs are employed with experienced users, i.e. inexperienced users found the FC took longer and was difficult to complete than a questionnaire.

Although no quantitative data exists on the cost to the investigator of employing FCs, it was shown that in a "real-life", industrial setting, FCs were designed considerably quicker than a questionnaire.

11.1.4 *Transferring laboratory validity to field validity*

Perhaps one of the most serious criticisms that can be made of the research conducted in this thesis, is that the accuracy (i.e. validity) of FC data was only assessed in laboratory situations. The important question that can therefore be asked is “does this accuracy transfer to field situations?”

In order to answer this question it would be necessary to accurately record every feature that a group of new users invoked with a system over a reasonable period of time in a “real-life” setting. Ideally this would be done by a combination of automatic data logging and unobtrusive observation. This data would then be compared with the answers that these users gave on a FC. There would be many practical problems involved for the investigator e.g. ensuring that they were present for every interaction and that their presence was unobtrusive etc. Given these problems it does not seem practical to answer the question with complete confidence.

However, since the subjects in these studies were unaware that they would be asked questions about the features in the system that they were using, there is no good reason to believe that the results reported would not transfer to “real-life” situations. It may again be the case that the best way to assess the accuracy of FC data in “real-life” situations is to use the obtained data to make design changes to the system and examine how these changes impact on users and their interaction.

11.2 Feature checklists and human memory

Throughout the thesis an attempt was made to try and understand the actual memory processes involved in FCs and how this might explain the high accuracy of subject responses. Each study’s discussion contained a section on relevant memory research. It now seems appropriate to summarise these discussions in order to develop an overall memory explanation of FCs.

In chapter 4 (study 1), a definition of the terms recall and recognition was given (Brown 1976), in which the crucial distinction between the two was that in recall the subject has to generate the target word/command, etc., whilst in recognition the target word/command is presented either alone or with others (false or genuine). Using this definition it could be argued that subjects using the FC were performing a recognition task, whilst those using the ORQ were performing unprompted recall. However, it was

pointed out that in this study, subjects had to decide whether or not they had selected the command (“Did You Use?”) and if so then how often had they selected it (“How often?”); this is clearly not analogous to traditional recall-recognition studies in which the subject’s task is simply to decide whether the target was actually presented.

Having said this however, the results of this study were explained in terms of a “generation-recognition” model of human memory. In concluding this discussion, it was mentioned that there are problems with G-R models (such as cases where recall is superior to recognition), and that the role of context should be considered in any discussion of FCs.

In chapter 5 (study 2), the idea that FCs involved more than simple recognition was developed more fully. It was stressed that the task for subjects in this study (i.e. choosing menu commands in order to play different versions of a computer game), was different from that in study 1 (i.e. simple selection of commands with no effect seen). As a result, it was likely that subjects’ encoding of information at learning in study 2 was more elaborate (e.g. based on task→action mappings) and related to context (e.g. visual aspects of the interface such as command font style, menu headings, etc.). In summary it was stated that, recognition was the first stage of a reconstructive process, i.e. recognition presupposes recall.

In chapter 6 (study 3), the “encoding-specificity hypothesis” (E-SH) was introduced as an explanation for the observed results (Tulving and Thompson 1973). It was argued that FCs designed to match the visual aspects of the interface more closely, provide subjects with a lot of the cues actually used at learning. Although a comparison of the E-SH and G-R models revealed weaknesses of both, it was stressed that both involved a reconstructive approach to memory which the FC attempted to simulate. It was again stated that unlike traditional research on human memory, the task involved in this thesis (i.e. completing a FC), required subjects to retrieve more information than that needed to simply decide whether a feature actually existed. It is likely that this process required subjects to recognise features based on identification rather than familiarity (Mandler 1980)

Finally, in this chapter it was argued that since research interest on recall and recognition has waned (Eysenck et. al. 1990) it might be relevant to discuss FCs in terms of the “transfer-appropriate processing” framework. This framework focused more on the match in mental operations performed at study and test; its basic premise is, that memory for a prior occurrence results from the overlap between the retrieval

processes induced by a memory test and the encoding operations undertaken during learning. The improved accuracy of FCs through better visual design suggests that there is considerable overlap between the memory processes required by the FC and the encoding operations subjects performed at learning (i.e. using “MacPaint”).

In chapter 7 (study 4), it was suggested that different FC columns require subjects to perform different amounts of processing, e.g. asking someone what a feature is for probably requires more reconstruction than asking someone whether a feature existed. Through a discussion of recent research on memory systems (e.g. Tulving 1985; Graf and Schacter 1985), it was argued that regardless of the FC column, the reconstruction required is explicit and involves both semantic and episodic memory. Table 11.1 outlines the types of memory system that each FC column is likely to be more related to. (N.B. answering each column will involve each system to differing degrees, the table is only a speculative proposal indicating which system is probably more dominant in providing information necessary for retrieval).

Table 11.1: Different FC columns and the memory system likely to be more involved

FC Column	Memory System
“Existed?”	Episodic/Semantic
“Used?”	Episodic
“How often?”	Episodic
“What for?”	Semantic or Episodic
“Need?”	Semantic*

* Although the “Need?” column asks subjects for an opinion rather than remembering something it is likely that their opinion will be based on knowledge of command function.

In answering the FC column “Existed?”, it is equally likely that subjects could do this by recalling an occasion (episode) in which they used the command or by knowing that the command is likely to exist even if they have never used it (e.g. “Help”). With the “Used?” and “How often?” columns it is probable that subjects are most likely to think of specific occasions in which they used the command, however, this is not to deny that some amount of recall of command function is also done (i.e. semantic memory). With the “What for?” column it is possible that naive subjects would answer this by thinking of occasions in which they used a command (episodic memory), whilst experienced

users are more likely to have assimilated a command's meaning into semantic memory and have less need to maintain an episodic record of it.

Finally it was argued that G-R models were more appropriate for explaining memory processes in this study than the “transfer-appropriate processing” framework; i.e. in this study subjects were only given the train station name as a cue (on the FC) and it is unlikely that the retrieval process overlapped significantly with subjects' encoding operations during actual use of the underground station.

In chapter 8 (study 5), an attempt was made to increase the accuracy of FCs by including semantic descriptions of each commands' function as an additional cue; the logic behind this being that when subjects learn menu commands and become competent at using them, they are likely to form internal task→action mappings of commands rather than (or as much as) recognising command names. Although the results did not support the use of semantic descriptions this may have been due to the quality of the descriptions rather than false reasoning. However, even if subjects did form internal task→action mappings, it is still likely that actual command selection involved retrieval of visual/spatial elements of the interface (Kaptelinin 1993); it is hoped that well designed FCs will cover many of these elements.

To conclude, therefore, FCs appear to be an accurate HCI measurement instrument; this accuracy can best be explained in terms of a “transfer-appropriate processing” framework of memory, in that the retrieval processes required to complete a FC overlap considerably with the encoding operations that subjects perform when operating software containing pictorial and textual commands.

11.3 Guidelines for feature checklist design and implementation

Since the studies conducted in this thesis involved different types of computer software in different conditions, we are now in a position to offer some guidelines about FC design and implementation. These are as follows:

- When the interface contains many pictorial features such as icons, FCs should be designed to match the actual screen interface as closely as possible, e.g. spatial layout, font style/size, etc.

- FCs should be as short and concise as possible, particularly the number of columns and instructions on how to fill them in (if necessary include an example of a FC with completed columns). The actual columns listed on the FC will depend on the interface under investigation, the information required and the purpose of the investigation.
- FCs should be administered either as a single measurement instrument for measuring command usage, or as the first in a series of instruments, as their performance seems to be blunted by others; this was explained in terms of “familiarity decay” (Mandler 1967).
- Subjects should be allowed as long a time as necessary to complete a FC.
- When FCs are used for bug detection it is important to inform subjects that the FC is intended for interface evaluation and not user-performance evaluation.
- When FCs are used for bug detection it is important to that the subjects are experienced users of the system under investigation.

11.4 Possible future research

Throughout this thesis a number of suggestions were mentioned for future research on FCs, and are collected here.

The most important suggestion for future research would be the implementation of FCs in an iterative design process using experienced users (this was the intention of chapter 10 however, the majority of users in this study were unexpectedly inexperienced users). Once design alterations were made (on the basis of FC information), it would be of interest to look at whether these design changes had significant effects on such things as error rates, user attitudes and usability components such as: guessability, EUP (Jordan 1992), etc.

As well as FCs listing features such as menu commands and icons, it would also be of interest to examine how appropriate it would be to list such things as dialog boxes, error messages, etc. FCs would then be useful for measuring subjects’ knowledge (of function and existence) and usage of such features. Given the results reported in this

thesis, there is reason to suggest that FCs would be an accurate instrument for providing information about these features.

Since visual realism was shown to increase FC accuracy, it may be useful to have FCs appear on the actual computer screen. If subjects could select commands in their normal manner and then have a table appear that asked them questions about usage, knowledge etc., there may be a subsequent increase in accuracy as well as a saving in the researcher's time. Since accessing FCs by this procedure is likely to be a similar process to the actual usage of features in the system in natural use, there may also be a subsequent increase in FC accuracy; i.e. on-screen FCs may access users' "procedural knowledge" (knowledge that lies behind complex actions and typically is rather resistant to attempts to make it conscious).

One area that was not explored in this thesis, was the idea of listing common errors or bugs as features on the FC. Obviously some understanding of system use would be required beforehand in order to know what may be potential bugs. Once potential bugs were known about however, it may be possible to investigate the frequency of occurrence of such bugs using FCs. An important aspect of this requiring careful consideration concerns the way possible bugs could be listed; as we have seen from chapter 8, descriptions of features (whether bugs or command function) is not a straightforward matter.

Leading of from this, is the idea of exploring the concept of semantic descriptions for menu commands more fully. Although the results reported in chapter 8 did not support their use, it was suggested that more careful consideration of these descriptions (using such things as pre-experimental validation by subjects) might improve the accuracy of subject responses (or at least the match between subject responses and recorded usage). Future research could explore this issue in more detail.

Finally, the results reported in chapter 7, suggest that FCs may be applicable to situations outside HCI. In particular areas containing many features and requiring detailed knowledge e.g. complex control panels. The range of this applicability could be explored more fully by looking at the use of FCs in differing ergonomics/human factors contexts.

References

- ANASTASI, A. (1979)** Psychological Testing 5th ed. MacMillan Publishing Co., New York.
- ANDERSON, J.R. and BOWER, G.H. (1972)** Human Associative Memory Washington, DC. Winston.
- BADDELEY, A. (1990)** Human Memory: Theory and Practice. Lawrence Erlbaum Associates Publishers.
- BAHRICK, H.P. (1970)** Two-Phase Model for Prompted Recall. Psychological Review, 77, 215-222.
- BARTLETT, F.C. (1932)** Remembering: A Study in Experimental and Social Psychology. Cambridge University Press. Cambridge, England.
- BAXTER, I. and OATLEY, K. (1991)** Measuring the learnability of spreadsheets in inexperienced users and those with previous spreadsheet experience. Behaviour and Information Technology. vol. 10 pp.475-490.
- BELSON, W. and DUNCAN, J. A. (1962)** A comparison of the check-list and the open-response questioning systems. Applied Statistics, 11, 120-32.
- BEVAN, N. and MacLEOD, M. (1994)** Usability measurement in context. Behaviour and Information Technology, vol. 13, nos. 1 and 2, 132-145.
- BLACKENBERGER and HAHN (1991)** Effects of icon design on HCI. International Journal of Man-Machine Studies, 35, 363-377.

BLAXTON, T.A. (1989) Investigating Dissociations among memory measures: support for a transfer-appropriate processing framework. Journal of Experimental Psychology: learning, memory and cognition, 15, 657-668.

BROWN, J. (1976) Recall and Recognition. John Wiley and Sons Ltd.

CALLAHAN, J., HOPKINS, D., WEISER, M. and SCHNEIDERMAN, B. (1988) An empirical comparison of pie vs. linear menus. Proceedings of the CHI'88 Conference, pp. 95-100

CHESSARI, J. and LINDGAARD, G. (1988) Providing meaningful pictorial information for users of technological equipment. Designing a Better World, Proceedings of the 10th International Conference IEA, Sydney, Australia.

COHEN, L. and HALLIDAY, M. (1982) Statistics For Social Scientists. Paul Chapman Publishing Ltd., London.

CRAIK, F.I.M (1983) On the transfer of information from temporary to permanent memory. Philosophical Transactions of the Royal Society of London, B, 302, 341-359.

DAVIS, R., SUTHERLAND, N.S. and JUDD, B.R. (1961) Information Content In Recognition and Recall. Journal of Experimental Psychology. 61, 422-429.

DRAPER, S.W. (1985) The Nature Of Expertise In UNIX. In B. Shackel (ed.) Human-Computer Interaction - INTERACT '84. North-Holland: Amsterdam pp. 465-471.

DRAPER, S.W. and BARTON, S.B. (1993) Learning By Exploration And Affordance Bugs. Interchi '93 Adjunct Proceedings (eds.) S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, T. White (ACM), pp 75-76.

DREYFUSS, H. (1972) Symbol Sourcebook. Van Nostrand Reinhold Company.

EATON, H.S. (1940) Word Frequency Dictionary. Dover Publications, Inc. New York.

EDGERTON, E. A. (1990) Verbal Reports Of Recall And Information Flow For An Everyday Task. Final Year Honours Thesis, University of Glasgow (Unpublished).

- EDGERTON, E. A. (1992)** A Comparison of The Feature Checklist And The Open Response Questionnaire In HCI Evaluation. Computing Science Research Report, GIST-1993-1, University Of Glasgow.
- EDGERTON, E. A. , LAFFERTY, C. and COOPER, G. (1992)** Visual Realism In Feature Checklist Design: Implications For Validity. Computing Science Research Report, GIST-1993-2, University Of Glasgow.
- EDGERTON, E. A. and LAFFERTY, C. (1992)** Using feature Checklists For Discovering User Knowledge of Glasgow Underground Stations. Computing Science Research Report, GIST-1993-3, University Of Glasgow.
- ESTES, W.K. and DA POLITO, F. (1967)** Independent Variation of Informal Storage and Retrieval Processes in Paired Associate Learning. Journal of Experimental Psychology. 75, 18-26.
- EYSENCK, M.W. and KEANE, M.T. (1990)** Cognitive Psychology: A Student's Handbook. Lawrence Erlbaum Associates Publishers.
- FLEXER, A.J. and TULVING, E. (1978)** Retrieval Independence in Recognition and Recall. Psychological Review. 85,153-171.
- GARDINER, M. M. and CHRISTIE, B. (1987)** Applying Cognitive Psychology To User- Interface Design. John Wiley and Sons Ltd.
- GARDINER, J.M. (1988)** Functional aspects of recollective experience. Memory and Cognition, 16, 309-313.
- GOOD, M., SPINE, T.M., WHITESIDE, J. and GEORGE, P. (1986)** User-derived impact analysis as a tool for usability engineering. In Mantel, M. and Oberton, P. (eds.) Human Factors in Computing Systems, CHI'86 Proceedings, special issue of SIGCHI Bulletin, pp. 241-246
- GRAF, P. and SCHACTER, D.L. (1985)** Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: learning, memory and cognition, 13, 45-53.
- GREGG, V. (1975)** Human Memory. Essential Psychology, A6, Methuen and Co. Ltd.

HAMMOND, N., MORTON, J., MACLEAN, A. and BERNARD, P. (1983) Fragments and Signposts: Users' Models Of The System. 10th International Symposium On Human Factors In Telecommunications. Helsinki, Finland.

JACOBY, L.L. (1983) Remembering the data: analyzing interactive processing in reading. Verbal Learning and Verbal Behaviour, 22, 485-508.

JONES , G.V. (1982) Test of the dual-mechanism theory of recall. Acta Psychologica, 50, 61-72.

JORDAN, P.W. (1992) Usability: A Multi-Component Definition. Computing Science Research Report GIST-1992-1, University Of Glasgow.

JORDAN, P.W. and O'DONNELL, P.J. (1992) The index of interactive difficulty. In E.J. Lovesey (ed., Contemporary Ergonomics 1992, London: Taylor and Francis.

KACMAR, C.J. and CAREY, J.M. (1991) Assessing the usability of icons in user interfaces. Behaviour and Information Technology, vol. 10, no. 6, pp. 443-457

KAPTELININ, V. (1993) Item recognition in menu selection: the effect of practice. Interchi '93 Adjunct Proceedings, pp. 183-184, (eds.) S.Ashlund, K.Mullet, A. Henderson, E. Hollnagel, T. White (ACM).

KARAT, J., McDONALD, E. and ANDERSON, M. (1986) A comparison of menu selection techniques: touch panel, mouse and keyboard. International Journal of Man-Machine Studies, 25.

KARAT, C.-M, CAMPBELL, R. and FIEGEL, T. (1992) Comparisons of empirical testing and walkthrough methods in user interface evaluation. CHI'92 Conference Proceedings. Reading, MA: Addison Wesley, 119-124.

KINTSCH, W. (1970) Learning, Memory and Conceptual Processes. New York: Wiley.

LANSDALE, M., SIMPSON, M. and STROUD,T. (1990) A comparison of words and icons as external memory aids in an information retrieval task. Behaviour and Information Technology, 9, pp. 111-113

- LEWIS, C., POLSON, P., WHARTON, C. and RIEMAN, J. (1990)** Testing a walkthrough methodology for theory based design of walk-up-and-use interfaces. CHI'90 Conference Proceedings. Reading, MA: Addison Wesley, pp. 235-242.
- LINDGAARD G. and PERRY L. (1986)** Words, Words, Words. What Role Do They Play In Interactive Message Handling? Telecom Australia Research laboratories, Branch paper No. 79.
- LINDEGAARD, G. (1987)** What's On The Menu - Or Better, What Should Be On The Menu? A Literature Review and Guidelines For Menu Design. Telecom Australia Research laboratories, Branch paper No. 142.
- LINDGAARD G and MILLAR J. (1989)** Testing The Usability Of Interactive Computer Systems. Ergonomics Society of Australia Workshop. HCI Australia 1989.
- LINTON, M. (1975)** Memory for real-world events. In D.A. Norman and D.E. Rumelhart (Eds.) Explorations in Cognition, Chapter 14, San Francisco: Freeman.
- MANDLER, G. (1967)** Organisation and Memory. In K.W. Spence and J.T. Spence (Eds.), The Psychology of Learning and Motivation: Advances in Research and Theory: Vol. 1. London: Academic Press.
- MANDLER, G. (1980)** Recognising: the judgement of previous occurrence. Psychological Review, 87, 252-271.
- MANDLER, G. and BOECK, W. (1974)** Retrieval Processes in Recognition. Memory and Cognition, 2, 613-615.
- MANDLER, G., PEARLSTONE, A. and KOOPMANS, H.S. (1969)** Effects of organisation and semantic similarity on recall and recognition. Journal of Verbal Learning and Verbal Behaviour, 8, 410-423.
- MAYES, J.T., DRAPER, S.W., MCGREGOR, A.M. and OATLEY, K. (1988)** Information flow in a user interface: the effect of experience and content on the recall of MacWrite screens. People and Computers IV. Cambridge University Press, pp 275-289.
- MAYO, E. (1933)** The human problems of an industrial civilisation. New York: Macmillan.

MICROSOFT CORPORATION (1990) Microsoft Windows: Guide to Programming. Microsoft Press, Redmond, WA.

MOLICH, R. and NIELSEN, J. (1990) Improving a human-computer dialog. Communications of the ACM, 33, 338-348.

MOLICH, R. (1994) Preventing user interface disasters. Behaviour and Information Technology, vol. 13, nos. 1 and 2, 154-159.

MORRIS, C.D., BRANSFORD, J.D. and FRANKS, J.J. (1977) Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behaviour, 16, 519-533.

NEAL, A.S. and SIMONS, R.M. (1983) Playback: a method for evaluating the usability of software and its documentation. In Janda, A. (ed.), CHI'83 Conference Proceedings, ACM, New York, pp. 78-82.

NEISSER, U. (1982) Memory Observed: Remembering In Natural Contexts. W. H. Freeman and Company. San Francisco.

NICKERSON, R.S. and ADAMS, M.J. (1979) Long-term memory for a common object. Cognitive Psychology, 11, 287-307

NIELSEN, J. (1994) Usability Laboratories. Behaviour and Information Technology, (Special Issue), vol. 13, nos. 1 and 2.

NIELSEN, J. and MOLICH, R. (1990) Heuristic evaluation of user interfaces. In, Carrasco Chew, J. and Whiteside, J. (eds.) CHI'90 Conference Proceedings.

NORMAN, D.A. (1988) The Psychology of Everyday Things. Basic Books Inc.: New York.

OPEN UNIVERSITY (1990) Human-Computer Interaction. Unit 5: Fundamentals of Design.

OPPERMANN, R., MURHNER, B., PAETAU, M., PIEPER, M., SIMM, H. and STELLMACHER, I. (1989) Evaluation of Dialog Systems. St. Augustin, Germany: GMD.

PARKIN, A.J. (1980) Levels of processing and the cue overload principle. Quarterly Journal of Experimental Psychology, 32, 427-434.

PARKIN, A.J. (1981) Determinants of cued recall. Current Psychological Research, 1, 291-300.

PARKIN, A.J. (1993) Memory: Phenomena, Experiment and Theory. Blackwell Publishers.

PERRY, L., LINGAARD, G. and WILSHIRE, C. (1986) Do Words Really Matter? Preferences and Performance With An Electronic Message Handling System (Parts I and II). Telecom Australia Research Laboratories, Report No. 7825 and 7826.

RABINOWITZ, J.C., MANDLER, G. and PATTERSON, K.A. (1977) Determinants of Recognition and Recall: Accessibility and Generation. Journal of Experimental Psychology: General, 106, 302-329.

RAVDEN and JOHNSON (1989) Evaluating The Usability Of Human - Computer Interfaces: A Practical Method. Chichester: Ellis Horwood.

REBER, A.S. (1985) The Penguin Dictionary of Psychology. Penguin Books. London.

RHEINGOLD, H. (1989) Icons at the interface: their usefulness. Interacting with Computers, 1, pp. 105-117

REITERER, H. (1992) EVADIS II: A new method to evaluate user interfaces. People and Computers VII. Cambridge University Press.

ROETHLISBERGER, F.J. and DICKSON, W.J. (1939) Management and the worker. Cambridge, Mass.: Harvard University Press.

ROSENBERG, J. (1982) Evaluating The Suggestiveness Of Command Names. Proceedings Human factors In Computer Systems. Gaithersburg, Maryland.

ROY D. F. (1991) Improving Recall By Eyewitnesses Through The Cognitive Interview. The Psychologist, vol. 4, September 1991.

SHARRATT, B. (1990) Human-Computer Interaction Unit 7 Evaluation. The Open University.

SHEPARD, R.N. (1967) Recognition memory for words, sentences and pictures. Journal of Verbal Learning and Verbal Behaviour, 6, 156-163.

SHERRY, D.F. and SCHACTER, D.L. (1987) The evolution of memory systems. Psychological Review, 94, 439-454.

SMILOWITZ, E.D., DARNELL, M.J. and BENSON, A.E. (1994) Are we overlooking some usability testing methods? A comparison of lab, beta, and forum tests. Behaviour and Information Technology, vol. 13, nos. 1 and 2, 183-190.

SMITH, S.M. (1986) Environmental context-dependent memory: recognition memory using a short-term memory task for input. Memory and Cognition, 14, 347-354.

SQUIRE, L.R. (1987) Memory and Brain. New York: Oxford University Press.

SUTCLIFFE, A.G. and SPRINGETT, M.V. (1992) From user's problems to design errors: Linking evaluation to improving design practice. In Monk, A., Diaper, D. and Harrison M.D. (eds.) People and Computers VII Proceedings of the HCI'92 Conference.

TELLES, M. (1990) Updating An Older Interface. In Carrasco, J. and Whiteside, J. CHI'90 Conference Proceedings.

TULVING, E. (1974) Cue Dependent Forgetting. American Scientist, 62, 74-82.

TULVING, E. (1983) Elements of Episodic Memory. New York: Oxford University Press.

TULVING, E. (1985) How many memory systems are there? 1984 APA Award Address. American Psychologist, 40, 385-398.

TULVING, E. and THOMPSON, D. M. (1971) Retrieval Processes In Recognition Memory: Effects of Associative Context. Journal of Experimental Psychology, 87, 116-124.

TULVING, E. and THOMPSON, D. M. (1973) Encoding Specificity And Retrieval Processes In Episodic Memory. Psychological Review, 80, 5, 352-373.

TYLDESLEY, D.A. (1990) Employing Usability Engineering in the Development of Office Products. In Preece, J. and Keller, L. (eds.), Human-Computer Interaction, The Open University. Prentice Hall.

WATKINS, M.J. (1973) When is Recall Spectacularly Higher Than Recognition? Journal of Experimental Psychology, 102, 161-163.

WATKINS, M.J. (1979) Engrams as cuegrams and forgetting as cue overload. In C.R. Puff (ed.), Memory Organization and Structure, New York: Academic Press.

WATKINS, M.J. and GARDINER, J.M. (1979) An Appreciation of Generate-Recognise Theory of Recall. Journal of Verbal Learning and Verbal Behaviour, 18, 687-704.

WHITESIDE, J., BENNETT, J. and HOLTZBLATT, K. (1987) Usability engineering: our experience and evolution. In Helander, M. (ed.), Handbook of Human-Computer Interaction, North-Holland: Amsterdam.

Appendices

Appendix 4.1

Computer experience questionnaire

Subject No: _____

Sex: _____

Age: _____

Please Tick Your Answer/s

1. How extensive would you say your experience with computers is?

- (a) none at all
- (b) some
- (c) a lot

2. Please indicate the type of experience, (hands-on) you have had with computers by ticking those boxes which apply.

- (a) Word-processing
- (b) Statistical Packages
- (c) Graphics Packages
- (d) Programming
- (e) Computer Games
- (f) Others (please state) _____

3. Please state which terminals you have worked on.

- (a) "Sun Workstation"
- (b) "IBM"
- (d) "Apple Macintosh"
- (e) "Apricot"
- (f) Others (please state) _____

4. Is this a normal day for you, (i.e. has anything unusual happened today that could affect your performance in the study?).

- (a) Yes (it is a normal day)
- (b) No (it is not a normal day)

Thank You

Appendix 4.2

Keyboard difficulty questionnaire

Subject No: _____

Sex: _____

Age: _____

Instructions

The following questions are an attempt to measure how easy or difficult you found it to select commands from the menus using the keyboard.

Please indicating the degree of difficulty you experienced by circling the appropriate number on the scale for each question.

1. How easy/difficult did you find it to highlight the menu bar using the <enter> key?

(Very Difficult) 0 1 2 3 4 5 6 (Very Easy)

2. How easy/difficult did you find it to open a menu using the appropriate number key (i.e. nos. 1-7)?

(Very Difficult) 0 1 2 3 4 5 6 (Very Easy)

3. How easy/difficult did you find it to highlight the appropriate command using the "arrow" keys?

(Very Difficult) 0 1 2 3 4 5 6 (Very Easy)

Appendix 4.2 (cont.)

4. How easy/difficult did you find it to select the appropriate command using the <enter> key?

(Very Difficult) 0 1 2 3 4 5 6 (Very Easy)

5. How easy/difficult did you find it to use the <clear> key to close dialogue boxes or menus?

(Very Difficult) 0 1 2 3 4 5 6 (Very Easy)

6. Was there any other aspect of the study that you found particularly easy/difficult. (please state).

Thank You

Appendix 4.3

Menu layout, command selections and responses required

	Command	No. of times used	Response
	File		
1	Open	1	esc
2	Open Any File	6	esc
3	Close	2	esc
4	Print	7	esc
5	Quit	0	N/A
	Edit		
6	Delete Forward	0	N/A
7	Glossary	7	esc
8	Commands	3	esc
9	Italic Cursor	1	none
10	Sort	5	none
	Format		
11	Show Ruler	0	N/A
12	Plain Text	4	none
13	Bold	6	none
14	Italic	2	none
15	Underline	4	none
	Font		
16	12 Point	2	none
17	24 Point	1	none
18	Helvetica	6	none
19	New York	3	none
20	Times	4	none
	Document		
21	Open Footer	0	N/A
22	Full Repaginate Now	3	none
23	Demote Heading	5	none
24	Expand Subtext	2	none
25	Collapse Selection	6	none
	Utilities		
26	Find	7	esc
27	Change	1	esc
28	Go To	5	esc
29	Spelling	0	N/A
30	Word Count	3	esc
	Window		
31	Help	5	esc
32	Show Clipboard	0	N/A
33	New Window	0	N/A
34	Untitled 1	7	none
35	All Caps	4	none

Appendix 4.4

"Microsoft WORD 4.2" O.R.Q.

Subject No: _____

Sex: _____

Age: _____

Instructions

Part I

In the column headed "COMMAND" (overleaf) would you please write down the names of all commands that you can remember seeing in yesterday's study. Please write down the command names that you can remember seeing, regardless of whether or not you actually used that command.

Part II

In the column headed "Q1 DID YOU USE?" (overleaf), would you please put a tick (✓) against each command that you definitely used and a cross (✗) against each command that you definitely did not use. If you are not sure or don't know please leave blank.

Part III

In the column headed "Q2 HOW OFTEN?" (overleaf), would you please write down the approximate number of times that you used that command (if at all), e.g. "3"

Appendix 4.4 (cont.)

"Microsoft WORD 4.2" O.R.Q.

COMMAND

Q1 DID YOU USE?

Q2 HOW OFTEN?

COMMAND	Q1 DID YOU USE?	Q2 HOW OFTEN?

Thank You

Appendix 4.5

"Microsoft WORD 4.2" F.C.

Subject No: _____

Sex: _____

Age: _____

Instructions

Can you remember which of the following commands (overleaf) you used in yesterday's study?

In the column headed "Q1 DID YOU USE?" please put a tick (✓) against each command that you definitely used and a cross (✗) against each command that you definitely did not use. If you are not sure or don't know please leave blank.

In the column headed "Q2 HOW OFTEN?" please write down the approximate number of times that you used that command (if at all), e.g. "3"

Please answer each question for every command in the list by filling in the appropriate space.

Appendix 4.5 (cont.)

"Microsoft WORD 4.2" F.C.

	COMMAND	Q1: DID YOU USE?	Q2: HOW OFTEN?
	File		
1	Open		
2	Open Any File		
3	Close		
4	Print		
5	Quit		
	Edit		
6	Delete Forward		
7	Glossary		
8	Commands		
9	Italic Cursor		
10	Sort		
	Format		
11	Show Ruler		
12	Plain Text		
13	Bold		
14	Italic		
15	Underline		
	Font		
16	12 Point		
17	24 Point		
18	Helvetica		
19	New York		
20	Times		
	Document		
21	Open Footer		
22	Full Repaginate Now		
23	Demote Heading		
24	Expand Subtext		
25	Collapse Selection		
	Utilities		
26	Find		
27	Change		
28	Go To		
29	Spelling		
30	Word Count		
	Window		
31	Help		
32	Show Clipboard		
33	New Window		
34	Untitled 1		
35	All Caps		

Appendix 4.6

Tukey test for comparing means

	Comparison Means	M1	M2	M3	M4
FC 1st	M1 = 30.4		M1-M2 = 3.5	M1-M3 = 15.6	M1-M4 = 13.5
FC 2nd	M2 = 26.9			M2-M3 = 12.1	M2-M4 = 10.0
ORQ 1st	M3 = 14.8				M3-M4 = -2.1
ORQ 2nd	M4 = 16.9				

$$T_{(0.05)} = 4.87 \times \sqrt{(12.236/9)} = 5.678$$

If the T value of 5.678 is smaller than the difference between two means, then the means are significantly different. Referring to our table of mean differences, we see that there are significant differences between the FCs and the ORQs. We can therefore conclude that the FC was significantly more accurate than the ORQ on correct usage scores in all cases, regardless of the order of presentation. There were no significant differences in the accuracy of correct usage scores between the same instruments when their order was varied (FC 1st vs. FC 2nd) or (ORQ 1st vs. ORQ 2nd).

Appendix 4.7

Incorrect usage scores (false positives vs. false negatives) for the F.C. presented first: T-test for related samples

Subject No.	"False Positives"	"False Negatives"	Differences (X ₁ - X ₂)	Differences Squared
	X ₁	X ₂	d	d ²
1	2	0	2	4
2	0	1	-1	1
3	1	0	1	1
4	0	0	0	0
5	6	0	6	36
6	3	1	2	4
7	1	0	1	1
8	2	0	2	4
9	4	0	4	16
			$\Sigma d = 17$	$\Sigma d^2 = 67$

$$t = 2.72$$

For eight degrees of freedom the value of t required for the 5 per cent significance (two-tailed) is 2.306. As the observed value of t is greater than 2.306, we can conclude that there is a significant difference between the number of "false positives" and the number of "false negatives" for the F.C. presented first.

Appendix 4.8

Spearman's rank correlation coefficient: recall score by usage frequency

Command	Usage	Recall	Usage (Rank)	Recall (Rank)	D	D ²
Open	1	10	26.5	16.5	10	100
Open Any File	6	15	6.5	6	0.5	0.25
Close	2	10	22.5	16.5	6	36
Print	7	15	2.5	6	-3.5	12.25
Quit	0	4	32	25.5	6.5	42.25
Delete Forward	0	0	32	34	-2	4
Glossary	7	11	2.5	14	-11.5	132.25
Commands	3	5	18.5	23.5	-5	25
Italic Cursor	1	5	26.5	23.5	3	9
Sort	5	6	10.5	21	-10.5	110.25
Show Ruler	0	2	32	29.5	2.5	6.25
Plain Text	4	4	14.5	25.5	-11	121
Bold	6	15	6.5	6	0.5	0.25
Italic	2	16	22.5	4	18.5	342.25
Underline	4	13	14.5	10.5	4	16
12 Point	2	17	22.5	2	20.5	420.25
24 Point	1	17	26.5	2	24.5	600.25
Helvetica	6	17	6.5	2	4.5	20.25
New York	3	14	18.5	8.5	10	100
Times	4	14	14.5	8.5	6	36
Open Footer	0	0	32	34	-2	4
Full Repaginate Now	3	11	18.5	14	4.5	20.25
Demote Heading	5	6	10.5	21	-10.5	110.25
Expand Subtext	2	3	22.5	27.5	-5	25
Collapse Selection	6	3	6.5	27.5	-21	441
Find	7	12	2.5	12	-9.5	90.25
Change	1	1	26.5	31.5	-5	25
Go To	5	11	10.5	14	-3.5	12.25
Spelling	0	2	32	29.5	2.5	6.25
Word Count	3	6	18.5	21	-2.5	6.25
Help	5	13	10.5	10.5	0	0
Show Clipboard	0	0	32	34	-2	4
New Window	0	1	32	31.5	0.5	0.25
Untitled 1	7	7	2.5	18.5	-16	256
All Caps	4	7	14.5	18.5	-4	16
						3150.5

$$r(s) = 1 - \frac{6 \sum D_2}{N(N_2 - 1)}$$

$$r(s) = 1 - \frac{6 \times 3150.5}{35 \times 1224}$$

$$r(s) = 0.559 \quad \text{This is significant at the 0.01 level (one-tailed)}$$

Appendix 4.9

Kendall's partial rank correlation coefficient: recall score by usage frequency

Variable (1) = recall score

Variable (2) = usage frequency

Variable (3) = language frequency

Rank ordering of commands by the experimental variables

Command Name	Recall Score	Usage Frequency	Language Frequency
Helvetica	1.0	4.5	14.0
Italic	2.0	10.5	14.0
Bold	3.5	4.5	8.0
Print	3.5	2.0	9.0
Times	5.0	8.5	3.5
Help	6.5	6.5	3.5
Underline	6.5	8.5	12.0
Find	8.0	2.0	3.5
Glossary	9.0	2.0	14.0
Open	10.5	12.5	3.5
Close	10.5	10.5	3.5
Sort	12.0	6.5	3.5
Quit	13.0	14.5	10.0
Spelling	14.0	14.5	11.0
Change	15.0	12.5	7.0

Rank correlation coefficients:- $r_{12} = 0.581$; $r_{13} = -0.304$; $r_{23} = -0.042$

$$\begin{aligned}
 \text{Equation: } r_{12.3} &= \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - [r_{13} \times r_{13}])(1 - [r_{23} \times r_{23}])}} \\
 &= \frac{0.581 - 0.013}{\sqrt{(0.908 \times 0.998)}} \\
 &= \frac{0.568}{0.952} \\
 &= 0.597
 \end{aligned}$$

It can be seen that with the effects of language frequency partialled out, there is a modest correlation that is still significant, between recall score and usage frequency; i.e. the more frequently a command was used in the study, the more likely it is to be recalled by subjects.

Appendix 4.10

Kendall's partial rank correlation coefficient: recall score by language frequency

Variable (1) = recall score

Variable (2) = language frequency

Variable (3) = usage frequency

Rank ordering of commands by the experimental variables

Command Name	Recall Score	Language Frequency	Usage Frequency
Helvetica	1.0	14.0	4.5
Italic	2.0	14.0	10.5
Bold	3.5	8.0	4.5
Print	3.5	9.0	2.0
Times	5.0	3.5	8.5
Help	6.5	3.5	6.5
Underline	6.5	12.0	8.5
Find	8.0	3.5	2.0
Glossary	9.0	14.0	2.0
Open	10.5	3.5	12.5
Close	10.5	3.5	10.5
Sort	12.0	3.5	6.5
Quit	13.0	10.0	14.5
Spelling	14.0	11.0	14.5
Change	15.0	7.0	12.5

Rank correlation coefficients:- $r_{12} = -0.304$; $r_{13} = 0.581$; $r_{23} = -0.042$

$$\begin{aligned}
 \text{Equation: } r_{12.3} &= \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - [r_{13} \times r_{13}])(1 - [r_{23} \times r_{23}])}} \\
 &= \frac{-0.304 - (-0.024)}{\sqrt{(1 - [0.338 \times 0.338])(1 - [0.002])}} \\
 &= \frac{-0.28}{0.9997} \\
 &= -0.28
 \end{aligned}$$

It can be seen that with the effects of usage frequency partialled out, there is a low negative correlation that is not significant, between recall score and language frequency. We can therefore conclude that recall score and language frequency are unrelated.

Appendix 4.11

Spearman's rank correlation coefficient: recognition score by usage frequency

Command	Usage	Recognition	Usage (Rank)	Recognition (Rank)	D	D ²
Open	1	13	26.5	26	.5	.25
Open Any File	6	17	6.5	11	-4.5	20.25
Close	2	10	22.5	33.5	-11	121
Print	7	18	2.5	4	-1.5	2.25
Quit	0	11	32	32	0	0
Delete Forward	0	13	32	26	6	36
Glossary	7	18	2.5	4	-1.5	2.25
Commands	3	14	18.5	21.5	-3	9
Italic Cursor	1	13	26.5	26	.5	.25
Sort	5	14	10.5	21.5	-11	121
Show Ruler	0	14	32	21.5	10.5	110.25
Plain Text	4	13	14.5	26	-11.5	132.25
Bold	6	18	6.5	4	2.5	6.25
Italic	2	17	22.5	11	11.5	132.25
Underline	4	16	14.5	15.5	-1	1
12 Point	2	17	22.5	11	11.5	132.25
24 Point	1	15	26.5	18	8.5	72.25
Helvetica	6	18	6.5	4	2.5	6.25
New York	3	13	18.5	26	-7.5	56.25
Times	4	17	14.5	11	3.5	12.25
Open Footer	0	14	32	21.5	10.5	110.25
Full Repaginate Now	3	18	18.5	4	14.5	210.25
Demote Heading	5	17	10.5	11	-.5	.25
Expand Subtext	2	10	22.5	33.5	-11	121
Collapse Selection	6	16	6.5	15.5	-9	81
Find	7	17	2.5	11	-8.5	72.25
Change	1	6	26.5	35	-8.5	72.25
Go To	5	17	10.5	11	-.5	.25
Spelling	0	12	32	30	2	4
Word Count	3	12	18.5	30	-11.5	132.25
Help	5	18	10.5	4	6.5	42.25
Show Clipboard	0	15	32	18	14	196
New Window	0	12	32	30	2	4
Untitled 1	7	18	2.5	4	-1.5	2.25
All Caps	4	15	14.5	18	-3.5	12.25
						2034

$$r(s) = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$r(s) = 1 - \frac{6 \times 2034}{35 \times 1224}$$

$$r(s) = 0.715 \quad \text{This is significant at the 0.01 level (one-tailed)}$$

Appendix 4.12

Kendall's partial rank correlation coefficient: recognition score by language frequency

Variable (1) = recognition score

Variable (2) = language frequency

Variable (3) = usage frequency

Rank ordering of commands by the experimental variables

Command Name	Recognition Score	Language Frequency	Usage Frequency
Helvetica	4.5	14.0	4.5
Italic	10.5	14.0	10.5
Bold	4.5	8.0	4.5
Print	2.0	9.0	2.0
Times	8.5	3.5	8.5
Help	6.5	3.5	6.5
Underline	8.5	12.0	8.5
Find	2.0	3.5	2.0
Glossary	2.0	14.0	2.0
Open	12.5	3.5	12.5
Close	10.5	3.5	10.5
Sort	6.5	3.5	6.5
Quit	14.5	10.0	14.5
Spelling	14.5	11.0	14.5
Change	12.5	7.0	12.5

Rank correlation coefficients:- $r_{12} = -0.265$; $r_{13} = 0.816$; $r_{23} = -0.042$

$$\begin{aligned}
 \text{Equation: } r_{12.3} &= \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - [r_{13} \times r_{13}])(1 - [r_{23} \times r_{23}])}} \\
 &= \frac{-0.265 - (-0.034)}{\sqrt{(1 - [0.666 \times 0.002])}} \\
 &= \frac{-0.231}{0.9993} \\
 &= -0.231
 \end{aligned}$$

It can be seen that with the effects of usage frequency partialled out, there is a low negative correlation that is not significant, between recognition score and language frequency. We can therefore conclude that recognition score and language frequency are unrelated.

Appendix 5.1

"Brickles 7.0" command task sheet

Subject No: _____

Sex: _____

Age: _____

Date: _____

INSTRUCTIONS

The following is a list of tasks (i.e.. command selections from menus) that you have to complete as part of the "Brickles 7.0" game you are about to play. Please go through each task in the order laid out.

In the process of carrying out these tasks you will play different versions of the "Brickles 7.0" game with different patterns, speeds, etc. At the end of each game please write your score in the space provided.

If you have any problems please ask the experimenter for help. (N.B. It might be a useful idea to score off each task as you perform them).

Appendix 5.1 (cont.)

Select	Show High Scores	from the File menu
Select	Slower	from the QuickOne menu
Select	Paddle Large	from the QuickOne menu
Select	Paddle Black	from the QuickTwo menu
Select	Ball Black	from the QuickTwo menu
Select	New Game - One Paddle	from the File menu

Now Click On The Screen To Play The Game

Score

Select	New Game - One Paddle	from the File menu
Select	Bkground Gray	from the QuickTwo menu
Select	Bricks White	from the QuickTwo menu
Select	Paddle Medium	from the QuickOne menu
Select	Copy	from the Edit menu

Now Click On The Screen To Play The Game

Score

Select	New Game - Two Paddles	from the File menu
Select	Paddle Large	from the QuickOne menu
Select	Bricks Bricks	from the QuickTwo menu
Select	Paddle White	from the QuickTwo menu
Select	Show High Scores	from the File menu
Select	Paste	from the Edit menu
Select	Slow	from the QuickOne menu

Now Click On The Screen To Play The Game

Score

Appendix 5.1 (cont.)

Select	New Game - Two Paddles	from the File menu
Select	Cut	from the Edit menu
Select	Sound OnOff	from the QuickOne menu
Select	Paddle XLarge	from the QuickOne menu
Select	Ball Large	from the QuickOne menu
Select	Paddle Black	from the QuickTwo menu
Select	Show High Scores	from the File menu

Now Click On The Screen To Play The Game

Score

Select	Ball XLarge	from the QuickOne menu
Select	Paddle White	from the QuickTwo menu
Select	Bkground Black	from the QuickTwo menu
Select	Ball Gray	from the QuickTwo menu
Select	Fast	from the QuickOne menu
Select	New Game - Four Paddles	from the File menu

Now Click On The Screen To Play The Game

Score

Select	New Game - One Paddle	from the File menu
Select	Show High Scores	from the File menu
Select	Cut	from the Edit menu
Select	Sound OnOff	from the QuickOne menu
Select	Slow	from the QuickOne menu
Select	Bricks Gray	from the QuickTwo menu
Select	Ball White	from the QuickTwo menu

Now Click On The Screen To Play The Game

Score

Appendix 5.1 (cont.)

Select	New Game - Two Paddles	from the File menu
Select	Slower	from the QuickOne menu
Select	Copy	from the Edit menu
Select	Ball Medium	from the QuickOne menu
Select	Bricks Bricks	from the QuickTwo menu

Now Click On The Screen To Play The Game

Score_____

Select	New Game - One Paddle	from the File menu
Select	Paddle Medium	from the QuickOne menu
Select	Bkground White	from the QuickTwo menu
Select	Ball Black	from the QuickTwo menu
Select	Show High Scores	from the File menu
Select	Paddle Black	from the QuickTwo menu

Now Click On The Screen To Play The Game

Score_____

Select	New Game - Two Paddles	from the File menu
Select	Clear	from the Edit menu
Select	Show High Scores	from the File menu

Now Click On The Screen To Play The Game

Score_____

Appendix 5.1 (cont.)

- | | | |
|--------|-----------------------|-------------------------------|
| Select | Bkground Gray | from the QuickTwo menu |
| Select | Slow | from the QuickOne menu |
| Select | New Game - One Paddle | from the File menu |

Now Click On The Screen To Play The Game

Score_____

- | | | |
|--------|-----------------------|-------------------------------|
| Select | New Game - One Paddle | from the File menu |
| Select | Paddle White | from the QuickTwo menu |
| Select | Slower | from the QuickOne menu |
| Select | Copy | from the Edit menu |
| Select | Show High Scores | from the File menu |

Now Click On The Screen To Play The Game

Score_____

- | | | |
|--------|------|--------------------|
| Select | Quit | from the File Menu |
|--------|------|--------------------|

Appendix 5.2

"Brickles 7.0" descriptive task sheet

Subject No: _____

Sex: _____

Age: _____

Date: _____

INSTRUCTIONS

The following is a group of tasks (i.e.. command selections from menus) that you have to complete as part of the "Brickles 7.0" game you are about to play. Please go through each group of tasks in the order laid out.

In the process of carrying out these tasks you will play different versions of the "Brickles 7.0" game with different patterns, speeds, etc.; please try to make these games as different as possible i.e. choose many different commands.

At the end of each game please write your score in the space provided.

If you have any problems please ask the experimenter for help.

(N.B. It might be a useful idea to score off each task as you perform them).

Appendix 5.2 (cont.)

- Look at the previous best scores
- Decrease the game speed
- Choose the second biggest bat size
- Choose the darkest colour for the bat
- Choose the darkest colour for the ball
- Choose the game with one bat

Now Click On The Screen To Play The Game

Score_____

- Choose the game with one bat
- Choose a darker background colour
- Choose the lightest colour for the wall
- Choose a smaller bat size
- Choose copy from the edit menu

Now Click On The Screen To Play The Game

Score_____

- Choose the game with two bats
- Choose a larger bat size
- Choose a new pattern for the wall
- Choose the lightest colour for the bat
- Look at the previous best scores
- Choose paste from the edit menu
- Choose a quicker game speed

Now Click On The Screen To Play The Game

Score_____

Appendix 5.2 (cont.)

- Choose the game with two bats
- Choose cut from the edit menu
- Switch off the noise effects
- Choose a bigger bat size
- Choose a bigger ball size
- Choose a darker colour for the bat
- Look at the previous best scores

Now Click On The Screen To Play The Game

Score_____

- Choose a bigger ball size
- Choose a lighter colour for the bat
- Choose the darkest colour for the background
- Choose a lighter colour for the ball
- Choose a quicker game speed
- Choose the game with four bats

Now Click On The Screen To Play The Game

Score_____

- Choose the game with one bat
- Look at the previous best scores
- Choose cut from the edit menu
- Switch the noise effects on
- Decrease the game speed
- Choose a light colour for the wall
- Choose the lightest colour for the ball

Now Click On The Screen To Play The Game

Score_____

Appendix 5.2 (cont.)

- Choose the game with two bats
- Decrease the game speed
- Choose copy from the edit menu
- Choose a smaller ball size
- Choose a new pattern for the wall

Now Click On The Screen To Play The Game

Score_____

- Choose the game with one bat
- Choose a smaller bat size
- Choose the lightest colour for the background
- Choose the darkest colour for the ball
- Look at the previous best scores
- Choose the darkest colour for the bat

Now Click On The Screen To Play The Game

Score_____

- Choose the game with two bats
- Choose clear from the edit menu
- Look at the previous best scores

Now Click On The Screen To Play The Game

Score_____

Appendix 5.2 (cont.)

Choose a darker background colour

Choose a quicker game speed

Choose the game with one bat

Now Click On The Screen To Play The Game

Score

Choose the game with one bat

Choose the lightest bat colour

Decrease the game speed

Choose copy from the edit menu

Look at the previous best scores

Now Click On The Screen To Play The Game

Score

Exit from the game

Appendix 5.3

Psychology summer school transport survey

(please tick appropriate answer)

(1) Have you ever used the Glasgow Underground in the last year?

Yes
No

If yes, then how often do you use the underground:-

very rarely
occasionally
frequently

(2) Have you ever used any of the bus services in the Glasgow area in the last year?

Yes
No

If yes, then how often do you use the bus services:-

very rarely
occasionally
frequently

(3) Have you ever used any of the train services in the Glasgow area in the last year?

Yes
No

If yes, then how often do you use the train services:-

very rarely
occasionally
frequently

Thank You

Appendix 5.4

"Brickles 7.0" standard F.C.

Subject No: _____

Sex: _____

Age: _____

Date: _____

INSTRUCTIONS

Can you remember what commands you selected from the menus in the "Brickles 7.0" study?

In the column headed **Did You Use?** please put a tick (✓) against each command that you definitely used and a cross (✗) against each command that you definitely did not use. If you are not sure or don't know please leave blank.

In the column headed **How often?** please write down the approximate number of times that you used that command (if at all), e.g. "3"

Please answer each question for every command in the list by filling in the appropriate space.

Appendix 5.4 (cont.)

Did You Use?

How often?

	Did You Use?	How often?
File		
New Game - One Paddle		
New Game - Two Paddles		
New Game - Four Paddles		
Show High Scores		
Clear High Scores		
Quit		
Edit		
Undo		
Cut		
Copy		
Paste		
Clear		
Formats...		
Quick One		
Sound OnOff		
Slower		
Slow		
Fast		
Faster		
Paddle Small		
Paddle Medium		
Paddle Large		
Paddle XLarge		
Ball Small		
Ball Medium		
Ball Large		
Ball XLarge		
Quick Two		
Paddle Black		
Paddle White		
Paddle Gray		
Ball Black		
Ball White		
Ball Gray		
Bkgrnd Black		
Bkgrnd White		
Bkgrnd Gray		
Bricks Black		
Bricks White		
Bricks Gray		
Bricks Bricks		

Appendix 5.5

"Brickles 7.0" visual F.C.

Subject No: _____

Sex: _____

Age: _____

Date: _____

INSTRUCTIONS

Can you remember what commands you selected from the menus in the "Brickles 7.0" study?

In the column headed **Did You Use?** please put a tick (✓) against each command that you definitely used and a cross (✗) against each command that you definitely did not use. If you are not sure or don't know please leave blank.

In the column headed **How often?** please write down the approximate number of times that you used that command (if at all), e.g. "3"

Please answer each question for every command in the list by filling in the appropriate space.

"Brickles 7.0" visual F.C.

File
New Game - One Paddle
New Game - Two Paddles
New Game - Four Paddles

Show High Scores
Clear High Scores

Quit

Did You Use?

How Often?

Edit
Undo

Cut
Copy
Paste
Clear

Formats...

Did You Use?

How Often?

"Brickles 7.0" visual F.C.

QuickOne
Sound OnOff
Slower
Slow
Fast
Faster
Paddle Small
Paddle Medium
Paddle Large
Paddle XLarge
Ball Small
Ball Medium
Ball Large
Ball XLarge

Did You Use?

How Often?

QuickTwo
Paddle Black
Paddle White
Paddle Gray
Ball Black
Ball White
Ball Gray
Bkgrnd Black
Bkgrnd White
Bkgrnd Gray
Bricks Black
Bricks White
Bricks Gray
Bricks Bricks

Did You Use?

How Often?

Appendix 5.6

Standard F.C. mean scores

	Actual Usage	Standard F.C.	D	D2
File				
New Game - One Paddle	6.91	3.73	- 3.18	10.11
New Game - Two Paddles	4.45	2.82	- 1.63	2.66
New Game - Four Paddles	1.09	1.09	0	0
Show High Scores	7.00	4.73	- 2.27	5.15
Clear High Scores	0.00	0.09	+ 0.09	0.01
Quit	1.00	1.00	0	0
Edit				
Undo	0.09	0.09	0	0
Cut	2.00	1.36	- 0.64	0.41
Copy	3.00	2.00	- 1.00	1.00
Paste	1.00	0.91	- 0.09	0.01
Clear	1.00	0.54	- 0.45	0.20
Formats...	0.00	0.00	0	0
Quick One				
Sound OnOff	2.27	2.00	- 0.27	0.07
Slower	3.18	2.45	- 0.73	0.53
Slow	2.54	1.91	- 0.63	0.40
Fast	1.18	1.00	- 0.18	0.03
Faster	0.00	0.27	+ 0.27	0.07
Paddle Small	0.18	0.64	+ 0.46	0.21
Paddle Medium	1.36	1.54	+ 0.18	0.03
Paddle Large	1.91	1.36	- 0.55	0.30
Paddle XLarge	1.00	1.27	+ 0.27	0.07
Ball Small	0.00	0.45	+ 0.45	0.20
Ball Medium	0.91	1.27	+ 0.36	0.13
Ball Large	1.09	1.64	+ 0.55	0.30
Ball XLarge	1.18	1.09	- 0.09	0.01
Quick Two				
Paddle Black	2.91	2.18	- 0.73	0.53
Paddle White	2.91	1.54	- 1.37	1.88
Paddle Gray	0.45	1.18	+ 0.73	0.53
Ball Black	2.00	2.36	+ 0.36	0.13
Ball White	1.18	1.54	+ 0.36	0.13
Ball Gray	1.00	1.54	+ 0.54	0.29
Bkgrnd Black	1.63	1.36	- 0.27	0.07
Bkgrnd White	1.18	2.18	+ 1.00	1.00
Bkgrnd Gray	1.81	1.54	- 0.27	0.07
Bricks Black	1.18	0.91	- 0.27	0.07
Bricks White	1.18	1.36	+ 0.18	0.03
Bricks Gray	1.18	1.45	+ 0.27	0.07
Bricks Bricks	1.73	1.73	0	0
			TOTAL	26.70

Appendix 5.7

Visual F.C. mean scores

	Actual Usage	Visual F.C.	D	D2
File				
New Game - One Paddle	6.50	6.10	- 0.40	0.16
New Game - Two Paddles	4.64	3.18	- 1.46	2.13
New Game - Four Paddles	1.00	1.00	0	0
Show High Scores	6.78	4.67	- 2.11	4.45
Clear High Scores	0.18	0.09	- 0.09	0.01
Quit	1.10	0.80	- 0.30	0.09
Edit				
Undo	0.00	0.00	0	0
Cut	2.09	1.45	- 0.64	0.41
Copy	2.81	1.64	- 1.17	1.37
Paste	1.00	1.20	+ 0.20	0.04
Clear	1.00	0.80	+ 0.20	0.04
Formats...	0.00	0.00	0	0
Quick One				
Sound OnOff	1.91	1.91	0	0
Slower	3.09	3.36	+ 0.27	0.07
Slow	2.90	3.09	+ 0.19	0.04
Fast	1.36	1.45	+ 0.09	0.01
Faster	0.18	0.82	+ 0.64	0.41
Paddle Small	0.36	2.00	+ 1.64	2.69
Paddle Medium	2.00	2.45	+ 0.45	0.20
Paddle Large	1.90	2.20	+ 0.30	0.09
Paddle XLarge	1.09	1.09	0	0
Ball Small	0.18	1.27	+ 1.09	1.19
Ball Medium	0.80	2.00	+ 1.20	1.44
Ball Large	1.18	1.18	0	0
Ball XLarge	0.82	0.73	- 0.09	0.01
Quick Two				
Paddle Black	2.64	3.27	+ 0.63	0.40
Paddle White	3.09	2.45	- 0.64	0.41
Paddle Gray	0.73	1.54	+ 0.81	0.66
Ball Black	2.09	3.64	+ 1.55	2.40
Ball White	1.45	2.36	+ 0.91	0.83
Ball Gray	0.80	1.00	+ 0.20	0.04
Bkgrnd Black	1.91	1.73	- 0.18	0.03
Bkgrnd White	1.09	3.27	+ 2.18	4.75
Bkgrnd Gray	1.64	2.18	+ 0.54	0.29
Bricks Black	0.09	1.18	+ 1.09	1.19
Bricks White	1.36	2.82	+ 1.46	2.13
Bricks Gray	0.64	1.91	+ 1.27	1.61
Bricks Bricks	1.91	2.09	+ 0.18	0.03
			TOTAL	29.62

Appendix 6.1

"MacPaint" standard F.C.

Subject No: _____

Sex: _____

Age: _____

Date: _____

Instructions

Can you remember which of the following commands, tools, patterns, etc. that you used in the "MacPaint" application?

In the column headed "Q1 USED?" please put a tick (✓) against each command, tool, pattern, etc. that you definitely used and a cross (✗) against each one that you definitely did not use.

If you are not sure or don't know please leave blank.

In the column headed "Q2 HOW OFTEN?" please write down the approximate number of times that you used that command, tool, pattern, etc. (if at all), e.g. "3"

Please answer each question for every command in the list by filling in the appropriate space.

Appendix 6.1 (cont.)

Used? How often?

File		
New		
Open...		
Close		
Save		
Save As...		
Revert		
Print Draft		
Print Final		
Print Catalog		
Quit		
Edit		
Undo		
Cut		
Copy		
Paste		
Clear		
Invert		
Fill		
Trace Edges		
Flip Horizontal		
Flip Vertical		
Rotate		
Goodies		
Grid		
FatBits		
Show Page		
Edit Pattern		
Brush Shape		
Brush Mirrors		
Introduction		
Short Cuts		

Appendix 6.1 (cont.)

Used? How often?

Font		
Athens		
Cairo		
Chicago		
Courier		
Geneva		
Helvetica		
London		
Los Angeles		
Monaco		
New York		
Palatino		
San Francisco		
Symbol		
Times		
Venice		
FontSize		
9 point		
10		
12		
14		
18		
24		
36		
48		
72		
Style		
Plain		
Bold		
Italic		
Underline		
Outline		
Shadow		
Align Left		
Align Middle		
Align Right		

Appendix 6.1 (cont.)

Used? How often?

Tool		
		
		
		
		
		
		
		
		
		
		
		
A		
		
		
		
		
		
		
		
		
Line		
		
		
		
		
		

Appendix 6.1 (cont.)

Used?

How often?

Pattern		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		
		

Subject No: _____
Sex: _____
Age: _____
Date: _____

Instructions

Can you remember which of the following commands, tools, patterns, etc. that you used in the MacPaint application?

In the column headed "Q1 USED?" please put a tick (✓) against each command, tool, pattern, etc. that you definitely used and a cross (X) against each one that you definitely did not use.

If you are not sure or don't know please leave blank.

In the column headed "Q2 HOW OFTEN?" please write down the approximate number of times that you used that command, tool, pattern, etc. (if at all), e.g. "3"

Please answer each question for every command in the list by filling in the appropriate space.

Appendix 6.2 (cont.)

Command Name	Used?	How Often?
Goodies		
Grid		
FatBits		
Show Page		
Edit Pattern		
Brush Shape		
Brush Mirrors		
Introduction		
Short Cuts		

Command Name	Used?	How Often?
Edit		
Undo		
Cut		
Copy		
Paste		
Clear		
Invert		
Fill		
Trace Edges		
Flip Horizontal		
Flip Vertical		
Rotate		

Command Name	Used?	How Often?
File		
New		
Open...		
Close		
Save		
Save As...		
Revert		
Print Draft		
Print Final		
Print Catalog		
Quit		

"MacPaint" visual F.C.

Appendix 6.2 (cont.)

Command Name	Used?	How Often?
Font		
Athens		
Cairo		
Chicago		
Courier		
Geneva		
Helvetica		
London		
Los Angeles		
Monaco		
New York		
Palatino		
San Francisco		
Symbol		
Times		
Venice		

Command Name	Used?	How Often?
FontSize		
9 point		
10		
12		
14		
18		
24		
36		
48		
72		

Command Name	Used?	How Often?
style		
Plain		
Bold		
Italic		
Underline		
Outline		
Shadow		
Align Left		
Align Middle		
Align Right		

Appendix 6.3

Extension of the Median Test

Median for all three groups = 54

	A	B	C
No. of scores above median	3* 0	3* 4	3* 4
No. of scores below median	3* 6	3* 2	3* 2

$$X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where, O_{ij} = observed number of cases categorised in i th row of j th column

E_{ij} = number of cases expected under H_0 to be categorised in i th row of j th column

$\sum_{i=1}^r \sum_{j=1}^k$

directs one to sum over all cells.

$\sum_{i=1}^r \sum_{j=1}^k$

$$\frac{(0-3)^2}{3} + \frac{(3-6)^2}{3} + \frac{(3-4)^2}{3} + \frac{(3-2)^2}{3} + \frac{(3-4)^2}{3} + \frac{(3-2)^2}{3}$$

$$\frac{9 + 9 + 1 + 1 + 1 + 1}{3}$$

22

3

= 7.333

Appendix 7.1

Glasgow underground route F.C.

Name: _____

Sex: _____ Age: _____

Have you ever used the Glasgow Underground in the last year?

YES NO (Please tick your answer)

If yes, then how often do you use the underground:-

FREQUENTLY OCCASIONALLY VERY RARELY

INSTRUCTIONS

This Feature Checklist is inquiring about the train stations on the Glasgow Underground.

Down the left-hand side of the feature checklist are the names of various Glasgow Underground stations, and along the top of the feature checklist are five headings which ask a different question for each underground station. The headings (and their meanings) are as follows:-

- | | |
|-----------------------------|--|
| Q1 Existed? | Did you know this underground station existed? |
| Q2 What for?
e.g. | Do you know what this station might be useful for generally?
museums, train/bus stations, colleges, sports facilities etc.? |
| Q3 Need? | Do you think you would ever have a need to use this
underground station? |
| Q4 Used? | Have you ever used this underground station? |
| Q5 How often? | How often do you use this underground station? |

Please answer each question for every underground station on the feature checklist by filling in the appropriate box with either:-

Q1 - 4 ✓ = Yes ✗ = No ? = Not Sure/Don't Know

Q5 1 = daily; 2 = weekly; 3 = monthly, 4 = yearly; 5 = never

Appendix 7.1 (cont.)

Glasgow underground route F.C.

	Q 1 Existed?	Q 2 What for?	Q 3 Need?	Q 4 Used?	Q 5 How often?
BUCHANAN STREET					
COWCADDENS					
CHARING CROSS					
ST. GEORGE'S CROSS					
KELVINBRIDGE					
HILLHEAD					
HYNDLAND					
KELVIN HALL					
PARTICK					
GOVAN					
IBROX					
BELLAHOUSTON					
CESSNOCK					
KINNING PARK					
SHIELDS ROAD					
WEST STREET					
WOODLANDS					
BRIDGE STREET					
ST. ENOCH					

If selected, would you be willing to take part in a 10-20 minute interview about the Glasgow Underground for which you would be paid £4.00?

YES NO (Please tick your answer)

Glasgow underground alphabetical F.C.

Name: _____

Sex: _____

Age: _____

Have you ever used the Glasgow Underground in the last year?

YES NO (Please tick your answer)

If yes, then how often do you use the underground:-

FREQUENTLY OCCASIONALLY VERY RARELY

INSTRUCTIONS

This Feature Checklist is inquiring about the train stations on the Glasgow Underground.

Down the left-hand side of the feature checklist are the names of various Glasgow Underground stations, and along the top of the feature checklist are five headings which ask a different question for each underground station. The headings (and their meanings) are as follows:-

- | | |
|----------------------|---|
| Q1 Existed? | Did you know this underground station existed? |
| Q2 What for? | Do you know what this station might be useful for generally?
e.g. museums, train/bus stations, colleges, sports facilities etc.? |
| Q3 Need? | Do you think you would ever have a need to use this underground station? |
| Q4 Used? | Have you ever used this underground station? |
| Q5 How often? | How often do you use this underground station? |

Please answer each question for every underground station on the feature checklist by filling in the appropriate box with either:-

Q1 - 4 ✓ = Yes ✗ = No ? = Not Sure/Don't Know

Q5 1 = daily; 2 = weekly; 3 = monthly, 4 = yearly; 5 = never

Appendix 7.2 (cont.)

Glasgow underground alphabetical F.C.

	Q 1 Existed?	Q 2 What for?	Q 3 Need?	Q 4 Used?	Q 5 How often?
BELLAHOUSTON					
BRIDGE STREET					
BUCHANAN STREET					
CESSNOCK					
CHARING CROSS					
COWCADDENS					
GOVAN					
HILLHEAD					
HYNDLAND					
IBROX					
KELVINBRIDGE					
KELVIN HALL					
KINNING PARK					
PARTICK					
SHIELDS ROAD					
ST. ENOCH					
ST. GEORGE'S CROSS					
WEST STREET					
WOODLANDS					

If selected, would you be willing to take part in a 10-20 minute interview about the Glasgow Underground for which you would be paid £4.00?

YES NO (Please tick your answer)

GLASGOW UNDERGROUND

INTERVIEW

Subject No: _____

Date: _____

Sex: _____

Age: _____

Appendix 7.3 (cont.)

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:

Do you know what this station might be used for generally?

Do you think you would ever have a need to use this station?

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:

Do you know what this station might be used for generally?

Do you think you would ever have a need to use this station?

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:

Do you know what this station might be used for generally?

Do you think you would ever have a need to use this station?

Appendix 7.3 (cont.)

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:
Do you know what this station might be used for generally?
Do you think you would ever have a need to use this station?

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:
Do you know what this station might be used for generally?
Do you think you would ever have a need to use this station?

PROMPTS

- Ever used?
- Location?
- Stations
- Work/Institutions
- Tourist attractions
- Shopping
- Entertainment
- Govt. offices, etc.

STATION:
Do you know what this station might be used for generally?
Do you think you would ever have a need to use this station?

Appendix 7.4

Existed scale

Did you know this underground station existed? Using the scale below, please indicate how confident you are about your answer:-

1

(very confident)

2

3

(semi confident)

4

5

(not at all confident)

Appendix 7.5

Need scale

Do you think that you would ever have a need to use this station? Using the following scale, please estimate how much of a need you think you might have:-

1	2	3	4	5
(strong need)		(average need)		(no need at all)

Appendix 7.6

EXISTED COLUMN ANSWERS: COHEN'S KAPPA ANALYSIS (used for summarising subjects' answer agreement)

COLUMN HEADING "EXISTED?"

		S-S Interview Answer		Total
		Yes	No	
F.C. Answer	Yes	0.72 (a)	0.07 (b)	0.79 (p1)
	No	0.02 (c)	0.19 (d)	0.21 (q1)
	Total	0.74 (p2)	0.26 (q2)	1.00

$$\text{Raw Agreement} = (a+d); = 0.91$$

$$\text{Occurrence Agreement} = a/(a+b+c); = 0.89$$

$$\text{Non-occurrence Agreement} = d/(b+c+d); = 0.68$$

$$\begin{aligned} \text{Cohen's Kappa} &= (a+d-p_1p_2-q_1q_2)/(1-p_1p_2-q_1q_2) \\ &= (0.91-0.58-0.05)/(1-0.58-0.05) \\ &= 0.28/0.37 \\ &= \mathbf{0.76} \end{aligned}$$

Appendix 7.7

Existed column answers: T-Test for correct answers (actual answers vs. chance answers)

Actual	Chance	D	D ²
15	15	0	0
18	15	3	9
18	15	3	9
17	15	2	4
14	15	-1	1
18	15	3	9
19	15	4	16
19	15	4	16
14	15	-1	1
17	15	2	4
16	15	1	1
16	15	1	1
16	15	1	1
19	15	4	16
14	15	-1	1
16	15	1	1
19	15	4	16
19	15	4	16
17	15	2	4
11	15	-4	16
		$\Sigma D = 31$	$\Sigma D^2 = 142$

$$t = \frac{\Sigma D}{\sqrt{[N\Sigma D^2 - (\Sigma D)^2] \div N - 1}}$$

$$= \frac{31}{\sqrt{(1879 \div 19)}}$$

$$= 3.117 \text{ (df = 19)}$$

The obtained value of $t = 3.117$, exceeds the critical value at the 5% level ($t = 1.729$) and the critical value at the 1% level (2.539). We can therefore conclude that the actual answers were significantly higher than that expected by chance.

Appendix 7.8

WHAT FOR COLUMN ANSWERS: COHEN'S KAPPA ANALYSIS
(used for summarising subjects' answer agreement)

COLUMN HEADING "WHAT FOR?"

		S-S Interview Answer		Total
		Yes	No	
F.C. Answer	Yes	0.58 (a)	0.03 (b)	0.61 (p1)
	No	0.03 (c)	0.36 (d)	0.39 (q1)
	Total	0.61 (p2)	0.39 (q2)	1.00

Raw Agreement = (a+d); = 0.94

Occurrence Agreement = a/(a+b+c); = 0.91

Non-occurrence Agreement = d/(b+c+d); = 0.86

Cohen's Kappa = (a+d-p1p2-q1q2)/(1-p1p2-q1q2)
 = (0.94-0.37-0.15)/(1-0.37-0.15)
 = 0.42/0.48
 = **0.87**

Appendix 7.9

NEED COLUMN ANSWERS: COHEN'S KAPPA ANALYSIS

(used for summarising subjects' answer agreement)

COLUMN HEADING "NEED?"

		S-S Interview Answer		Total
		Yes	No	
F.C. Answer	Yes	0.53 (a)	0.11 (b)	0.64 (p1)
	No	0.01 (c)	0.34 (d)	0.35 (q1)
	Total	0.54 (p2)	0.45 (q2)	1.00

$$\text{Raw Agreement} = (a+d); = 0.87$$

$$\text{Occurrence Agreement} = a/(a+b+c); = 0.81$$

$$\text{Non-occurrence Agreement} = d/(b+c+d); = 0.74$$

$$\begin{aligned} \text{Cohen's Kappa} &= (a+d-p_1p_2-q_1q_2)/(1-p_1p_2-q_1q_2) \\ &= (0.87-0.35-0.16)/(1-0.35-0.16) \\ &= 0.36/0.49 \\ &= \mathbf{0.74} \end{aligned}$$

Appendix 8.1

Word-processing exercise 1

Use the different methods of highlighting text that you have been shown to carry out the following tasks.

The words in **Bold** are commands listed in the menus.

[1] Change the font style of the title "Berlin Factsheet", to **New York** and increase the font size to **18 Point**.

[2] Change the format of the title "Berlin Factsheet", to **Bold**.

[3] Change the format of the sub-heading "Pre 1945", to **Underline**.

[4] Change the phrase "1930s to "thirties". To do this highlight 1930's and type in the word "thirties".

[5] Change the format of all the year dates (e.g. "1945"), to **Italic**.

[6] Move the location of the fourth paragraph to the end of the document. To do this highlight all of the fourth paragraph (including the blank line above it), and select **Cut** from the Edit menu. Then position the insertion point at the desired location and selecting **Paste** from the Edit menu.

[7] Change all the occurrences of the word "Russian" to "Soviet". To do this select the **Replace** command from the Edit menu and type in the word that is to be replaced and the replacement word in the appropriate boxes. Then click the "Find Next" button to begin the search.

Appendix 8.2

Word-processing exercise 2

Please perform the following tasks on the document "Berlin Factsheet". Remember that you can always undo your last task by selecting **Undo** from the Edit menu.

[1] Use the Tab key on the keyboard (->/) to indent the beginning of each paragraph. To do this position the insertion point at the desired location then press the Tab key

[2] Highlight the whole document using **Select All** from the Edit menu and click the double spacing icon in the ruler

[3] Whilst the whole document is still highlighted (use **Select All** again if it is not), click on the full justification icon in the ruler

[4] Highlight the title "Berlin Factsheet" and select **Cut** from the Edit menu. Select **Header** from the View menu and then **Paste** from the Edit menu. Highlight the title and click on the centre justification icon, (N.B. if the header window does not have its own ruler you must give it one by selecting **Ruler** from the View menu).

[5] Select Footer from the View menu and click the page number icon and then the centre justification icon in the ruler of the Footer window, (N.B. if the footer window does not have its own ruler you must give it one by selecting **Ruler** from the View menu).Close the Footer window by clicking on its "close box".

[6] Place the insertion point before the fourth paragraph and select **Page Break** from the Insert menu.

[7] Select **Print Preview** from the File menu in order to see what your printed document is going to look like.

[8] Select Close from the File menu and then click Yes in the Save dialog box.

<p>IMPORTANT :- Headers and Footers do not show in the window of your document as you are working in it; if you want to check that they are correct use the Print Preview command.</p>
--

Appendix 8.3

Word-processing exercise 3

Exercise 3

Please perform the following tasks on the document "Berlin Factsheet". Remember that you can always undo your last task by selecting **Undo** from the Edit menu.

[1] Move the insertion point to the beginning of the document. Select **Spelling** from the Tools menu in order to check that all words are correct in your document. If you don't wish to change the selected word click Ignore in the dialog box. If you do wish to change the selected word to the suggestion given click Change in the dialog box. When the computer has finished the spelling check click OK.

N.B. If you wish to check the spelling of one individual word, highlight that word and then select Spelling from the Tools menu.

[2] When the spelling check is completed highlight any individual word and select **Thesaurus** from the Tools menu. If you want to replace the selected word with the computer alternative, click on that alternative in the dialog box and then click Replace. Click Close when you have finished with the Thesaurus.

[3] Move the insertion point to the beginning of the document. Select **Word Count** from the tools menu. Click Count in the dialog box to see how many words are in your document. Close when you have finished.

Appendix 8.4

"Microsoft WORD 5.0" standard F.C.

Computer No: _____
Sex: _____
Age: _____
Date: _____

The following feature checklist asks you 5 questions about commands that you may or may not have used during the psychology summer school word-processing class. Please read the following instructions carefully.

Instructions

In the column headed **Existed?** please put a tick (✓) against each command if you think that command actually existed. If you do not think that this command existed please put a cross (✗) against it. If you are not sure or don't know whether that command existed please put a question mark (?) against that command.

In the column headed **Used?** please put a tick (✓) against that command if you used it. If you don't think you used that command please put a cross (✗) against it. If you are not sure or don't know whether you used that command please put a question mark (?) against that command.

In the column headed **How often?** please put down a number representing the approximate amount of times you think you actually used that command during the class; e.g. (4).

In the column headed **What for?** please put a tick (✓) against that command if you know what that command actually does. If you do not know what that command does please put a cross (✗) against that command. If you are not sure or don't know what that command does please put a question mark (?) against that command.

In the column headed **Need?** please put a tick (✓) against that command if you have ever needed to use that command. If you have never needed to use that command please put a cross (✗) against that command. If you are not sure or don't know whether you have ever needed to use that command please put a question mark (?) against that command.

Please answer each question for every task description in the list by putting the appropriate answer in the relevant space.

Appendix 8.4 (cont.)

File	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					
Quit					

Edit	Existed?	Used?	How often?	What for?	Need?
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					

Appendix 8.4 (cont.)

View	Existed?	Used?	How often?	What for?	Need?
Normal					
Outline					
Page Layout					
Ribbon					
Ruler					
Hide/Show					
Header					
Footer					

Insert	Existed?	Used?	How often?	What for?	Need?
Page break					
Table...					
Date					
File...					
Picture...					

Format	Existed?	Used?	How often?	What for?	Need?
Character...					
Paragraph...					
Section...					
Document...					
Border...					
Table Cells...					
Table Layout...					
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					

Appendix 8.4 (cont.)

Font	Existed?	Used?	How often?	What for?	Need?
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Athens					
Cairo					
Chicago					
Courier					
Geneva					
Helvetica					
London					
Los Angeles					
Monaco					
New York					
Palatino					
San Francisco					
Symbol					
Times					
Venice					
Zapf Dingbats					

Appendix 8.4 (cont.)

Tools	Existed?	Used?	How often?	What for?	Need?
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					
Repaginate Now					
Preferences...					
Commands...					

Window...	Existed?	Used?	How often?	What for?	Need?
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 8.5

"Microsoft WORD 5.0" semantic F.C.

Computer No: _____
Sex: _____
Age: _____
Date: _____

The following feature checklist asks you 4 questions about tasks that you may or may not have performed during the psychology summer school word-processing class. Please read the following instructions carefully.

Instructions

In the column headed **Existed?** please put a tick (✓) against each command if you think that command actually existed. If you do not think that this command existed please put a cross (✗) against it. If you are not sure or don't know whether that command existed please put a question mark (?) against that command.

In the column headed **Used?** please put a tick (✓) against that command if you used it. If you don't think you used that command please put a cross (✗) against it. If you are not sure or don't know whether you used that command please put a question mark (?) against that command.

In the column headed **How often?** please put down a number representing the approximate amount of times you think you actually used that command during the class; e.g. (4).

In the column headed **What for?** please put a tick (✓) against that command if you know what that command actually does. If you do not know what that command does please put a cross (✗) against that command. If you are not sure or don't know what that command does please put a question mark (?) against that command.

In the column headed **Need?** please put a tick (✓) against that command if you have ever needed to use that command. If you have never needed to use that command please put a cross (✗) against that command. If you are not sure or don't know whether you have ever needed to use that command please put a question mark (?) against that command.

Please answer each question for every task description in the list by putting the appropriate answer in the relevant space.

Appendix 8.5 (cont.)

TASK DESCRIPTION	Perform?	How often?	Possible?	Need?
Create a file				
Begin to edit a file				
Finish editing a file				
Keep the changes you have made				
Keep the changes you have made in a different file				
Show what the document will look like on paper				
Show the layout of the file to be printed				
Put the document on paper				
Combine information from another file				
Finish editing and leave program completely				
Cancel/Repeat last editing done				
Remove piece of text to clipboard				
Duplicate piece of text and move to clipboard				
Insert text that was removed or duplicated				
Erase highlighted text				
Highlight everything in the document				
Search for a specified piece of text				
Search for and alter a specified piece of text				
Move to a specified page				
Store a frequently used piece of text/document				
Display the document in its usual view				
Add and arrange underlying structure of document				
Show where the paper page edges will be				
Show/Hide tools for formatting, font styles, etc.				
Show/Hide tools for spacing, indenting, etc.				
Cover/Reveal text markers				
Edit the top border to appear on each page of the printed version of the document				
Edit the bottom border to appear on each page of the printed version of the document				

Appendix 8.5 (cont.)

TASK DESCRIPTION	Perform?	How often?	Possible?	Need?
Begin a new page				
Insert a figure with columns and rows for entering data into				
Insert the current date				
Insert contents of another document at this point				
Insert selected drawing at this location				
Display options for different text styles				
Change formats/measurements of a group of sentences				
Change format/measurements of part of file				
Change format/measurements of whole file				
Add/Remove boundary lines around specified text				
Change size of rows/columns in table				
Add/Remove rows or columns in table				
Display list of stored format patterns				
Change to this kind of format style				
Change to this kind of format style				
<i>Change to this kind of format style</i>				
<u>Change to this kind of format style</u>				
Make print this size				
Make print this size				
Make print this size				
Make print this size				
Make print this size				
Make print this size				
Make the print look like this				
Make the print look like this (abcdefghi)				
Make the print look like this				
Make the print look like this				
Make the print look like this				

Appendix 8.5 (cont.)

TASK DESCRIPTION	Perform?	How often?	Possible?	Need?
Make the print look like this				
Make the print look like this				
Make the print look like this				
Make the print look like this				
Make the print look like this				
Make the print look like this				
Make the print look like this				
Make the print look like this (αβχδεφγη)				
Make the print look like this				
Make the print look like this (☉☉*** ***)				
Check that all words in the document are correct				
Check that the word order and syntax in the document is correct				
Look at a list of alternative words that could be used to replace the word selected				
Calculate amount of character groups in document				
Assign sequence numbers to paragraphs				
Rearrange text alphabetically or numerically				
Recalculate page breaks				
Display available size parameter options				
Display all actions available for inclusion or exclusion in menus				
Give aid to user				
Display text removed or duplicated for later use				
Create a separate view of current document that can be accessed				
Show current document/bring document to top				

Appendix 8.6

"MICROSOFT WORD 5.0" WORD-PROCESSING INTERVIEW

Subject No: _____

Date: _____

Sex: _____

Age: _____

Appendix 8.6 (cont.)

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

Appendix 8.6 (cont.)

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

PROMPTS

COMMAND:
Do you know what this command is for?
Do you think you would ever have a need to use this command?

Appendix 8.7

"Microsoft WORD 5.0" standard F.C. (abbreviated)

Name: _____

Sex: _____

Age: _____

Date: _____

The following feature checklist asks you 5 questions about commands that you may or may not have used during the psychology summer school word-processing class. Please read the following instructions carefully.

Instructions

In the column headed **Used?** please put a tick (✓) against that command if you used it. If you don't think you used that command please put a cross (✗) against it. If you are not sure or don't know whether you used that command please put a question mark (?) against that command.

In the column headed **How often?** please put down a number representing the approximate amount of times you think you actually used that command during the class; e.g. (4).

Please answer each question for every task description in the list by putting the appropriate answer in the relevant space.

Appendix 8.7 (cont.)

File Used ? How often?

File	Used ?	How often?
New		
Open...		
Close		
Save		
Save As...		
Print Preview...		
Page Setup...		
Print...		
Print Merge...		
Quit		

Edit Used ? How often?

Edit	Used ?	How often?
Undo/Redo		
Cut		
Copy		
Paste		
Clear		
Select All		
Find...		
Replace...		
Go To...		
Glossary...		

Appendix 8.7 (cont.)

View Used ? How often?

View	Used ?	How often?
Normal		
Outline		
Page Layout		
Ribbon		
Ruler		
Hide/Show		
Header		
Footer		

Insert Used ? How often?

Insert	Used ?	How often?
Page break		
Table...		
Date		
File...		
Picture...		

Format Used ? How often?

Format	Used ?	How often?
Character...		
Paragraph...		
Section...		
Document...		
Border...		
Table Cells...		
Table Layout...		
Style...		
Plain Text		
Bold		
<i>Italic</i>		
<u>Underline</u>		

Appendix 8.7 (cont.)

Font	Used ?	How often?
9 Point		
10 Point		
12 Point		
14 Point		
18 Point		
24 Point		
Athens		
Cairo		
Chicago		
Courier		
Geneva		
Helvetica		
London		
Los Angeles		
Monaco		
New York		
Palatino		
San Francisco		
Symbol		
Times		
Venice		
Zapf Dingbats		

Appendix 8.7 (cont.)

Tools	Used ?	How often?
Spelling...		
Grammar...		
Thesaurus...		
Word Count...		
Renumber...		
Sort		
Repaginate Now		
Preferences...		
Commands...		

Window...	Used ?	How often?
Help		
Show Clipboard		
New Window		
Untitled 1		

Appendix 8.8

"Microsoft WORD 5.0" combined F.C.

Name: _____

Sex: _____

Age: _____

Date: _____

The following feature checklist asks you 5 questions about commands that you may or may not have used during the psychology summer school word-processing class. Each command is also followed by a description of its function. Please read the following instructions carefully.

Instructions

In the column headed **Used?** please put a tick (✓) against that command if you used it. If you don't think you used that command please put a cross (✗) against it. If you are not sure or don't know whether you used that command please put a question mark (?) against that command.

In the column headed **How often?** please put down a number representing the approximate amount of times you think you actually used that command during the class; e.g. (4).

Please answer each question for every task description in the list by putting the appropriate answer in the relevant space.

Appendix 8.8 (cont.)

File	Used ?	How often?
New Create a file		
Open... Begin to edit a file		
Close Finish editing a file		
Save Keep the changes you have made		
Save As... Keep the changes you have made in a different file		
Print Preview... Show what the document will look like on paper		
Page Setup... Show the layout of the file to be printed		
Print... Put the document on paper		
Print Merge... Combine information from another file		
Quit Finish editing and leave program completely		

Appendix 8.8 (cont.)

Edit Used ? How often?

Undo/Redo Cancel/Repeat last editing done		
Cut Remove piece of text to clipboard		
Copy Duplicate piece of text and move to clipboard		
Paste Insert text that was removed or duplicated		
Clear Erase highlighted text		
Select All Highlight everything in the document		
Find... Search for a specified piece of text		
Replace... Search for and alter a specified piece of text		
Go To... Move to a specified page		
Glossary... Store a frequently used piece of text/document		

Appendix 8.8 (cont.)

View	Used ?	How often?
Normal Display the document in its usual view		
Outline Add and arrange underlying structure of document		
Page Layout Show where the paper page edges will be		
Ribbon Show/Hide tools for formatting, font styles, etc.		
Ruler Show/Hide tools for spacing, indenting, etc.		
Hide/Show Cover/Reveal text markers		
Header Edit the top border to appear on each page of the printed version of the document		
Footer Edit the bottom border to appear on each page of the printed version of the document		

Insert	Used ?	How often?
Page break Begin a new page		
Table... Insert a figure with columns and rows for entering data into		
Date Insert the current date		
File... Insert contents of another document at this point		
Picture... Insert selected drawing at this location		

Appendix 8.8 (cont.)

Format	Used ?	How often?
Character... Display options for different text styles		
Paragraph... Change formats/measurements of a group of sentences		
Section... Change format/measurements of part of file		
Document... Change format/measurements of whole file		
Border... Add/Remove boundary lines around specified text		
Table Cells... Change size of rows/columns in table		
Table Layout... Add/Remove rows or columns in table		
Style... Display list of stored format patterns		
Plain Text Change to this kind of format style		
Bold Change to this kind of format style		
<i>Italic</i> Change to this kind of format style		
<u>Underline</u> Change to this kind of format style		

Appendix 8.8 (cont.)

Font	Used ?	How often?
9 Point Make print this size		
10 Point Make print this size		
12 Point Make print this size		
14 Point Make print this size		
18 Point Make print this size		
24 Point Make print this size		
Athens Make the print look like this		
Cairo Make the print look like this (abcdefghi)		
Chicago Make the print look like this		
Courier Make the print look like this		
Geneva Make the print look like this		
Helvetica Make the print look like this		
London Make the print look like this		
Los Angeles Make the print look like this		
Monaco Make the print look like this		
New York Make the print look like this		

Appendix 8.8 (cont.)

Font (cont.)

Used ?

How often?

<p>Palatino Make the print look like this</p>		
<p>San Francisco Make the print look like this</p>		
<p>Symbol Make the print look like this (αβχδεφγηι)</p>		
<p>Times Make the print look like this</p>		
<p>Denice Make the print look like this</p>		
<p>Zapf Dingbats Make the print look like this (☉☺*~*~*~*~*~)</p>		

Appendix 8.8 (cont.)

Tools

Used ?

How often?

Spelling... Check that all words in the document are correct		
Grammar... Check that the word order and syntax in the document is correct		
Thesaurus... Look at a list of alternative words that could be used to replace the word selected		
Word Count... Calculate amount of character groups in document		
Renumber... Assign sequence numbers to paragraphs		
Sort Rearrange text alphabetically or numerically		
Repaginate Now Recalculate page breaks		
Preferences... Display available size parameter options		
Commands... Display all actions available for inclusion or exclusion in menus		

Window...

Used ?

How often?

Help Give aid to user		
Show Clipboard Display text removed or duplicated for later use		
New Window Create a separate view of current document that can be accessed		
Untitled 1 Show current document/bring document to top		

Appendix 8.9

Need scale

Do you think that you would ever have a need to use this command? Using the following scale, please estimate how much of a need you think you might have:-

1
(strong need)

2

3
(average need)

4

5
(no need at all)

Appendix 8.10

WHAT FOR COLUMN ANSWERS: COHEN'S KAPPA ANALYSIS

(used for summarising subjects' answer agreement)

COLUMN HEADING "WHAT FOR?"

		S-S Interview Answer		Total
		Yes	No	
F.C. Answer	Yes	0.56 (a)	0.07 (b)	0.63 (p1)
	No	0.03 (c)	0.34 (d)	0.37 (q1)
	Total	0.59 (p2)	0.41 (q2)	1.00

$$\text{Raw Agreement} = (a+d); = 0.90$$

$$\text{Occurrence Agreement} = a + (a+b+c); = 0.66$$

$$\text{Non-occurrence Agreement} = d + (b+c+d); = 0.44$$

$$\text{Cohen's Kappa} = (a+d-p_1p_2-q_1q_2) \div (1-p_1p_2-q_1q_2)$$

$$= (0.90-0.37-0.15) \div (1-0.37-0.15)$$

$$= 0.38 \div 0.48$$

$$= \underline{\underline{0.79}}$$

Appendix 8.11

NEED COLUMN ANSWERS: COHEN'S KAPPA ANALYSIS
(used for summarising subjects' answer agreement)

COLUMN HEADING "NEED?"

		S-S Interview Answer		Total
		Yes	No	
F.C. Answer	Yes	0.39 (a)	0.05 (b)	0.44 (p1)
	No	0.13 (c)	0.43 (d)	0.56 (q1)
	Total	0.52 (p2)	0.48 (q2)	1.00

Raw Agreement = (a+d); = 0.82

Occurrence Agreement = a ÷ (a+b+c); = 0.68

Non-occurrence Agreement = d ÷ (b+c+d); = 0.71

Cohen's Kappa = (a+d-p1p2-q1q2) ÷ (1-p1p2-q1q2)
 = (0.82-0.23-0.27) ÷ (1-0.23-0.27)
 = 0.32 ÷ 0.50
 = **0.64**

Appendix 8.12

COHEN'S KAPPA ANALYSIS (RELIABILITY)
(used for summarising subjects' answer agreement)

RELIABILITY (Standard F.C. at Time A and Time B)

		Time B		Total
		Yes	No	
Time A	Yes	0.31 (a)	0.04 (b)	0.35 (p1)
	No	0.11 (c)	0.55 (d)	0.66 (q1)
	Total	0.42 (p2)	0.59 (q2)	1.00

Raw Agreement = (a+d); = 0.86

Occurrence Agreement = a + (a+b+c); = 0.67

Non-occurrence Agreement = d + (b+c+d); = 0.79

Cohen's Kappa = $(a+d-p_1p_2-q_1q_2) \div (1-p_1p_2-q_1q_2)$
 = $(0.86 - 0.15 - 0.39) \div (1 - 0.15 - 0.39)$
 = $0.32 \div 0.46$
 = **0.70**

Appendix 8.13

Subjects' answers to the F.C. question "do you know what this command actually does?"

COMMAND NAME	Number of Subjects that answered "Yes"
New	8
Open...	15
Close	14
Save	15
Save As...	8
Print Preview...	13
Page Setup...	7
Print...	15
Print Merge...	2
Quit	14
Undo/Redo	12
Cut	15
Copy	13
Paste	15
Clear	5
Select All	12
Find...	8
Replace...	12
Go To...	3
Glossary...	4
Normal	2
Outline	2
Page Layout	8
Ribbon	0
Ruler	15
Hide/Show	4
Header	14
Footer	14
Page break	14
Table...	4

Appendix 8.13 (cont.)

COMMAND NAME	Number of Subjects that answered "Yes"
Date	7
File...	5
Picture...	4
Character...	1
Paragraph...	5
Section...	1
Document...	3
Border...	3
Table Cells...	0
Table Layout...	1
Style...	2
Plain Text	11
Bold	14
<i>Italic</i>	15
<u>Underline</u>	14
9 Point	11
10 Point	11
12 Point	11
14 Point	11
18 Point	12
24 Point	11
Athens	11
Cairo	11
Chicago	11
Courier	10
Geneva	11
Helvetica	9
London	11
Los Angeles	11
Monaco	10
New York	12
Palatino	10

Appendix 8.13 (cont.)

COMMAND NAME	Number of Subjects that answered "Yes"
San Francisco	11
Symbol	9
Times	9
Denice	10
Zapf Dingbats	8
Spelling...	15
Grammar...	10
Thesaurus...	15
Word Count...	15
Renumber...	4
Sort	3
Repaginate Now	2
Preferences...	2
Commands...	2
Help	8
Show Clipboard	4
New Window	5
Untitled 1	4

Appendix 8.14

Subjects' answers to the F.C. question "do you think you would ever have a need to use this command?"

COMMAND NAME	Number of Subjects that answered "Yes"
New	5
Open...	11
Close	10
Save	15
Save As...	6
Print Preview...	12
Page Setup...	3
Print...	5
Print Merge...	1
Quit	14
Undo/Redo	9
Cut	13
Copy	7
Paste	13
Clear	3
Select All	10
Find...	4
Replace...	8
Go To...	0
Glossary...	1
Normal	3
Outline	2
Page Layout	7
Ribbon	1
Ruler	15
Hide/Show	3
Header	14
Footer	14
Page break	12
Table...	2
Date	2

Appendix 8.14 (cont.)

COMMAND NAME	Number of Subjects that answered "Yes"
File...	2
Picture...	1
Character...	1
Paragraph...	5
Section...	2
Document...	2
Border...	2
Table Cells...	1
Table Layout...	1
Style...	3
Plain Text	9
Bold	14
<i>Italic</i>	14
<u>Underline</u>	14
9 Point	1
10 Point	1
12 Point	4
14 Point	4
18 Point	11
24 Point	4
Athens	2
Cairo	1
Chicago	1
Courier	1
Geneva	3
Helvetica	1
London	2
Los Angeles	2
Monaco	0
New York	9
Palatino	1
San Francisco	3

Appendix 8.14 (cont.)

COMMAND NAME	Number of Subjects that answered "Yes"
Symbol	1
Times	3
Uenice	1
Zapf Dingbats	2
Spelling...	15
Grammar...	4
Thesaurus...	14
Word Count...	2
Renumber...	1
Sort	1
Repaginate Now	0
Preferences...	1
Commands...	3
Help	3
Show Clipboard	2
New Window	2
Untitled 1	2

Appendix 8.15

Standard F.C. (1st, time A) vs. semantic F.C. (1st, time A):
T-test for independent samples (separate variance)

Standard F.C.

$$N_1 = 8$$

$$M_1 = 72.4$$

$$SD_1 = 4.4$$

Semantic F.C.

$$N_2 = 7$$

$$M_2 = 51.4$$

$$SD_2 = 6.3$$

$$t = \frac{M_1 - M_2}{\sqrt{([SD_1 \times SD_1]/N_1 + [SD_2 \times SD_2]/N_2)}}$$

$$t = \frac{21}{\sqrt{(19.36/8 + 36.69/7)}}$$

$$t = \frac{21}{\sqrt{(2.42 + 5.24)}}$$

$$t = \frac{21}{2.768}$$

$$t = 7.588$$

Appendix 9.1

Computer use questionnaire

From: Eddie Edgerton
To:
Date: 14th September 1993

Dear _____,

I would be very grateful if you could give me some information about the word-processing package that you normally use. I need the information for a study that I might be doing as the final part of PhD.

What I would like to know is:-

(1) the type of computer that you normally use when you are word-processing, e.g. "Apple Macintosh", "IBM PC", etc.

(2) the type of word-processing package that you normally use, e.g. "Microsoft WORD 5.0", "WORD for Windows", "Claris Works", "MacWrite", etc.

(3) how often do you normally use this word-processing package?

- | | |
|------------------------|--------------------------|
| every day | <input type="checkbox"/> |
| every 2-3 days | <input type="checkbox"/> |
| once a week | <input type="checkbox"/> |
| once a month | <input type="checkbox"/> |
| less than once a month | <input type="checkbox"/> |

I would be grateful if you could return this completed form to me as soon as possible;
Room No. S605, Adam Smith Building.

Thanks, Eddie

Appendix 9.2

"Microsoft WORD 5.0" F.C.

NAME: _____ AGE: _____ DATE: _____

The following feature checklist asks you various questions about the menu commands in the word-processing package "Microsoft WORD 5.0". Please read the following instructions carefully before attempting to complete the feature checklist. If you are unsure about what the questions mean or how to fill in your answers, please don't hesitate to ask me.

INSTRUCTIONS

Column 1: Existed? - "Did you know this command existed?"

✓ = "Yes" ✗ = "No" ? = "Unsure"

Column 2: Used? - "Have you ever used this command?"

✓ = "Yes" ✗ = "No" ? = "Unsure"

Column 3: How often? "How often do you use this command (approximately)?"

0 = never used it

1 = less than once a month

2 = once a month

3 = once a week

4 = every 2-3 days

5 = every day

Column 4: What for? - "Do you know what this command does?"

✓ = "Yes" ✗ = "No" ? = "Unsure"

Column 5: Need? - "How often do you have any need for this command?"

0 = no need at all

1 = less than once a month

2 = once a month

3 = once a week

4 = every 2-3 days

5 = every day

[N.B. You can only answer this question if you know what the command does]

Please answer for every command in each column by putting the appropriate answer in the relevant space.

Appendix 9.2 (cont.)

File	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					
Quit					

Edit	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					

View	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Normal					
Outline					
Page Layout					
Ribbon					
Ruler					
Hide/Show					
Header					
Footer					

Appendix 9.2 (cont.)

Insert	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Page Break					
Table...					
Date					
File...					
Picture...					

Format	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Character...					
Paragraph...					
Section...					
Document...					
Border...					
Table Cells...					
Table Layout...					
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					

Appendix 9.2 (cont.)

Font	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Athens					
Cairo					
Chicago					
Courier					
Geneva					
Helvetica					
London					
Los Angeles					
Monaco					
New York					
Palatino					
San Francisco					
Symbol					
Times					
Venice					
Zapf Dingbats					

Appendix 9.2 (cont.)

Tools	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					
Repaginate Now					
Preferences...					
Commands...					

Window...	1 Existed?	2 Used?	3 How often?	4 What for?	5 Need?
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 9.3

FC assessment form

Dear _____,

You may (or may not) remember that I asked you to take fill in an instrument called a feature checklist on "Microsoft WORD 5.0", a couple of months ago. As a quick refresher, I have attached a copy of the feature checklist that I sent to you (**don't worry, I'm not asking you to fill it in again!**).

What I would like you to do for me, is to try and cast your mind back to when you filled in the feature checklist (all those months ago) and try and answer the following questions. I know you might find this hard to do, but I would be very grateful if you could attempt this as accurately as possible. I'm afraid I can't offer you any money, only eternal thanks and perhaps a drink once my Ph.d is accepted. Here are the questions:

(1) Overall, how difficult/easy did you find it to fill in the feature checklist? (Please circle).

<u>Very difficult</u>					<u>Very easy</u>
1	2	3	4	5	

(2a) Did you find some columns (i.e. questions) harder to answer than others? Please rate the columns (1 = easiest to answer, 5 = hardest to answer; N.B. if you think some columns were equal then give them the same number).

Existed?
Used?
How Often?
What For?
Need?

(3) How long did it take you (approximately) to fill in the feature checklist? (Please tick).
less than 10 mins
10-20 mins
20-30 mins
more than 30 mins

(4) Do you have any additional comments you would like to make about the feature checklist?
Please state: _____

Please send this completed form back to me in the enclosed self-addressed envelope.
Thanks again,
Eddie

Appendix 9.4

"Microsoft WORD 5.0" F.C. - system bug detection answers

Explanation of categorisations:-

Existed?-

0, 1, or 2	= very few users knew this command existed
3, 4, or 5	= some users knew this command existed
6, 7, or 8	= most users knew this command existed

Used?-

0, 1, or 2	= very few users have ever used this command
3, 4, or 5	= some users have used this command at least once
6, 7, or 8	= most users have used this command at least once

How often?-

$>0 < 2$	= command is used less than once a month
$\geq 2 < 3$	= command is used more than once a month but less than once a week
$\geq 3 \leq 5$	= command is used at least once a week

What for?-

0, 1, or 2	= very few users know what this command does
3, 4, or 5	= some users know what this command does
6, 7, or 8	= most users know what this command does

Need?-

$>0 < 2$	= command is needed by users less than once a month
$\geq 2 < 3$	= command is needed more than once a month but less than once a week
$\geq 3 \leq 5$	= command is needed by users at least once a week

Appendix 9.4 (cont.)

	Existed?	Used?	How often?	What For?	Need?
New	8	8	4.50	8	4.50
Open...	8	8	4.62	8	4.75
Close	8	8	4.62	8	4.62
Save	8	8	4.75	8	4.75
Save As...	8	8	4.12	8	4.12
Print Preview...	8	8	3.38	8	3.25
Page Setup...	8	8	2.88	8	3.25
Print...	8	8	4.62	8	4.50
Print Merge...	8	1	0.25	2	1.50
Quit	8	8	4.75	8	4.75
Undo/Redo	8	8	2.75	8	2.88
Cut	8	8	3.62	8	3.75
Copy	8	8	4.12	8	4.12
Paste	8	8	4.12	8	4.12
Clear	7	6	1.38	6	1.83
Select All	7	7	3.25	8	3.38
Find...	8	8	2.38	8	2.62
Replace...	8	6	1.38	6	1.83
Go To...	6	4	1.00	4	1.75
Glossary...	5	1	0.50	4	1.50
Normal	8	5	2.25	6	3.00
Outline	7	4	1.38	5	1.60
Page Layout	8	7	2.25	7	2.57
Ribbon	5	3	1.00	4	2.00
Ruler	8	8	2.38	8	2.25
Hide/Show	8	6	1.62	6	2.17
Header	8	8	2.00	8	1.75
Footer	8	8	2.38	8	2.25
Page Break	8	7	2.12	7	2.57
Table...	7	4	1.00	6	1.50
Date	5	3	0.62	4	1.25
File...	8	5	0.88	6	1.17
Picture...	7	7	2.12	7	2.29
Character...	7	5	1.38	7	1.57
Paragraph...	6	3	0.62	5	1.00
Section...	6	3	0.75	4	1.50
Document...	8	6	1.25	7	1.57

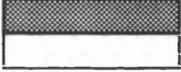
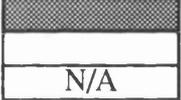
Appendix 9.4 (cont.)

	Existed?	Used?	How Often?	What For?	Need?
Border...	7	4	0.50	5	0.40
Table Cells...	5	5	1.12	5	1.60
Table Layout...	5	4	1.00	4	2.00
Style...	8	6	1.88	7	2.14
Plain Text	8	7	4.38	8	4.50
Bold	8	8	4.00	7	4.29
<i>Italic</i>	8	8	3.50	8	3.50
<u>Underline</u>	8	8	3.75	8	3.88
9 Point	8	7	1.50	8	1.61
10 Point	8	7	2.00	8	2.12
12 Point	8	8	4.38	8	4.50
14 Point	8	8	1.75	8	1.62
18 Point	8	8	1.12	8	1.12
24 Point	8	8	1.12	8	1.12
Chicago	8	7	1.38	8	1.38
Courier	8	6	1.50	8	1.50
Geneva	8	7	2.25	8	2.50
Helvetica	8	6	0.88	8	1.00
Monaco	8	5	0.88	8	1.00
New York	8	6	1.12	8	1.25
Palatino	8	6	2.25	8	2.38
Symbol	8	3	0.25	8	0.25
Times	8	8	2.12	8	2.50
Spelling...	8	8	3.38	8	3.12
Grammar...	7	3	0.62	7	0.86
Thesaurus...	8	5	1.25	7	1.43
Word Count...	8	7	2.00	8	1.88
Renumber...	6	0	0.12	4	0.25
Sort	8	3	1.58	4	3.00
Repaginate Now	7	2	0.75	3	1.25
Preferences...	7	5	0.88	5	1.20
Commands...	7	3	0.62	3	1.33
Help	8	5	1.12	8	1.12
Show Clipboard	8	8	1.88	8	1.62
New Window	7	6	1.25	7	1.57
Untitled 1	7	7	1.88	7	1.57

Appendix 9.5

"Microsoft WORD 5.0" F.C. - individual subject answers

Explanation of shading categories:-

<u>FC Column</u>	<u>Shading</u>	<u>Meaning</u>
Existed?		= subject did not know this command existed = subject did know this command existed
Used?		= subject has never used this command = subject has used this command at least once
How Often?		= subject uses this command less than once a month = subject uses this command at least once a month
What for?		= subject does not know what this command does = subject does know what this command does
Need?		= subject needs to use this command less than once a month = subject needs to use this command at least once a month = not applicable, i.e. subject is unaware of command function

Appendix 9.5 (cont.)

Subject No. 1	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					N/A
Glossary...					N/A
Normal					
Outline					
Page Layout					
Ribbon					N/A
Ruler					
Hide/Show					
Header					
Footer					
Page Break					
Table...					
Date					N/A
File...					
Picture...					
Character...					
Paragraph...					N/A
Section...					N/A
Document...					N/A

Appendix 9.5 (cont.)

Subject No. 1	Existed?	Used?	How often?	What for?	Need?
Border...					N/A
Table Cells...					N/A
Table Layout...					N/A
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					N/A
Repaginate Now					N/A
Preferences...					N/A
Commands...					N/A
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 9.5 (cont.)

Subject No. 2	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					N/A
Go To...					N/A
Glossary...					N/A
Normal					
Outline					
Page Layout					
Ribbon					N/A
Ruler					
Hide/Show					N/A
Header					
Footer					
Page Break					
Table...					N/A
Date					N/A
File...					
Picture...					
Character...					N/A
Paragraph...					
Section...					N/A
Document...					

Appendix 9.5 (cont.)

Subject No. 2	Existed?	Used?	How often?	What for?	Need?
Border...					
Table Cells...					N/A
Table Layout...					N/A
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					N/A
Sort					N/A
Repaginate Now					
Preferences...					
Commands...					
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 9.5 (cont.)

Subject No. 3	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					N/A
Select All					
Find...					
Replace...					N/A
Go To...					N/A
Glossary...					N/A
Normal					N/A
Outline					N/A
Page Layout					N/A
Ribbon					
Ruler					
Hide/Show					N/A
Header					
Footer					
Page Break					
Table...					
Date					
File...					
Picture...					
Character...					
Paragraph...					
Section...					
Document...					

Appendix 9.5 (cont.)

Subject No. 3	Existed?	Used?	How often?	What for?	Need?
Border...					N/A
Table Cells...					
Table Layout...					
Style...					N/A
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					N/A
Sort					N/A
Repaginate Now					N/A
Preferences...					
Commands...					N/A
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 9.5 (cont.)

Subject No. 4	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					N/A
Select All					
Find...					
Replace...					
Go To...					N/A
Glossary...					N/A
Normal					N/A
Outline					N/A
Page Layout					
Ribbon					N/A
Ruler					
Hide/Show					
Header					
Footer					
Page Break					
Table...					
Date					N/A
File...					N/A
Picture...					N/A
Character...					
Paragraph...					
Section...					
Document...					

Appendix 9.5 (cont.)

Subject No. 4	Existed?	Used?	How often?	What for?	Need?
Border...					N/A
Table Cells...					
Table Layout...					N/A
Style...					
Plain Text					
Bold					
<i>Italic</i>					
Underline					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					N/A
Word Count...					
Renumber...					N/A
Sort					
Repaginate Now					
Preferences...					N/A
Commands...					N/A
Help					
Show Clipboard					
New Window					N/A
Untitled 1					

Appendix 9.5 (cont.)

Subject No. 5	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					
Normal					
Outline					
Page Layout					
Ribbon					
Ruler					
Hide/Show					
Header					
Footer					
Page Break					
Table...					
Date					
File...					
Picture...					
Character...					
Paragraph...					
Section...					
Document...					

Appendix 9.5 (cont.)

Subject No. 5	Existed?	Used?	How often?	What for?	Need?
Border...					
Table Cells...					
Table Layout...					
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					N/A
Repaginate Now					
Preferences...					N/A
Commands...					N/A
Help					
Show Clipboard					
New Window					
Untitled 1					

Appendix 9.5 (cont.)

Subject No. 6	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					
Normal					
Outline					N/A
Page Layout					
Ribbon					
Ruler					
Hide/Show					
Header					
Footer					
Page Break					
Table...					N/A
Date					N/A
File...					N/A
Picture...					
Character...					
Paragraph...					N/A
Section...					N/A
Document...					

Appendix 9.5 (cont.)

Subject No. 6	Existed?	Used?	How often?	What for?	Need?
Border...					
Table Cells...					N/A
Table Layout...					N/A
Style...					
Plain Text					
Bold					
<i>Italic</i>					
Underline					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					N/A
Thesaurus...					
Word Count...					
Renumber...					N/A
Sort					
Repaginate Now					N/A
Preferences...					
Commands...					N/A
Help					
Show Clipboard					
New Window					
Untitled 1					N/A

Appendix 9.5 (cont.)

Subject No. 7	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					
Normal					
Outline					
Page Layout					
Ribbon					N/A
Ruler					
Hide/Show					
Header					
Footer					
Page Break					
Table...					
Date					
File...					
Picture...					
Character...					
Paragraph...					
Section...					
Document...					

Appendix 9.5 (cont.)

Subject No. 7	Existed?	Used?	How often?	What for?	Need?
Border...					
Table Cells...					
Table Layout...					
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					
Repaginate Now					N/A
Preferences...					
Commands...					
Help					
Show Clipboard					
New Window					
Untitled 1					

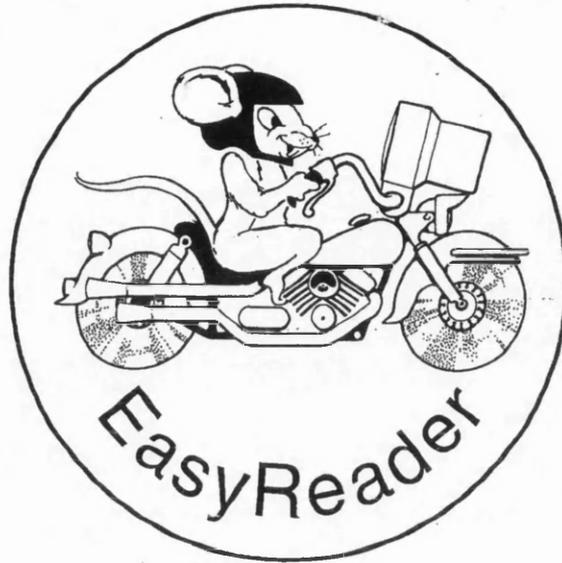
Appendix 9.5 (cont.)

Subject No. 8	Existed?	Used?	How often?	What for?	Need?
New					
Open...					
Close					
Save					
Save As...					
Print Preview...					
Page Setup...					
Print...					
Print Merge...					N/A
Quit					
Undo/Redo					
Cut					
Copy					
Paste					
Clear					
Select All					
Find...					
Replace...					
Go To...					
Glossary...					
Normal					
Outline					
Page Layout					
Ribbon					
Ruler					
Hide/Show					
Header					
Footer					
Page Break					N/A
Table...					
Date					
File...					
Picture...					
Character...					
Paragraph...					N/A
Section...					N/A
Document...					

Appendix 9.5 (cont.)

Subject No. 8	Existed?	Used?	How often?	What for?	Need?
Border...					
Table Cells...					
Table Layout...					
Style...					
Plain Text					
Bold					
<i>Italic</i>					
<u>Underline</u>					
9 Point					
10 Point					
12 Point					
14 Point					
18 Point					
24 Point					
Chicago					
Courier					
Geneva					
Helvetica					
Monaco					
New York					
Palatino					
Symbol					
Times					
Spelling...					
Grammar...					
Thesaurus...					
Word Count...					
Renumber...					
Sort					
Repaginate Now					N/A
Preferences...					
Commands...					
Help					
Show Clipboard					
New Window					
Untitled 1					

EasyReader Questionnaire



This questionnaire is part of a monitoring programme aimed at assessing NAM's Electronic Distribution Module "*EasyReader*". The monitoring programme is part of a continuing commitment by the management of NAM TWE to introduce and maintain *EasyReader* in the most effective and efficient manner. It also provides the opportunity for users of *EasyReader* to give feedback about any aspect of *EasyReader* that they may wish to.

All questions on the questionnaire can be answered by either:

- circling the appropriate number on the scale shown.
- ticking the appropriate answer box.
- writing your answer in the space provided.

It is important that you answer all questions on the questionnaire, as clearly and as accurately as possible. If you have any problems or queries with the questionnaire please call 05920-63891/62566 and ask for Jan de Zeeuw.

The information gained from this questionnaire and other aspects of the monitoring programme will be used to give all users feedback about the implementation of *EasyReader*.

Thank you for your time and effort

Appendix 10.1 (cont.)

6. How easy/difficult did you find it to perform the following tasks? (Please circle).

	<u>Very difficult</u>				<u>Very easy</u>
	1	2	3	4	5
Start <i>EasyReader</i> (i.e. double clicking on icon)	1	2	3	4	5
Open <i>EasyReader</i> (i.e. selecting a book from the collection window)	1	2	3	4	5
Use the table of contents	1	2	3	4	5
Use the "go back" button	1	2	3	4	5
Use the table icons	1	2	3	4	5
Use the link icons	1	2	3	4	5
View drawings/illustrations	1	2	3	4	5
Print from <i>EasyReader</i> (including user feedback forms)	1	2	3	4	5
Exit from <i>EasyReader</i>	1	2	3	4	5
Create and use annotations	1	2	3	4	5
Use keywords to find documents	1	2	3	4	5
Swap out a CD	1	2	3	4	5
Use "on-line" help	1	2	3	4	5

7. Have you ever been unable to perform any of the following tasks? (Please tick).

	<u>Yes</u>	<u>No</u>	<u>Don't know</u>
Start <i>EasyReader</i> (i.e. double clicking on icon)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Open <i>EasyReader</i> (i.e. selecting a book from the collection window)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use the table of contents	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use the "go back" button	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use the table icons	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use the link icons	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
View drawings/illustrations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Print from <i>EasyReader</i> (including user feedback forms)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Exit from <i>EasyReader</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Create and use annotations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use keywords to find documents	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Swap out a CD	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use "on-line" help	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix 10.1 (cont.)

SYSTEM BUGS/ERROR RATES

13a. Have you ever experienced any problems/difficulties using *EasyReader*? (Please tick one).

Yes No Don't know

13b. If "Yes", what have been the most frequent problems/difficulties that you have experienced using *EasyReader*? (Please state in order of frequency i.e. most frequent first).

Problem 1: _____

Problem 2: _____

Problem 3: _____

13c. How serious have the problems listed above been? (Please circle).

	<u>Very Serious</u>					<u>Not at all serious</u>
Problem 1:	1	2	3	4	5	
Problem 2:	1	2	3	4	5	
Problem 3:	1	2	3	4	5	

14. Where do you normally obtain help from? (Please tick any that apply).

EasyReader user guide

Other users

TWE/42 personnel

On-screen help system

EasyReader quick reference card

Other (Please state): _____

15. How useful/effective has this help information been? (Please circle). Leave blank if you have never tried this help method.

	<u>Not at all useful</u>					<u>Very useful</u>
<i>EasyReader</i> manual	1	2	3	4	5	
Other users	1	2	3	4	5	
TWE/42 personnel	1	2	3	4	5	
On-screen help system	1	2	3	4	5	
<i>EasyReader</i> quick reference card	1	2	3	4	5	
Other	1	2	3	4	5	

Appendix 10.1 (cont.)

TRAINING/SUPPORT DOCUMENTATION

16a. Did you attend an *EasyReader* training course? (Please tick one).

Yes No Don't know

16b. If "Yes", then how useful/appropriate was the training course? (Please circle).

Not at all useful Very useful
1 2 3 4 5

16c. If "No", would you like to attend an *EasyReader* training course? (Please tick one).

Yes No Don't know

17a. Do you think the training course could have been improved? (Please tick one). Leave blank if you did not attend a training course.

Yes No Don't know

17b. If "Yes", in what ways would you have improved the training course?

Please state: _____

18. Would you like additional training? (Please tick one).

Yes No Don't know

19a. Did you receive *EasyReader* support documentation i.e. user guide and quick reference guide. (Please tick one).

Yes No Don't know

19b. If "Yes", how useful/appropriate was the support documentation that you received as part of the training course, i.e. user guide and quick reference guide? (Please circle).

Not at all useful Very useful
1 2 3 4 5

20a. Do you think the support documentation could have been improved? (Please tick one).

Yes No Don't know

20b. If "Yes", in what ways would you have improved the support documentation?

Please state: _____

Appendix 10.1 (cont.)

USER ATTITUDES

21. What was your attitude to *EasyReader* before using it in the work environment? (Please circle).

<u>Negative</u>			<u>Neutral</u>		<u>Positive</u>
1	2	3	4	5	

22. What is your attitude to *EasyReader* now that you have used it in the work environment? (Please circle).

<u>Negative</u>			<u>Neutral</u>		<u>Positive</u>
1	2	3	4	5	

23a. Do you feel you received enough information about *EasyReader* ? (Please tick one).

Yes No Don't know

23b. If "No", how could this be improved, i.e. what kind of information would you like? (Please tick any that apply).

Legal requirements

Reasons for "electronic" documentation

Tips on improving performance

Others (Please state): _____

24. Overall, how confident do you feel using *EasyReader*? (Please circle).

<u>Not at all confident</u>				<u>Very confident</u>
1	2	3	4	5

25a. Have you ever found out of date information on *EasyReader*, that the documents/contents controller is not aware of? (Please tick one).

Yes No Don't know

25b. If "Yes", did you report this out of date information? (Please tick one).

Yes No Don't know

26. How confident were you that the paper documents used previously, provided the most recent documentation? (Please circle).

<u>Not at all confident</u>				<u>Very confident</u>
1	2	3	4	5

27. How confident are you that *EasyReader* provides the most recent documentation? (Please circle).

<u>Not at all confident</u>				<u>Very confident</u>
1	2	3	4	5

Appendix 10.1 (cont.)

34. How useful have these different feedback methods been? (Please circle). Leave blank if you have never used this method.

	<u>Not at all useful</u>				<u>Very useful</u>
	1	2	3	4	5
User feedback form	1	2	3	4	5
TWE/42 personnel	1	2	3	4	5
Other: _____	1	2	3	4	5

35. Have you felt encouraged to give feedback about *EasyReader*? (Please tick one).

Yes No Don't know

36. How familiar are you with the document management process as described in the Document Management Well Engineering Manual? (Please circle).

<u>Not at all familiar</u>					<u>Very familiar</u>
1	2	3	4	5	

37. How familiar are you with the following roles/names? (Please circle).

	<u>Not at all familiar</u>				<u>Very familiar</u>
	1	2	3	4	5
Contents Controller	1	2	3	4	5
Document Controller	1	2	3	4	5

38. Are you aware that every manual now has a contents controller who is responsible for that manual? (Please tick one).

Yes No Don't know

39a. Are you a contents controller? (Please tick one).

Yes No Don't know

39b. If "Yes", are you aware of your role in this position? (Please tick one).

Yes No Don't know

Appendix 10.1 (cont.)

ADDITIONAL COMMENTS

40. Do you have any additional comments about EasyReader that you would like to make? If so, please write them in the space below:

Once you have completed this questionnaire please return it in the enclosed self addressed envelope by Friday 29/4/94, to Jan De Zeeuw c/o S. McCraith, TWE/42, Assen.

Thank you for completing this questionnaire.

Appendix 10.2

SEMI-STRUCTURED INTERVIEW QUESTIONS.

1. What are the good and bad aspects of EasyReader compared with paper documents?

Prompts:

- speed
- out of date information
- access
- searching for information
- losing documents

2. Has your attitude/opinion to EasyReader changed as a result of using it, and if so why?

Prompts:

- same prompts as above?
- how strong were your attitudes/opinions before?
- did you expect your attitudes/opinions to change?
- what changes to EasyReader would make you more/less positive?

3. Have you received enough support/help with EasyReader?

- documentation
- training course
- help information

4. Do you feel you know enough about why EasyReader was introduced or would you like to know more?

Prompts:

- why was it introduced?
- why electronic documentation?
- do you agree?

5. Do you have any suggestions that might increase the acceptance of EasyReader by users?

Prompts:

- extra training?
- more consultation with users?
- more/better computers, etc.?

EasyReader Feature Checklist



The following feature checklist asks you various questions about features (menu commands or icons) in EasyReader. The purpose of the feature checklist is to assess the *EasyReader* interface and to identify ways that might increase the effectiveness of your interaction with the system. This is not a test of how good or bad you are at using *EasyReader*.

Instructions on how to answer the feature checklist questions are given overleaf. However, before reading the instructions, please fill in your personal details in the space below:

Name: _____

Age: _____

Job Title: _____

Date: _____

How long have you been with Shell/NAM? (Please state): _____

Please indicate what you think your level of computer experience is on the following scale:

Not at all experienced

Very experienced

1

2

3

4

5

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure

1 = less than once a month 4 = every 2-3 days

2 = once a month 5 = every day

Appendix 10.3 (cont.)

Instructions

(1) In the "Existed?" column, please indicate whether or not you knew the menu command or icon existed. You may respond by using either of the following methods:

✓ = Yes ✗ = No ? = Unsure

(2) In the "Used?" column, please indicate whether or not you have ever used this menu command or icon. You may respond by using either of the following methods:

✓ = Yes ✗ = No ? = Unsure

(3) In the "How often?" column, please indicate how often you normally use this menu command or icon (approximately). You may respond by using either of the following methods:

0 = never used it 3 = once a week
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

(4) In the "What for?" column, please indicate whether or not you know what the menu command or icon does. You may respond by using either of the following methods:

✓ = Yes ✗ = No ? = Unsure

(5) In the "Need?" column, please indicate how often you think you might need to use this menu command or icon. You may respond by using either of the following methods:

0 = no need at all 3 = once a week
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

(N.B. you can only answer this question if you know what the command does.)

Please answer each question for every feature in the checklist by putting the appropriate answer in the relevant space. An example of a completed section of a feature checklist is shown below:

Edit	Existed?	Used?	How often?	What for?	Need?
<u>U</u> ndo	?	✗	0	?	0
<u>C</u> ut	✓	✗	0	✓	2
<u>C</u> opy	✓	✓	4	✓	4

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

Appendix 10.3 (cont.)

File	Existed?	Used	How often?	What for?	Need?
<u>O</u> pen					
<u>C</u> lose					
<u>S</u> ave					
Save <u>A</u> s...					
<u>E</u> xport...					
<u>P</u> rint...					
<u>P</u> rint Setup...					
<u>P</u> references...					
<u>E</u> xit					

Edit	Existed?	Used	How often?	What for?	Need?
<u>U</u> ndo					
<u>C</u> ut					
<u>C</u> opy					
Paste					
Copy <u>S</u> GML					

Book	Existed?	Used	How often?	What for?	Need?
<u>G</u> o Back					
Find					
Search <u>F</u> orms...					
<u>N</u> ext					
<u>P</u> revious					
<u>C</u> lear Search					
Search <u>H</u> istory...					
<u>S</u> earch Panel					
<u>O</u> pen Object					
<u>D</u> elete Object					
<u>M</u> anage Annotations...					
Create <u>A</u> nnotations...					

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

Appendix 10.3 (cont.)

Collection	Existed?	Used	How often?	What for?	Need?
<u>F</u>ind					
<u>R</u>efine Search					
<u>C</u>lear Search					
<u>V</u>iew <u>M</u>ultiple					
<u>S</u>earch Panel					

Journal	Existed?	Used	How often?	What for?	Need?
<u>N</u>ew					
<u>S</u>tart <u>R</u>ecording					
<u>S</u>top Recording					
<u>T</u>ake Snapshot					
<u>S</u>how					
<u>N</u>ext					
<u>P</u>revious					
<u>R</u>ename...					
<u>D</u>elete...					

View	Existed?	Used	How often?	What for?	Need?
<u>E</u>nlarge					
<u>R</u>educe					
<u>O</u>riginal Size					
<u>T</u>OC					
<u>M</u>ain					

Window	Existed?	Used	How often?	What for?	Need?
<u>N</u>ew Window					
<u>C</u>ascade					
<u>T</u>ile					
<u>A</u>rrange <u>I</u>cons					
<u>C</u>lose <u>A</u>ll					

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

Appendix 10.3 (cont.)

Help	Existed?	Used	How often?	What for?	Need?
<u>R</u> eader Guide					
<u>B</u> ook Window					
<u>U</u> sing the TOC					
<u>S</u> earching					
<u>A</u> dvanced Searching					
<u>H</u> ypertext Navigation					
<u>A</u> nnotations					
<u>R</u> ecording <u>L</u> ocations					
<u>P</u> rinting					
<u>A</u> bout...					

Button Bar	Existed?	Used?	How often?	What for?	Need?
					
					
					
					
					

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

Appendix 10.3 (cont.)

Book Icons Existed? Used? How often? What for? Need?

Drawing Tools Existed? Used? How often? What for? Need?

0 = never used/no need at all 3 = once a week ✓ = Yes; ✗ = No; ? = Unsure
 1 = less than once a month 4 = every 2-3 days
 2 = once a month 5 = every day

Appendix 10.4

EasyReader FC Information Sheet

As part of the monitoring programme for *EasyReader* we have asked you to complete two different instruments obtaining different types of information, i.e. a feature checklist and a questionnaire (that you received earlier). We would like to try and measure how difficult and how time consuming it was for you to complete the feature checklist and the questionnaire. Please indicate this on the following scales:

1. How difficult was it for you to complete the FEATURE CHECKLIST?

Very difficult

1

2

3

4

Not at all difficult

5

2. How difficult was it for you to complete the QUESTIONNAIRE?

Very difficult

1

2

3

4

Not at all difficult

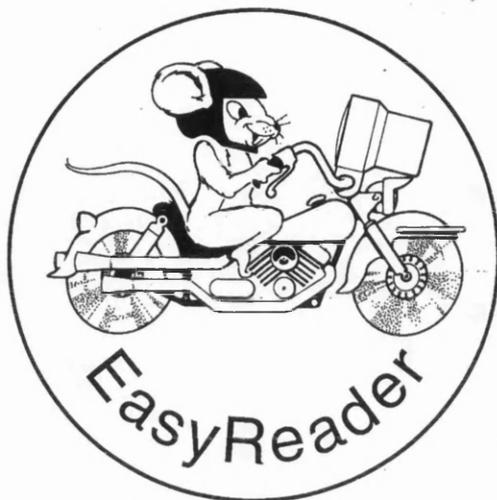
5

3. Overall which instrument took longer to complete? (Please tick).

Feature Checklist

Questionnaire

EasyReader Final Questionnaire



This questionnaire is the final part of the *EasyReader* monitoring programme.

This aim of this questionnaire is to follow-up the previous questionnaire and obtain more detail about your areas of concern with *EasyReader*. This extra detail will be used to try and improve the aspects of *EasyReader* that you, as users have expressed concern about.

This questionnaire is much shorter than the previous questionnaire that you received since we are only looking at specific aspects of the *EasyReader* system.

Remember the information gained from this questionnaire will be used to improve the *EasyReader* system and will benefit all users of *EasyReader*.

All questions on the questionnaire can be answered by either:

- circling the appropriate number on the scale shown.
- ticking the appropriate answer box.
- writing your answer in the space provided.

It is important that you answer all questions on the questionnaire, as clearly and as accurately as possible. If you have any problems or queries with the questionnaire please call 05920-63891/62566 and ask for Jan de Zeeuw.

Thank you (again) for your time and effort

Appendix 10.5 (cont.)

8. How confident are you that the findings of this monitoring programme will be used to improve *EasyReader*? (Please tick one).

Not at all
confident

Very confident

1

2

3

4

5

9a. Are you satisfied with the quality of support that you receive from TWE/42? (Please tick one).

Yes

No

Don't know

9b. If "No", then why? (Please state).

Your use of *EasyReader*

10a. Has your use of *EasyReader* increased in the past two months? (Please tick one).

Yes

No

Don't know

10b. If "Yes", then why do you think this is so? (Please tick any that apply).

Better knowledge of *EasyReader*

More confident using *EasyReader*

Job requires greater use of *EasyReader*

Other (please state): _____

11. Have you become more aware of the functional advantages of *EasyReader* over the past two months, e.g. "word searches", "annotations", etc.? (Please tick one).

Yes

No

Don't know

EasyReader hardware

12a. Are you satisfied with the hardware provided? (Please tick one).

Yes

No

Don't know

Appendix 10.5 (cont.)

12b. If "No", then why? (Please state).

13. Would improvements to any of the following increase your use of EasyReader? (Please tick any that apply).

	<u>Yes</u>	<u>No</u>	<u>Don't know</u>
Faster machines	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Faster Network access	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Larger screens	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Other (please state): _____

Your awareness of *EasyReader* document management

14. have you become more aware of the role of TWE/42 in Document Management over the past two months? (Please tick one).

Yes No Don't know

15. Have you become more aware of the procedure for using the User Feedback form over the past two months? (Please tick one).

Yes No Don't know

16. Have you become more aware of the role of the Document Controller over the past two months? (Please tick one).

Yes No Don't know

17. Have you become more aware of the role and responsibility of the Contents Controller over the past two months? (Please tick one).

Yes No Don't know

18. Have you become more aware of the future plans for *EasyReader* e.g. new CD every 3-4 months, etc.? (Please tick one).

Yes No Don't know

Appendix 10.5 (cont.)

Additional comments

19. Do you have any additional comments about EasyReader that you would like to make? If so, please write them in the space below:

Once you have completed this questionnaire please return it in the enclosed self addressed envelope by Friday 29/4/94, to Jan De Zeeuw c/o S. McCraith, TWE/42, Assen.

It is our intention to follow-up the information gained from this monitoring programme, i.e. areas that users have expressed concern about will be addressed wherever possible; this may include the option of extra training courses, better equipment, etc.

Thank you yet again for completing this final questionnaire.

Appendix 10.6

T-test for independent samples: Difficulty rating for FC and Questionnaire (Q) by inexperienced users

FC	FC ²	Q	Q ²
2	4	4	16
3	9	3	9
2	4	3	9
1	1	3	9
2	4	2	4
3	9	3	9
3	9	4	16
1	1	2	4
$\Sigma FC = 17$	$\Sigma FC^2 = 41$	$\Sigma Q = 24$	$\Sigma Q^2 = 76$

$$t = 2.198$$

Average t at 0.05 level = 2.145

We therefore conclude that our value of 2.198 is significant at the 0.05 level, being larger than the average t value of 2.145 (i.e. inexperienced users rated the FC significantly harder to complete than the questionnaire).