UNIVERSITY
of
GLASGOW

Department of
Computing Science

Putting icons in context: the influence of contextual information on the usability of icons

Jackie Moyes

A Thesis submitted for the degree of Doctor of Philosophy

to

The University of Glasgow

October 1995

Abstract

Previous research has been driven by the idea that a good icon is one which consists of a form that is instantly recognisable as representing the underlying referent. As a result many of the design decisions suggested have been based round this premiss. However, the empirical method used to support these design decisions has often tended to ignore two factors of everyday interface usage that have an important influence on the ability of an icon to communicate its meaning.

The first factor that most researchers tended to ignore was that an icon is usually displayed simultaneously with a number of other icons. Rarely will you find icons displayed in isolation, yet many methodologies ask subjects to judge an icon's ability to communicate its meaning without any additional information from the interface to support it.

A second issue is that, in everyday usage, most users interact with icons over an extended period of time. Therefore it is not necessarily essential that the icon's meaning should be easy to guess, only that it be particularly easy to learn.

The ultimate aim of this thesis, therefore, is to suggest that icon research should no longer be concerned with measuring the degree to which a particular representational form of a solitary icon can increase performance on a subject's initial exposure to the icon. Instead research should be widened to consider how attributes in the interface interact, and how this interaction may vary over time as user knowledge of the interface increases.

A total of six experiments were performed. Each experiment had four conditions, manipulating the abstractness of icon shape and consistency of icon position. After a training period conditions changed without warning and the conclusions depended on whether or not performance was disrupted by the change. Results from the experiments suggested overwhelmingly that issues such as what attributes users relied upon, and when they relied upon them, were far more complicated than initially predicted.

The conclusions of the thesis question the generalisability of many long held assumptions about icon design. Findings show that once an icon is placed into an interface it is difficult to predict which of its attributes the user will rely on. It seems that this will not be determined by how representational the attribute is, but rather how discriminable it is in comparison to other attributes, once the icon is placed within a particular interface. Results also show that users are aware of multiple attributes apparent in the interface and may switch the attribute that they rely on several times within the entire user learning curve. However, it is impossible to say how much users learn, since the experimental data has shown that learning appears to continue after the classic performance measures of reaction time and error rate have reached asymptote. Finally, the results highlight the flaws in current evaluation paradigms;

namely that most are guilty of testing icons in isolation and only in the initial stages of the learning curve.

The experimental results, combined with the extensive literature review in Chapter 2, elicit a number of interesting future research questions, as well as providing designers with some general suggestions as rules-of-thumb for interface design. The thesis concludes by elaborating on these points.

Table of Contents

Main chapters

Detailed Table of Contents

Table of Figures

Acknowledgements

It is a great pleasure to acknowledge my debt to the large number of people involved, directly or indirectly, in my research which led to the production of this thesis. Firstly, I thank my supervisor, Steve Draper, whose discussion, criticism and guidance have been invaluable throughout the past four years. I am also eternally grateful to my second supervisor, Phil Gray, for his useful comments, inspiration and patient support regarding the many ideas and plans I assaulted him with over the years. Thanks also go to Paddy O'Donnell for his practical suggestions and expert knowledge in experimental design and statistics.

The Computing Science Department and, in particular, the joint psychology Glasgow Interactive Systems cenTre (GIST) have provided an excellent environment in which to study and carry out research. I am especially indebted to all the staff and students who so willingly offered to be subjects - without them none of this would exist.

Thanks also to Colin Prosser for his initial guidance and enthusiasm.

Finally, to family and friends, thank you for your help and for your never ceasing support and encouragement.

Funding for the research reported here was provided by DECcollege.

Declaration of Authorship

The material presented in this thesis is the product of the author's own independent research carried out at the Computing Science Department, University of Glasgow under the supervision of Phil Gray and Steve Draper.

The exception to this is the material presented in Chapter 9. The experiment discussed in this chapter was designed in conjunction with Patrick W. Jordan. Although collaboration on this experiment occurred, the motivation behind this research for each researcher was entirely independent. Therefore the main focus of the findings from this experiment are also independent.

1. Introduction

1. 1 Thesis aims

The motivation behind this thesis is to show that traditional evaluation methods for icons fail to take into account how a user of a real interface uses the entire interface or uses information gathered over time in order to acquire and understand an icon's meaning. There are two main arguments to support this claim.

Firstly, previous research has generally evaluated an icon's usability while displaying the icon in isolation. This is an entirely unrealistic situation. In the majority of interfaces an icon would be surrounded by other icons. It would be expected that relationships between attributes present in both the target icon and the surrounding icons would have the potential to provide users with additional information as to the icon's meaning. Thus an icon's meaning, which in isolation is difficult for a user to interpret, may become obvious when placed within a certain context. Conversely, it is equally appropriate to argue that some icons will become more difficult to identify when placed within a particular interface. It may even be possible to claim that an icon's ability to communicate its meaning is *entirely* dependent on the context in which the icon is situated. This thesis indicates that this may indeed be the case and raises the issue of whether discriminability of the relevant attributes within an interface may be more important than its ability to convey meaning.

The second argument is that users interact with icons over an extended period of time. Therefore the usability of an icon should not be determined by how easily a user can identify an icon's meaning on first exposure to it, but should instead be measured throughout the entire learning curve. Therefore, as previous research has shown (e.g. Rogers, 1986), when initially presented with the icon set, users attempt to identify and rely upon attributes which most effectively communicate the icon's meaning. This thesis shows, however, that over time users may move their attention from the initial learned attribute(s) to rely instead upon attributes that offer greater performance advantage or less cognitive effort.

The result of this analysis is to identify and explore some of the neglected factors determining an icon's usability in practice, in context and over time. The ultimate aim, therefore, is to question the validity of experimental paradigms which restricted their analysis to icons in isolation with no user interaction allowed. Consequently, by questioning these paradigms the thesis also questions the assumptions derived from experimentation using these paradigms. Although this thesis does not conclude by presenting the designer with a procedure for design, it does succeed in offering a number of issues that are essential to consider when creating an iconic interface.

1.2 Motivation

The motivation for the work reported here arose from the realisation that most previous research on computer-based icons were simply continuations of work carried out in the 1960's and 70's on road traffic and public information signs (e.g. Dewar and Ellis, 1977) Many researchers had failed therefore to realise that there were several important differences between computer-based icons and signs used elsewhere. Three main differences are:

Computer-based icons often represent abstract concepts

Everyday signs tend to represent tangible objects. Such objects have obvious pictorial representations. However computer-based icons generally represent functions or objects belonging to the computer architecture or the operating system. With no obvious pictorial representations the question of how this information is conveyed to the user becomes a more pervasive issue. The main criticism of previous research is that it has continued to focus on icon form for the solution to this problem.

Computer-based icons are rarely displayed in isolation

Road traffic signs are commonly displayed alone. Only occasionally will they be accompanied by a textual sign when more specific detail is required. Computer-based icons are rarely displayed in isolation from other icons. This has the added benefit of presenting the user with many contextual cues and attributes not present when the icon is viewed in isolation. These attributes have the potential of being manipulated so that they can actually aid the user in his or her attempt to understand what an icon means, thus ultimately improving the usability of the interface. It is clearly apparent from the research performed so far that many designers are oblivious to this potentially rich communication source, and therefore continue to develop and analyse design theories relating unnaturally to the usability of isolated icons.

Users interact with icons over extended periods of time

Users of graphical interfaces actively interact with computer-based icons, unlike road traffic signs which users passively perceive while driving. Interaction and subsequent feedback allow the user to learn what a previously un-guessed icon actually means. Apart from some public access systems (e.g. health information systems) much computer software is intended for use over an extended period of time. This allows many users the possibility of achieving a high level of experience. Previous work has not considered the possibility that what may be important for design on the first occasion of use may not be an important design feature when designing a usable interface for experienced users.

It was from these basic criticisms that the thesis aims as described in section 1.1 originated.

1.3 Thesis Outline

Through several empirical investigations this thesis answers some questions basic to our understanding of how context influences the usability of icons throughout the learning curve. Chapter 2 provides an extensive review of previous work on icon design and directly related issues such as the influence of task, previous knowledge, cultural influence and graphic design on usability.

Chapter 3 introduces the model of usability developed by Jordan et al (1991) which is used to divide the learning curve into three identifiable parts: guessability; learnability and experienced user performance. At the end of this chapter the main thesis hypothesis is stated, providing the questions that will be considered by the empirical investigations.

Chapter 4 describes the experimental paradigm used by the six experiments. The paradigm was established after consideration of previous methodologies, and is validated by comparing its design with paradigm recommendations made by Kantowitz (1992).

Chapter 5 introduces the initial investigation made to test the claim that users who rely on representational icon form will not learn position and that users who are presented with abstract form will learn icon position. Results support this hypothesis.

Chapter 6 describes experiments 2 to 5. As each experiment is performed there is an evident shift in conclusions drawn, moving away from the hypothesis suggested for experiment 1 towards the idea that users in the representational condition do initially rely on form, but switch to using position since it offers a greater performance advantage and is far easier to discriminate.

The last experiment is described in chapter 7. The experiment was placed in a separate chapter since it had several important differences in its design. It was also a far larger experiment than any of the others performed previously (involving more trials and therefore longer experimental sessions). As a result the amount of data collected is very large and is described in some detail in the discussion section of this chapter.

Chapter 8 provides an overview of the results collected from all six experiments, concluding by proposing a revised hypothesis. In chapter 9 this revised hypothesis is tested on an experiment that had been performed prior to any of the other experiments, designed to consider the effects of consistency on usability. Since essentially the same paradigm was used for experiments 1 to 6, it was important to test out the new hypothesis on results from a 'non-contrived' method. Results achieved from the consistency experiment were found to be explicable in terms of the revised hypothesis.

Chapter 10 concludes the thesis by raising criticisms of some of the assumptions made by previous research on icon design and questioning the paradigms used by previous research to derive these assumptions. Possible areas for future research are introduced and some design considerations, generated by this piece of research, are suggested for designers.

2. Review of Icon Research

2.1 Introduction

Iconic interfaces have only been in use for a relatively short period of time, making their first appearance some time in the mid 1970's, and developing prominence in the early 1980's with the development of the Star User Interface (Smith et. al., 1982). As a result there is a moderate volume of research in the area. Consequently, it is possible to include in this thesis a rather extensive literature review, covering more than just the work directly influencing the thesis hypothesis. The advantage of such a review is that it allows the reader to obtain a clearer idea of the diversity of research in the area of icon design, before focusing on the particular topic that this thesis is concerned with.

2.2 A clear definition of an icon

A useful starting point would be to define what this author means by the word 'icon'. It is crucial to be clear about what an icon is, its essential properties, and how it is distinguished from other types of sign, in order to become aware of the particular issues affecting computer based icons, and thus how these issues influences design decisions. This is not as straightforward as may first appear. Evidence suggests that many researchers in the area of icon design have failed to fully appreciate these interface specific issues before commencing research. As a result, work done previously appears too focused and restrictive to be of practical help to designers. This is an important issue which will be returned to and discussed in greater depth at various points throughout the thesis.

2.2.1 What is an icon?

The following definition of an icon by Yvonne Rogers (1986) is presented as an example of a typical definition by a researcher in the field of icon design:

> Pictographic symbols generally the size of a postage stamp, which are displayed on the screen. Their function is to represent underlying system objects, data structures and processes in a form that corresponds to the real world.

By comparison, the Concise Oxford English dictionary (1991) defines an icon as:

1. A devotional painting or carving, usually on wood, of Christ or another holy figure, especially in the Eastern church.
2. An image or statue.
3. *Computing* a symbol or graphic representation on a VDU screen of a program, option or window, especially one of several for selection.
4. *Linguistics* a sign which has characteristics in common with the thing it signifies.

The dictionary definitions of interest are definitions 3 and 4. These highlight the important differences between icons as defined as existing on a computer screen compared with icons as originally defined (as used for example by linguists and semioticians). Three important differences can be specified.

Firstly, when referring to the more traditional definition of an icon (4) the dictionary refers to the icon as having something in common with the object it signifies. However for computer-based icons (3) it merely refers to the relationship between the icon and its referent as being symbolic. This difference is perhaps due to the contrast between what computer-based icons and traditional icons generally represent. For example, the sign in Figure 2.1 represents a warning to drivers that the road ahead narrows. Drivers are able to see the road ahead and also its width. However, computer-based icons generally represent functions or objects belonging to the computer architecture or the operating system. These tend not to be tangible. For instance, how would you represent memory? What does software look like? How would you represent an operating system? The point is not that these entities cannot be represented pictorially, but rather that it may be far more difficult to achieve a direct mapping between the representation and the referent due to the abstract nature of the referent.

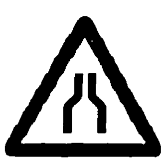


Figure 2.1: British road traffic sign representing the warning "road narrows"

As a result, many of the icons found at the interface do not share perceivable properties with the underlying objects or functions that they represent. This is perhaps why the desk-top metaphor became such a popular interface design in office based systems. Instead of trying to represent the actual properties, they were represented through metaphorically similar properties to which office workers could relate. The alternative solution is to use abstract icons. This is favoured by Gittins (1986), who claims that we should select technical icons in preference to natural objects to reduce the problems of mis-interpretation due to cultural variation. However, by using abstract icons (such as the ISO symbol represented in Figure 2.2) designers are not representing concepts, as Rogers suggests, in a "form that corresponds to the real world". A computer-based icon can only be accurately defined as a graphical sign present on the screen standing for either a command, object or function.

As suggested by dictionary definitions 3 and 4, this is a very much looser definition of the term icon than that used traditionally. Research in the area of semiotics (e.g. Peirce, 1931) define only those representations that directly represent the object or function as iconic. Using this definition, the ISO symbol in figure 2.2 would not be described as iconic, since it does not possess "characteristics in common with the thing it signifies".

Perhaps instead of the word 'icon' being used to describe all classification types, the word 'sign' would have been more accurate. 'Sign' is defined in the Concise Oxford English Dictionary (1991) as "something that indicates a fact" or "an arbitrary mark or device that stands for a word, phrase etc." The word 'signs' is also used by Peirce (and by semioticians in general) as the collective term for representations. The word 'icon' has been so widely used to refer to signs found within the interface, however, that more people understand what is being referred to when the word 'icon' is used (although technically incorrect), than when the word 'sign' is used.

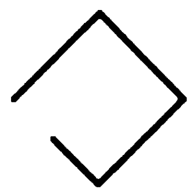


Figure 2.2: The International Standards Organisation (ISO) sign representing the command-action 'store'

The second point distinguishing computer-based icons from traditional icons, as identified by the dictionary definition, is that computer-based icons are usually "one of several". In other words what makes computer based icons different from signs in general is that they are rarely viewed in isolation. This is an important distinction which again is missing from Rogers' definition. As a direct result of being viewed within a collection of icons, the number of contextual cues available to the user increases. Generally users gain from the sign's surroundings some information relating to the underlying meaning being represented by the sign. For example, if the icon is situated within a word processing package, then the user may expect the icon to relate to some word processing function, to underline text, etc. Similarly, if the sign that represents the washing instruction "do not tumble dry" (Figure 2.3), is found on the label of a shirt, users can infer that it must represent some sort of information relating to the shirt to which the label belongs.

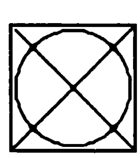


Figure 2.3: Sign used to represent the washing instruction "do not tumble dry"

Furthermore by having a number of icons present simultaneously, users can infer with greater accuracy what an icon may represent, by looking at the alternatives and deducing a possible meaning.

The third difference highlighted by the dictionary definition is that computer-based icons are used "for selection". Thus users of graphical interfaces do not just passively observe the icons that are present, but are able to interact with them. Therefore if they are not sure of what the icon represents, by interacting with the icon, (which usually involves placing the pointer over the icon and clicking the mouse a set number of times), users will often gain feedback relating to what the icon does. So for example, the 'zoom' icon shown in Figure 2.4, may not be immediately clear. However by clicking on it once, the size of the window is either enlarged or reduced, depending on its previous state. The user therefore learns to associate the 'zoom' icon with this function.

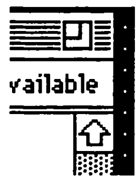


Figure 2.4: The 'zoom' icon on the Macintosh desktop

In some cases the connection may be a bit more obscure. If for example subjects are unsure of what the MacPaint selection rectangle tool stands for, they firstly have to click on the icon to select it. They then need to know that they should click and drag the mouse to create the rectangle. Furthermore, even if users get this far they then need to know that the rectangle should enclose some already existing graphic to have any effect. Only once they have done this will they discover the function that the icon represents, by either cutting, copying, deleting or dragging the rectangle.

The main point to keep in mind is that users do not passively perceive the computer-based sign as they pass it in a car. Instead users spend long periods exposed to the icon, and being allowed to interact with it. Therefore, although direct or lawful links between the representation and the referent may be important for first time guessability, humans are good at forming associations, and thus almost any connection may be learnable.

Rogers' definition of an icon fails to take account of this distinction. However she is not alone in failing to acknowledge the features highlighted by the Oxford English dictionary as being important to the intrinsic definition of what a computer-based icon is. As will be described in the following sections, most research has seen the inability to design icons in a form representational to the real world as a reason for not using iconic interfaces. Few researchers have considered how communication could be aided by presenting icons within a suitable context, allowing contextual relationships to become apparent. Furthermore, few researchers have looked at user interaction with icons over an extended period of time, thus not considering the possibilities of design for long-term use.

Inherent in the definition of what an icon is, is the understanding of its purpose. By knowing what an icon is meant to do, we can attempt to achieve this through design.

2.2.2 What is an icon's purpose?

The main purpose of an icon is to communicate its meaning to the user. Therefore the purpose of the computer-based ISO sign in Figure 2.2 is to communicate that it represents the command-action 'store' and that the user should activate this icon when wishing to perform this command. It could also be argued that an icon's purpose is to support useful user actions. However textual representations have the same purpose, so why should someone chose an iconic interface in preference to a textual one? According to many researchers icons possess additional properties unique to pictorial representation, thus making them a more desirable means of communication than alternative methods. For example, it has been claimed that icons have the ability to reduce complexity and therefore make the interface easier to learn (Lodding, 1983). It has also been noted that icons have the ability to convey certain types of information more directly and with more immediacy than a verbal equivalent. This affects both the icon's initial comprehension and subsequent retention (Hemenway, 1982). Scott and Findlay (1991) have suggested that search times for graphical icons are faster than search times for words. It has also been claimed that the use of iconic interfaces would make the electronic 'world' seem less alien (Smith, Irby, Kimball and Verplank, 1982), and that icons have the potential to exhibit a universality not found in natural language (Blattner et al., 1989). Furthermore, Rohr and Keppel (1984), suggested that icons are capable of presenting information in a more spatially condensed and holistic form[1] .

There are perhaps as many research reports claiming to deny the existence of these features (e.g. Whiteside et al., 1985). The main controversy appears to stem from the point raised by Norman (1993):

> The most appropriate format depends upon the task, which means that no single format can ever be correct for all purposes (Norman, 1993, pp. 60)

However, the aim of this thesis is not to argue about the virtues of one type of interface over another, but instead to analyse how the meaning of an icon can be communicated successfully assuming that the designer has already decided to use an iconic interface. Suggestions as to how meaning can be transmitted successfully are considered in more detail in the next section.

---

[1] Holistic refers to the ability to communicate information in a single icon which would require several words to convey textually.

2.3 Previous research: how can the icon's purpose be achieved?

There are several methods of conveying an icon's meaning to the user. A survey of previous research suggests that there are perhaps five ways in which to increase the likelihood of successful communication. These factors are highly inter-related. They are as follows:

- Being aware of user characteristics
- Being aware of the perceived task
- Cultural influences
- Graphic design techniques
- Successful mapping of the representation to its referent

The first two issues are very much background issues relating to user perception and knowledge of the task and the interface that the designer should be aware of. The latter three are more direct icon design issues and therefore, in order to give the reader a clearer idea of the state of current research, are discussed in greater depth in the following sections.

2.3.1 User characteristics

How even the most basic of tasks is perceived is very much user-dependent. According to Draper (1993a, pp. 207) "Even if, instead of just saying 'Prepare a business letter', we give them a printout of the target letter to be achieved, different kinds of people will simply notice, or not notice different aspects of the layout" (e.g. font size, margin widths, tabs, etc.).

There are, however, several common user characteristics that allow the designer to make a number of generalisations about the likely tasks performed. For example, there is a difference between designing for the general public and designing for a specific user group. Hence a user assuming a particular role (i.e. author, referee, editor) is expected to carry out the set of tasks associated with that role. One higher-order grouping of both tasks and roles is a job. A job can be defined in terms of the set of roles that a person is expected to take on and the tasks associated with them. For example, the job of lecturer might include the roles of administrator, teacher and researcher.

Clearly there are cases where it is difficult to categorise some task as being entirely within any role. Similarly there may be occasions where the same task might be found within several roles. However, in general, the way users see the computer application, and the user interface, is often determined by the role they are performing or have been trained to do (Bannon and Bødker, 1991).

Designing for a specific user group has an important effect on icon design. With specific roles comes experience of specific tools and practices the general public may not be aware of. So, for example, if a package is designed to help typographers with their work then it is likely that there are specific tools that typographers will require. If the icons representing these tools have a direct association between the pictorial representation and the tools as they look in real life then typographers, trained to use such tools, would recognise this mapping immediately. If, however, members of the general public were shown the application, they may be unable to comprehend what the icon represented, since they lack this specific knowledge.

This type of design would be in contrast to the design of systems used by the general public (e.g. public information systems), where the intention of the user or amount and type of previous experience is unpredictable. In such cases it is likely that the designer should try to transfer knowledge from an existing skill or task that most users may have in the past performed, i.e. through use of metaphor.

Previous experience is an important factor affecting task performance with an iconic interface. Experience can be general (i.e. knowing that all word processors have files which can be created, saved, deleted and changed), or it can be specific (i.e. the names of commands in a specific word processing application, the positions of icons, and the sequences of actions). Pollock (1988) demonstrated that people who are required to learn to use a second word processor after having learned a different word processor, apply knowledge they have acquired about using the first processor to the second. Sometimes this transfer will be appropriate and sometimes it will not. There is widespread evidence that in those cases where there is an appropriate transfer of existing knowledge, from one task situation to another, there are considerable savings of training time to be made, in terms of length of practice required to achieve a given level of task performed. Subjects who are able to transfer knowledge from one interface to another tend to achieve a higher level of performance sooner and with less training effort than would otherwise be the case.

According to Pollock, both general and specific knowledge can be transferred to the task domain. However, it seems likely that general knowledge can only be utilised in task performance through specific knowledge. If specific knowledge appropriate to the new task context is not available, then specific knowledge from the existing task contexts may be inappropriately transferred. Pollock suggests that appropriate specific knowledge transfer can be facilitated by the use of consistency between the representations and spatial layouts in first and second word processors, and also by the use of metaphor in the representation (e.g. the use of a spinning top to represent 'go to the top').

Individual differences, even within targeted user groups, can also have an effect. For example, the degree to which the user is motivated to learn the interface and the number of distractions present in a user's surroundings can both have an affect on the level of attention the user employs on the interface. As a result both have an effect on how successful the interface appears to the user. Therefore, it is important for an iconic interface designer to note that differing user characteristics result in differing levels of success of any design. Where software is being designed for a specific user group this problem may not be so evident. However, for software designed for the general public there is no way to predict the characteristics or degree of previous experience any user is likely to have. The only way to reduce the likelihood of a completely incomprehensible interface for most of the population is by testing the interface on a large number of diverse users. By doing so it would be hoped that any characteristics of the interface unclear to a specific type of user within the entire user population would be encountered, and therefore adjusted.

Therefore it could be argued that the physical design of the icon (e.g. whether the form is considered to represent the object or function) is going to be highly dependent on the user's previous experience. The category an icon may be classed as falling into (e.g. to say that one icon is more representational than another) may vary between different users. This therefore warns about making generalisations in the principles of design.

2.3.2 The perceived task

Often the task that the user plans to perform, or perceives he should perform, is going to influence what he expects to see on the interface. This will have an influence on the effectiveness with which an icon's attributes are able to communicate their meaning. For example, a novice Macintosh user may wish to perform the task of ejecting a disk. Looking at the Macintosh interface no icon appears relevant to that task, since he recognises the trash can as representing a place for discarding unwanted data, and therefore sees it as representing the tool he would use should he wish to erase the disk rather than eject it.

Many task analysis theories have been developed by HCI researchers to gain a better idea of how users are likely to interact with the interface (e.g. GOMS, Card et al, 1983; TAGs, Payne and Green, 1986; and KAT, Johnson and Johnson, 1989). According to Smith, Irby, Kimball, Verplank and Harslem (1982), task analysis involves establishing who the users are, their goals in performing the task, the information they use in performing it, the information they generate, and the methods they employ. Icons relate to this analysis by being both part of the information that users rely on when performing the task, and also being part of the method by which users would perform the task (e.g. by selecting the appropriate icon). Thus, the nature of the task will have a profound effect on how the icon is designed.

At their lowest level tasks can be described as a series of functions being executed upon one or more objects. Icons can represent both functions and objects. Depending on which they are, the emphasis on their design should be different. Bannon and Bødker (1991) suggest that objects should be designed to be the focus of the user's attention, whereas a good tool or function must be used almost unconsciously by the person performing the task. They present as an example the scenario of a carpenter who focuses his attention on driving the nail, while holding the hammer and moving it through operations. Only when the fluent hammering stops and the hammer does not respond to the actions of the carpenter does the hammer become an object in itself, a situation which Bødker (1989) describes as conceptualisation. In such a situation the user's attention moves away from the object to the tool, thus highlighting a failure in the design of the tool. Therefore design principles for icons as objects and icons as functions should be different. Designers must ensure that both are unambiguous from their surroundings, but ensure that objects map closely with the tasks as perceived by the user.

The task the user may perform could also determine which attribute of the icon identifies as meaningful to that task. For example, an object icon can potentially show information about what the icon represents, whether it is on of the most frequently used commands, if it is the last command that was performed by the user, the size of the file that the icon represents, and the icon's behaviour. The dilemma faced by designers is how much information should be available presuming, as has already been suggested, that which task the user is likely to want to perform is not always possible to predict. The easiest, most flexible solution would be to represent all information. However, Norman (1986) suggests, perhaps too adamantly, that a lot of information presented on the interface at any one time can be confusing and cause clutter. It seems likely that Norman's scenario refers to cases similar to that highlighted by Mackinley (1986), where richly detailed interfaces are used to present information that is not so detailed; in other words cases where the representation contains more attributes that the represented function or object possesses. The lesson to be taken from McKinley's example, therefore, is that as long as the representation contains only attributes found in the underlying object nothing should be assumed to be clutter, and performance is less likely to be degraded. Therefore, by guarding against unintentional presentation of misleading information, a very detailed interface could be created with the potential of supporting a multitude of tasks.

2.3.3 Cultural influence

Cultural expectations and standards will have an extremely important influence on the successful communication of an icon's meaning. It has been assumed that because icons can resemble the objects or functions that they represent, their meanings will automatically be transparent. Jervell and Olsen (1985) suggested that "a good icon will awake the same reaction in most people irrespective of background, education and nationality". This appears to be rather optimistic. The International Standards Organisation (ISO, 1981) consider a sign to be successful if 75% or more people can recognise or understand the sign. Familant and Detweiler (1991) suggest that even icons that have a direct relationship between representation and referent, there is still a degree of arbitrariness associated with the relationship (something that is generally overlooked by researchers and designers).

Research relating to the effects of cross cultural conventions on interaction is rather sparse, mainly due to the cost of implementing any such research. However, the need for research in this area is becoming increasingly important. According to Russo and Boor (1993) in 1991 five of the six biggest US computer companies derived over 50% of their income from international sales. This trend is expected to continue as political and economic changes open new markets around the globe.

Cultural conventions can have an effect on icon design at several levels. At the graphical design level, typical icons tend to conform to conventions established in North America and Europe and known as Western graphic conventions. These conventions, for example, allow users to interpret a series of marks on a two-dimensional screen as conveying a sense of distance for the viewer, or three-dimensionality (see Figure 2.5).

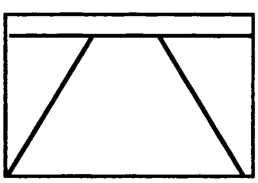


Figure 2.5: An example of an image that Western cultures may identify as conveying distance.

Deregowski (1989) has suggested that this ability is not universal.

Colour is another important graphical design feature that is affected by cultural conventions. For example in Egypt red symbolises death, in the USA it represents danger, while in China red symbolises happiness. Similarly in Egypt, yellow is equated with prosperity, while in the USA it signifies cowardice. Therefore it is probably wrong to assume that meaning communicated through colour is necessarily going to be universally recognised.

The flow of information is important within an icon. Western cultures read from left to right, therefore for example, the play button on most tape or CD players is usually labelled in Western cultures by an arrow head pointing to the right (see figure 2.6). If the play button were instead presented with an arrow head pointing left, we would perceive this as the wrong way round. However, Arabic cultures read from right to left, and therefore the play button may appear more natural in Arabic countries if displayed in reverse.

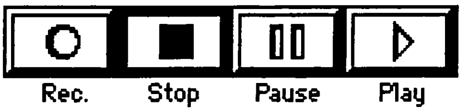


Figure 2.6: Typical 'western' tape or CD player symbols

The representation or image itself is the next level that can be affected by cultural conventions. The meaning conveyed by the icon can either be perceived differently by different cultures, or not perceived at all. Examples of the latter condition are fairly common. This is particularly true for icons that rely on metaphorical or historical representations. An example of a culturally specific metaphor is the use of a 'piggy bank' to represent the command 'save' (see figure 2.7). Saving money in a container formed in the shape of a pig is a particularly Western tradition, and therefore would not be understood in non-western cultures.

Figure 2.7: Image of a piggy bank, used to represent the command 'save'

According to Russo and Boor (1993) metaphoric representations are particularly culturally-dependent. These representations are created by representing the linguistic referent in a graphical image that has the same name. For example, the use of a spinning top to mean 'go to the top', or on the Macintosh, an icon displayed in the form of an insect is used to represent a fatal error (or 'bug' as it is known in English). These links may not exist when the command is translated.

Historical representations are representations that at one point in time did have some direct mapping with the underlying referent, however as time has gone on this no longer applies. For example, in certain European countries a sign depicting a cornet is used to represent mail, since at some time it was used by postmen to announce their arrival. A more recent example is that of the finder icon on the Macintosh (see figure 2.8), which depicts the outer casing of an Apple Macintosh. However Macintosh machines have changed so much since this icon was created that the icon now only represents the appearance of the machines at the lower price range, therefore to new users the icon should become increasingly less representational, and increasingly more abstract.

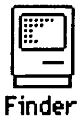


Figure 2.8: Macintosh representation of the Finder

When looking at icons that convey different meanings between cultures there is often a difference between comprehension and acceptability. An example of cultural differences in comprehension is that of the business card. According to Russo and Boor (1993), in Japan business cards are presented at the start of a meeting to establish seniority, thus allowing appropriate interaction to follow. In the USA however, they are used to convey information about the business or representative. Therefore if an icon depicting a business card were used on a network, Japanese users would recognise the image, but perhaps understand the icon to indicate status rather than business information..

There are, however, certain icons that are comprehended differently by different cultures, and as a result are found to be offensive (for either social, religious or legal reasons). As a rule designers should be especially careful when designing icons that contain religious symbols, or the human body (particularly if depicting women or hand gestures). An inappropriate use of a simple graphic can result in a customer being offended or insulted. For example, in some Arabic countries all signs that use graphical representations of the human body are required to display them without showing heads or shoulders. Therefore the typical male/female signs commonly used for pedestrian crossing signs in the UK, would be unacceptable in such countries. Similarly the USA hand signal for 'OK' is understood as a vulgar gesture in Germany and Brazil.

Even in Western society some computing conventions, such as displaying a sign in inverse video when it has been selected, are not conventions familiar to most cultures. Similarly many of the objects depicted at the interface are actually artifacts of the computer and its related environment (context of use). Therefore there are likely to be conventions that will have to be learned by any culture. At some point in history all man-made conventions were new, an example being the standardisation of a twenty-four hour clock, which now feels to most people to be an entirely natural convention. As Hutchins, Hollan and Norman (1985) suggest:

> If we restrict ourselves to only build an interface that allows us to do things that we can already do and in ways we already think, we will miss the most exciting potential of new technology: to provide new ways to think and to interact with a domain.

This would suggest therefore, that in order to progress, new conventions must be both developed and learned by all cultures.

2.3.4 Graphic design limitations

This is a factor of icon design which has quite clear guidelines. Currently one of the main drawbacks is that what may look good on paper may not work so well when transferred onto the screen. As technology improves and becomes cheaper, current limitations (e.g. resolution) should be removed, allowing the designer greater scope in the design used. As for the previous section, the main aim of reviewing work in this area is to give the reader an idea of how much of an influence graphic design has in the user's comprehension of an icon.

Graphical design can perhaps be described as the physical level of icon design. This section attempts to review what the visual characteristics of the icon are. It must be noted however, that this section specifically relates to the graphical design techniques used to design icons. It therefore does not include techniques such as the use of perspective, transparency, opacity, light and shadow which can be employed when designing the interface as a whole (Staples, 1993).

It is important to reflect on why graphic design is an important component of icon design. What benefits are to be gained by ensuring that the graphic design of the icon is correct? For example, does it increase the level of discriminability of an icon? Does it increase the degree of user recognition? To increase either of these measures of performance would be beneficial. Graphic design probably has a strong influence on the degree of discriminability of an icon. However, discriminability is highly context-specific, with the potential to vary quite dramatically between sets of icons. Therefore a highly discriminable icon can only be categorised as discriminable relative to the set of icons in which it was situated. The most important aim of the graphic designer is almost certainly to ensure that the representational form of the icon is clearly represented with maximum simplicity and clarity. According to Gittins (1986), "Rapid perception and recognition of symbols are affected by various aspects of their figural form". If the information is based on the clearest and most stable pattern of figural form, then it increases the likelihood that the user will recognise the graphical presentation, understand the representation, and ultimately understand the meaning that is being communicated by the interface designer.

The following sub-section will review the principles of graphic design relevant to icon design, after which, interface-specific methods will be considered and discussed.

2.3.4.1 General graphic design points to consider

Many principles covered in this section were first studied empirically by the Gestalt psychologists (e.g. Koffka, 1935). Indeed, according to Easterby (1970) "Much of the perceptual theory which would appear to be of any practical value in symbols design derives from their ideas and experiments." The Gestaltists proposed that the aim of these principles was to combine the parts of the stimulus into a whole, thus producing maximum simplicity and clarity. This is also the main aim of contemporary graphical interface designers.

The key question is whether or not people recognise the icon as the concept that it is supposed to represent. The following sub-sections contain summaries of five general graphic design issues which have the potential to influence the recognisability of icon form.

1. Visual Integration of Constituent Elements

The Gestalt psychologists formulated a number of principles of perceptual organisation to describe how certain perceptions are more likely to occur than others. The following principles are primarily concerned with the grouping of components within the icon.

Proximity

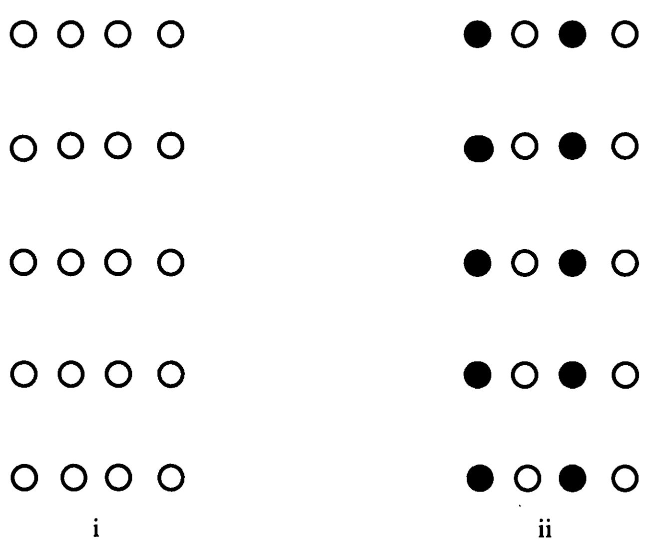


Figure 2.9: Examples of grouping by proximity and similarity

One of the most important factors determining the perceptual organisation of a graphic is the proximity of the elements within it. The closer elements are to each other, the more likely they will be grouped together. An example of this is Figure 2.9(i), where the observer views the dots forming rows because they are nearer horizontally then they are vertically.

Similarity

A second grouping principle is similarity. Things that look 'similar' are grouped together. Grouping by similarity can sometimes have the power to override grouping by proximity. For example, in Figure 2.9(ii), similarity in the brightness of the dots overrides the effect of proximity. How similar must items be in order to be grouped together? Olson and Attneave (1970), investigated the process of grouping by similarity by asking subjects to indicate where the 'odd' quadrant lay within a circular display of simple pattern elements. They found that the quadrant was most easily spotted if the elements within it differed in slope from the rest of the display (e.g. < ^). The quadrant was most difficult to stop if the elements differed in their configuration, but not in the slopes of their component parts (e.g. < >). These findings suggest that the variables that influence grouping are not necessarily the same as those which the user might judge to be conceptually similar. Therefore Figure 2.10(ii) might be considered to be more similar to Figure 2.10(i), when viewed as a single pair than Figure 2.10(iii). However, when viewed in large numbers, it is the difference in the common orientations that are more salient and therefore Figure 2.9(iii) is harder to distinguish from Figure 2.10(i).

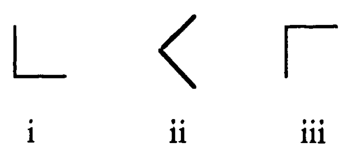


Figure 2.10: Elements from the Olson and Attneave experiments (1970)

In order to improve discriminability within the interface such design information is essential. The principles of similarity and proximity do not necessarily only apply to the design of individual icons, but to the design of the interface as a whole. Therefore, icons in close proximity or with similar appearance would be considered to be conceptually related.

Good Continuation

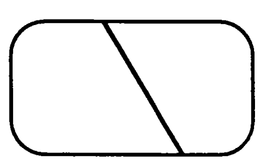


Figure 2.11: Example of good continuation, suggested by Easterby (1970)

Gestaltists argue that perceptual organisation will tend to preserve smooth continuity rather than yielding abrupt changes. Quite dissimilar objects may be perceived as "belonging together" as a result of a combination of proximity and good continuity. Figures with a strong tendency to smooth continuous outline are believed to be more stable and clearer. An example is the use of the icon to represent an eraser. Figure 2.11 is seen as one rectangle with rounded corners as opposed to 2 angular shapes joined together since the eye groups the straight lines together as opposed to the angular shapes.

Closure

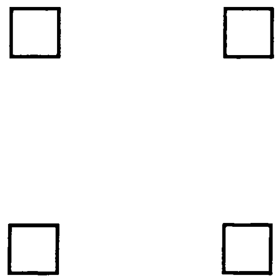


Figure 2.12: Typical Gestalt example of closure.

In Figure 2.12 the four squares are seen as forming a square rather than a cross. We imagine that the boxes represent the corners of the square rather than the end points of a St. Andrews cross (the Scottish flag). According to the Gestaltists this is because if it is geometrically possible to organise things perceptually as "closed" figures then one will generally do so, rather than perceiving it as an open figure. This is such a strong perceptual phenomena that Gittins (1984) warns "If it is necessary to have an icon which is unclosed, be aware of the user's tendency to perceive closure where none exists". Bartlett (1931), found in several studies that subjects asked to produced unclosed figures invariably drew the figures as closed. An icon showing an example of closure can be seen in Figure 2.13 where the figure is perceived as a black and white circle, with a black rectangle behind the white section of the circle.

Figure 2.13: An icon used to represent 'disk' (from the research report by Maissel, 1990)

Symmetry

Symmetry again encourages grouping. In the real world, most objects are approximately bilaterally symmetrical (e.g. the human face). Therefore we tend to perceive things that are symmetrical as objects. As a consequence symbols should be as symmetrical as possible, providing that asymmetry adds no further meaning to the figure.

Figure/Ground

Figure/ground refers to the way in which a figure is separated from the rest of the visual field. Many of the ambiguous pictures designed by Escher relied specifically on creating figure/ground ambiguity. It is for this reason that Gittins (1986) stresses that "the ground form of a figure should be clear and stable". This can achieved by employing four more principles established by Gestalt psychologists.

Relative Size

All other things being equal, the smaller of the two areas will be seen as the figure against a larger background. Therefore in Figure 2.14(i) the black area will be seen as the figure, and the white area will be seen as the background.

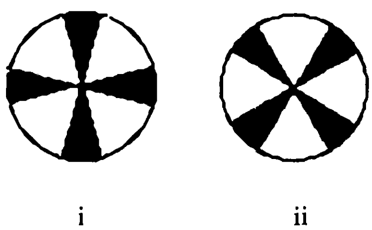


Figure 2.14: Examples of relative size (i) and orientation (ii)

Surroundedness

Stability is further increased, if the larger area completely surrounds the smaller area (see Figure 2.15).

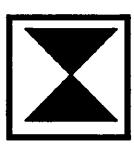


Figure 2.15: 'Delete' icon

The white area of the icon representing 'delete' completely surrounds the black area and therefore is seen as the background. The black area is seen as the figure.

Orientation

If figure 2.14(i) is oriented so that the white area is arranged around the horizontal and vertical axes, then it is easier to see this larger area as a figure, rather than a background (see Figure 2.14(ii)). Therefore according to Easterby (1970), the prevailing outlines of the symbol should follow as far as possible the main spatial axes of the horizontal and the vertical.

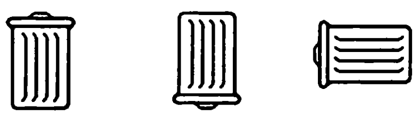


Figure 2.16 Various orientations of the Macintosh icon representing the trash can

In icon design, the orientation of the icon may be far more important in some cases than others. For example, the Macintosh trash can (even when oriented to the horizontal and vertical axes) may convey a different meaning if presented upside down rather than standing the right way up (see Figure 2.16). For example, the upside down icon may suggest the trash can being emptied, whereas the upright icon could suggest that the trash can is able to accept items to be deleted. For abstract icons such as the icon representing the command 'store to disk' in Figure 2.17 orientation may not be so crucial.

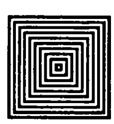
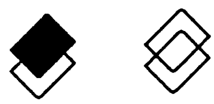


Figure 2.17: The abstract representation for 'store to disk' (Maissel, 1990)

Contour

For a figure to be defined it must possess a contour. Contour can be generated either by outline or contrast boundary. In both cases the surface figure is identifiable by contrast variations at the contour edges. Whereas solid figures have only one contour to contrast with the background, outline drawings have an inside and an outside contour. It may be that the line-bounded form is potentially ambiguous since there is a surface figure defined by the 'inside' of the line as well as the 'outside', whereas the contrast-bounded figure only has an 'outside' boundary.

The principle of psychological superiority of the contrast-bound figure over the line form, was first studied experimentally by Gottschaldt (reported by Koffka, 1935). By exploiting these perceptual qualities of line and contrast-bound figures, the graphic designer can emphasise one part of a symbol figure at the expense of another. Thus with simple two dimensional icons we can make one part (contrast-bounded) appear to lie on top of another "less important" part (line-bounded). An example can be found in Figure 2.18. Figure 2.18(i) shows a version of the icon representing the command 'copy'. The contrast-bound figure appears to be on top of the line-bound one. In Figure 2.18(ii), the idea of one figure on top of another is not so evident.



i ii

Figure 2.18: Using the icon representing the command 'copy' to show the superiority of contrast-bound figures, as studied by Gottschaldt

Whether the icon contour is contrast-bound, or line-bound, Arend et al (1987) have suggested that outline is a global feature. There theory proposes that the global features of figures (e.g. the general outline, size, colour) can be responded to considerably faster than local features (e.g. the detailed lines and structures within figures). Hence icons which differ from each other in global features should be searched and selected faster than icons that have similar outline shapes, but differ from each other in local features. The reason for this is assumed to be because of a "global superiority effect" (Pomerantz, 1983), in which the perception of global features in a figure is more rapid than the perception of local features.

2. Level of Detail

The previous sub-section suggests that a distinctive outline is essential for quick and efficient icon recognition. However, as Rogers (1989a) makes clear, an icon is also required to be meaningful and memorable. She suggests that in order to design meaningful icons, distinctive local features that can depict the state of the various underlying referents will need to be incorporated into the design. It would appear therefore that the most efficient icon design would be one which contained both a distinctive outline shape, and a sufficient amount of local features, presenting the user with information relating to the underlying referent.

The position of local features is important. As Easterby (1970) points out, if the detail is enclosed within the boundary of the shape then it is quite a good icon. If it is outside, it distracts from the impact (discriminability) created by the outline, and therefore makes the icon a poor one.

Given that computer-based icons are generally small, the amount of detail that can be displayed in each icon is limited. However, this is not so great a problem if cartoons are used. Cartoons are often designed so that only the essential details remain. Ryan and Schwartz (1956) found that in some cases cartoon-like depictions of objects were more accurately recognised by subjects than either line drawings or photographs. This finding was reinforced by Biederman and Ju (1988).

3. Direction of Reading

As described in the section on cultural influence, the reading direction of the composite elements of a symbol is important. In Western culture it is common practice to read from left to right, and top to bottom. Therefore when showing transformations (see Figure 2.19) it is usual for observers to expect to see the original situation at the left followed by the transformation to the right. By following the standard reading practices of the culture the icon should be understood far more clearly.

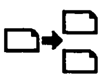


Figure 2.19: An alternative representation for the command 'copy'

4. Colour

Colour should be used judiciously and for a specific purpose. Combinations of certain colours can be very difficult for the observer to discriminate, especially if they are highly saturated (e.g. red next to blue). Furthermore, the presence of two or more bright colours means that the eye will continuously have to re-focus over the image resulting in eye strain and fatigue. This can distract users' attention from the task, and therefore become a negative feature. Therefore Strijland (1993), recommends that colours should be limited to a small palette, comprising of several greys.

The use of colour simply to make the icons more life-like does not improve their discriminability. Biederman and Ju (1988) compared recognition of a fully coloured photograph with recognition of a simplified line drawing of common objects (e.g. banana, chair, camera), and found no difference in reaction time or errors. Furthermore, searching for a particular icon from a whole set of individually coloured-in icons may prove to be more of an laborious task than if the icons were achromatic (Rogers, 1986).

Colour can, however, be used effectively to encode information. Rogers (1986) suggests that colour can be put to best use at the interface by using it to distinguish between icons that are related to each other, e.g. similar operations or file types. By using colour to divide icons into related subgroups, time taken to search and detect various tasks can be decreased.

However, Strijland (1993) suggests that there are drawbacks to using colour as a coding mechanism. Firstly if, for example, different hues are used to indicate the size of a file, then a key or legend will be required to explain the relationship between hue and file size. Secondly, colour perception varies considerably between users. Around 10% of the male population is known to be colour blind. Ageing can also have an effect on a user's ability to perceive colour. Lastly, if software is designed to be used on a range of machines, designers need to be aware that some machines may have monochrome monitors and therefore any information transmitted via colour will be lost.

5. Connotative factors

Connotative (affective) factors can be conveyed through graphical techniques. For example, angularity is often associated with hardness, curved horizontal upward strokes with gaiety, curved downward strokes with sadness, etc. (cf. Barnard & Marcel, 1977, Werner & Kaplan, 1967). Colour can also be used (e.g. red is associated with danger/anger, green with safety etc.). However, as mentioned in the previous section, colour associations are highly culturally-dependent.

2.3.4.2 Design principles specific to computer-based icons

By using the design guidelines in the previous section the designer may have created a promising icon design on paper. However, what appears to be an effective, discriminable graphic on paper is often not so effective when transposed onto the interface. Therefore, several interface specific factors need to be considered.

1. Low Resolution

Resolution can make a design much more crude and less elegant. Typical screens have a resolution from 70 to 120 dots per inch, (dpi). With these resolutions, jagged lines are inevitable. Therefore it is best to avoid odd angled lines (see figure 2.20). Similarly, designers need to be careful to select circle sizes that have the least jagged outlines.

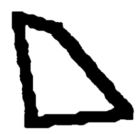


Figure 2.20: Example of 'jagged' lines found on the computer interface

Alternatively, jagged lines can be avoided by using a technique called anti-aliasing. For this technique the designer blends the line colour with the background colour. This, however, only works with a colour or grayscale system not black and white. There are two problems with anti-aliasing. Firstly, it tends to make the icons look blurry. Some users actually prefer this, although many don't. Secondly, anti-aliasing only works if you know the colour of the background. Backgrounds that change require a change in the anti-aliasing. Only computers with specific anti-aliasing systems, are capable of doing this.

2. Pixel Shape

Pixel shape can be square on some systems and rectangular on others. According to Strijland (1993) it is important for the designer to know the shape of the pixel before designing the icons. If the icons are to be used on any machine, two versions of the icon may need to be designed.

3. Selected State

The main difference between icons designed on paper and those designed on the screen is that those on the screen can be interacted with. It is important therefore to consider how the current 'state' of the icon can be displayed. For example, in a Macintosh interface, icon state can be either normal, open, off-line, or selected. It is particularly important that selected icons are distinguishable from the rest of the icon set. Individual icons that have equal amounts of the unselected colour (e.g. white) and the selected colour (e.g. black), will not show much contrast when selected (e.g. the black pixels will become white and vice versa). Similarly, if on the interface the unselected colour of some icons is white (changing to black if selected), and for other icons the unselected colour is black (changing to white if selected) then when either type of icon is selected it will fail to be stand out visually. Therefore it is best to have all icons on the interface either completely black or completely white (see Figure 2.21).

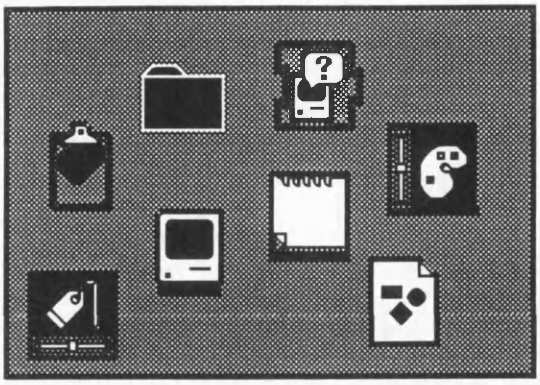


Figure 2.21: A collection of Macintosh icons. The folder is the selected icon. However, because other icons contain black, the icon does not stand out at all. (Example taken from the tutorial paper by Strijland on Icon Design, 1993)

4. Style Consistency

Keeping the graphical style of the icons consistent throughout the interface helps to establish the idea of corporate identity. A consistent style may involve the use of "a constant scale, limited size variations, the orientation of figures with respect to text, limited use of colours, limited variation of line weights, and the treatment of borders for figures and pictograms" (Marcus, 1984).

To establish corporate identity, most current systems use distinct graphical styles, (i.e. Macintosh uses a cartoon-like 2-D style of icon, while NeXT uses a photographic 3-D style).

2.3.4.3 Summary of graphic design principles

The principles described so far suggest guidelines to ensure that icons are presented in their clearest forms, thus increasing the likelihood of successful recognition by the user. However, in cases where there are many icons displayed at one time, it is important to consider what factors are essential to ensure rapid recognition. To assess what factors influence recognition, Arend et al (1987) found that for fast menu selection icons should not be visually similar within a given set, but should possess distinct global features with respect to each other.

Arend et al.'s findings, however, contradict the principles of style consistency. The aim of style consistency is to keep icons visually similar, thus limiting the degree to which icons can be visually distinct and thus discriminable from one another. The issue of discriminability and its importance in icon design is raised in greater detail in chapter 10.

On a similar theme Byrne (1993) hypothesised that the simpler the icon, the faster the visual search will be. Unfortunately, the experiment supporting this claim was only carried out with one set of icons for each condition (e.g. simple or complex) and therefore the results may actually be context dependent. Byrne does suggest, however, that complex icons may be of some use, when there are few icons on display, or when speed of discrimination is not important.

Experimental work is continuing in this area, and a clearer understanding of all the factors involved in determining the speed of recognition seems likely in the near future.

2.3.5 Mapping and categorising icon form to the underlying form

Given the increasing abundance of iconic forms, for example on computers, automatic teller machines, home appliances, etc., it is unsurprising that attempts have been made to develop taxonomic systems for organising and classifying icons (e.g. Gaver, 1986; Gittins, 1986; Blattner et al., 1989; Rogers, 1989a, 1989b; Webb, Sorenson and Lyons, 1989; Maissel, 1990). On inspection, however, there appear to be several apparent shortcomings with these taxonomies, limiting their usefulness in the evaluation and design of icons. It is from these limitations that the original thesis hypothesis originates. Therefore it is perhaps important to consider these taxonomies in some detail.

2.3.5.1 Philosophical classification systems

The origins of iconic classification are rooted in the area of semiotics, the science of signs. Therefore, it is perhaps interesting to consider the taxonomies developed for general signs to allow comparisons to be made between these and the taxonomies developed for computer-based signs. One of the main founders of the field of semiotics was the American philosopher Charles S. Peirce (1839-1914). He developed his own taxonomy of signs, classifying them into three categories; symbol, index and icon.

Symbols are distinguished by having an essentially arbitrary relationship with the object that they symbolise. An example would be the ISO symbol representing the command-action 'store' (see Figure 2.2). Such pictorial symbols can be compared to characters that have been developed in verbal language, where there is no physical or analogous correspondence between the characters used to form words and their intended meaning. In such cases therefore, associations must be learned.

The second type of sign, index, is related to the associated underlying object in some non-arbitrary way. The most frequent example used is the association between smoke and fire (i.e. the signal smoke logically implies the existence of the object fire). Another example would be a bird's footprints in the snow logically implying the presence of the object (the bird).

The third type of sign is called an icon. The essential feature of an icon is that it purposefully shares properties with the object to which it refers. An example would be the icon shown in Figure 2.22) representing the input device the mouse.

Figure 2.22: An iconic representation of a mouse

Another philosopher to develop a taxonomy of symbols was Fred Dretske (1988). Like Peirce, he suggests that signs can be categorised as falling into three groups; Type I, Type II and Type III. Type I is similar to Peirce's classification of symbolic icons, where the relationship between the sign and the object it represents is entirely arbitrary. Dretske stresses that in order for the representation to work some sort of convention must be established and maintained by the users otherwise misrepresentation will occur.

Between Type I and Type II representational systems is a category Dretske calls signs. Signs are similar to the category Peirce defines as indices. According to Dretske:

> we don't have to let tracks in the snow, bird songs, fingerprints, and cloud formations stand for the things we take them to indicate. There is a sense in which, whether we like it or not, these tracks, prints, songs and formations indicate what they do quite independent of us, of how we exploit them for investigative purposes, and of whether we even recognise their significance at all.

Although the examples given so far have all been taken from nature, these signs can also be man made. For example, a fuel gauge pointer indicating that the fuel tank is full, will indicate this fact independent of whether we recognise this relationship or not.

Dretske's Type II grouping is equivalent to the representations Peirce described as icons. Dretske regards these as "Conventional Systems of Representation". He defines Type II representations as natural signs. The important feature of natural signs is that they have their own intrinsic indicators and thus have the ability typically to indicate more than just one relationship, all of which exist irrespective of whether or not users ever recognise them. Therefore they usually tend only to be used to represent one particular condition, depending on what condition the users find of particular interest or importance. Dretske gives the example of the electronic fuel gauge. This not only indicates the amount of fuel left in the tank, but also the level of downward force on the bolts holding the tank to the car's frame, and the amount of electrical current flowing in the wires connecting the gauge to the fuel tank. Even though all these indications are present, we tend only to take one of these conditions to be what the gauge represents i.e. we assign the function of indicating the amount of petrol in the tank to the gauge. Therefore any misrepresentations will be limited to what the sign has the function of indicating. The important distinguishing feature of Type II representations is they are indicators that have specifically been assigned a representative role to perform. It is this specific role assignment that distinguishes Type II representation systems from signs.

Type III representations are what Dretske calls "Natural Systems of Representation". Like Type II representations, they have their own intrinsic indicators, however they also have their own intrinsic elements and functions, i.e. what they represent is independent from human choice. This type of representation, he claims, cannot be found to exist within the real world. It can however apply to beliefs, since he sees beliefs as having a function which is not imposed from outside. Although the existence of Type III representations is an important concept, computer-based representations can never possess the essential requirements to be considered a Type III representation.

Peirce	Dretske	Definition
Symbol	Type I	Arbitrary relationship with the object that they represent
Indice	Sign	Related to the associated underlying object in some non-arbitrary way
Icon	Type II	Purposefully shares properties with the object to which it refers
	Type III	Contains intrinsic indicators, but has a function not externally imposed

Figure 2.23: Summary of philosophical classification systems

Comparing these two philosophical examples (see figure 2.23), it is evident that the main emphasis of sign classification has been on how closely the representation contains properties similar to the properties contained in the referent. However, even comparing only two taxonomies, it is clear that category boundaries are not obvious. This lack of distinction encourages the potential for difference in interpretation and division of classes between different taxonomies. This is particularly apparent when comparing taxonomies developed for computer-based signs.

2.3.5.2 Initial computer-based classification systems

The main attack in this chapter is made against the many attempts to develop computer-based classifications. The main criticism is that this type of research is too restrictive and too impractical to be generally applicable in icon design.

The proposed classification system set out by David Gittins (1986) is shown in figure 2.24.

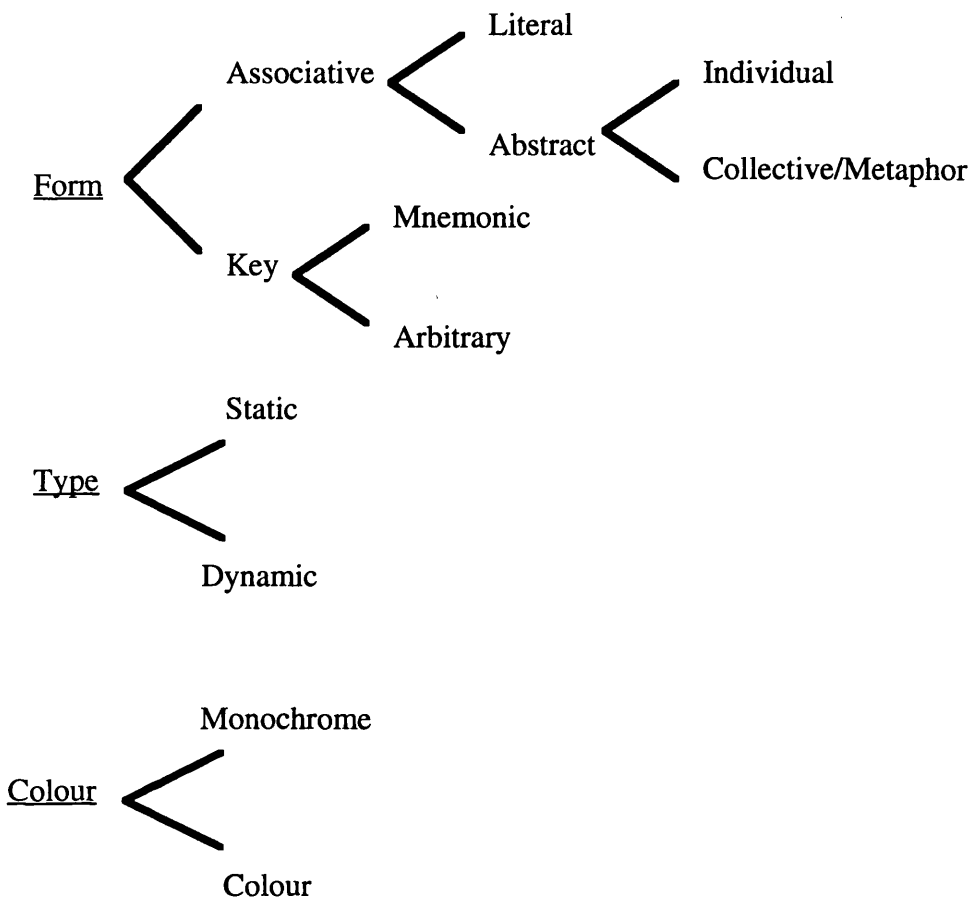


Figure 2.24: A diagrammatic representation of the sign taxonomy suggested by Gittins (1986)

Gittins suggests that icons can be classified in three ways; by their form, type and colour. Unfortunately Gittins failed to extend his discussion on type or colour and gives no indication why he includes these in his taxonomy while excluding other possible attributes (e.g. outline shape, size, position, etc.).

Icon form is classified by Gittins into two categories. In the first condition the icon form indicates a characteristic of the underlying object (associative) and in the second condition the form acts as a cognitive key to the object (key).

Although associative form is further divided into literal and abstract signs, Gittins again fails to make any further comment about these distinctions. It does seem clear, however, that Gittins' associative icons can be identified as being similar to Peirce's icons or Dretske's Type II representations. The only statement he does make regarding this branch of the hierarchical tree is that "a collection of icons may share some common design features which reinforce the effectiveness of the mapping. In this case they are collectively known as a 'metaphor'." An example would be the trash can, folders and other office equipment upon which the designers of the Xerox Star computer system based their interface, and which has since been collectively named the desk-top metaphor (Smith et al, 1982).

The key icon form is sub-divided into the categories mnemonic and arbitrary. Mnemonic icons are those icons designed to remind the user of the commonly accepted name of the object. As an example Gittins suggested that a pictographic icon of a guillotine could be used to represent the process 'execute'. Arbitrary icons on the other hand are characterised by their arbitrary design, from which it is not possible to infer the nature of the underlying object. This is again similar to Peirce's symbols and Dretske's Type I representations, and therefore requires convention or standardisation to be enforced to communicate meaning.

An alternative classification system was developed by Yvonne Rogers (1986, 1989a, 1989b) and Rogers and Oborne(1985, 1987). This system distinguishes between function and form. Function relates to the function that the icon is being used for. This can include the representation of concrete system objects, tools for drawing packages, messages for errors etc. The form of an icon can be classified into either concrete objects (e.g. folders), symbols (e.g. arrows, lines) or a combination of the two. However form can be classified further, into the way in which it represents the underlying concept or referent. There are four classifications, the first of which is resemblance icons. According to Rogers resemblance icons depict the underlying referent through an analogous image. The second group, called exemplar icons, depict a typical example for a general class of objects. Generally the image depicted will be composed of the most salient attributes associated with that class. This has been most commonly used within the area of public services, an example being the use of a knife and fork to represent "restaurant services" (see figure 2.25).

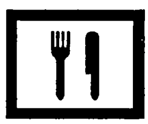


Figure 2.25: Public service sign representing "restaurant services"

The third type of icon is the symbolic icon, which is used to convey an underlying referent that is at a higher level of abstraction than the image itself. Rogers illustrates this using an image of a fractured wine glass to represent fragility. This category could perhaps be similar to what Gittins was trying to convey by his class of abstract icons, which was one of the sub-categories in the associative condition of his classification system. Rogers' last category, arbitrary icons, bear no similarity to the referent.

Comparing Rogers' taxonomy to those discussed previously, resemblance icons are similar to Peirce's icons, Dretske's Type II representations and Gittins' associative icons. Arbitrary icons correspond with Peirce's symbolic relationships, Dretske's Type I Representations and Gittins' abstract icons. The two other classifications (exemplar and symbolic) appear to be more specific forms of Peirce's definition of indices.

Gaver's (1986, 1989) classification system is created for auditory icons. He suggests that "perceptual mappings are not between the computer reality and the outward manifestations that are accessible to users, but between these manifestations and the model world of the computer". In common with previous classification systems, Gaver's system consists of three categories which he names symbolic, metaphorical and iconic. This classification system relates to that of Peirce and Gaver stresses that concepts have also been adapted from both Heil (1983) and Bates (1979).

Symbolic mappings between the display entity (symbol) and the model world are essentially arbitrary, depending on convention for meaning. Metaphorical mappings make use of similarities between the entities. This can occur in one of two ways. Structure mapping occurs when similarities between the structure of the icon and object are exploited, whereas metonymic mapping occurs when a feature is used to indicate a whole, for example, a crown could be used to represent a king.

Iconic (or nomic) mappings are based on Peirce's icons. In this condition the icon's characteristics are causally related to the object or function that it represents. The icon need not be a pictorial representation, but can be a sketch, caricature, outline or recorded sound.

Blattner et al. (1989) also distinguish three types of icon in their taxonomy. Representational icons are simple pictures of familiar objects. Abstract icons are entirely composed of geometric forms and shapes. Semi-abstract icons are either a combination of both representational and abstract icons, or are representational icons so simplified that they can be considered abstract forms. This classification system which Blattner et al. use as a framework for auditory icons (or 'earcons') is differentiated from previous taxonomies by the fact that it is based more on the physical representation of the sign rather than on the representation of the underlying concept. This can be attributed to the influence of Marcus (1984) who favoured the implementation of graphic design principles into interface design.

This focus on the pictorial or physical appearance of signs is also apparent in the work of Jonathan Maissel (1990). His aim is to develop a method to evaluate and identify the graphical characteristics of easily interpretable signs. The paradigm involves selecting a particular referent, such as the trash can, and presenting it in various representations and styles to discover which is the preferred style (in the case of the trash can there are 47 variants). Maissel's taxonomy is perhaps more detailed that Blattner et al's, since the variants are classified into types of metaphor and artistic style. Metaphor is classified into four categories: direct (no metaphor) representation, analogy, linguistic/mnemonic (similar to Gittins definition of mnemonic icons) and abstract/symbolic. The actual classes are similar to all the classification systems discussed previously, but in common with Blattner et al., instead of classifying the representation, Maissel uses the distinctions to classify the graphical style of the icon rather than the referent itself. Therefore Gaver's taxonomy would classify the trash can icon as metaphorical whatever the graphical style, whereas Maissel's taxonomy would class Figure 2.26(i) as a direct representation and 2.26(ii) as abstract/symbolic.

i ii

Figure 2.26: Two alternative representations for the trash can icon.

2.3.5.3 Problems with the initial classification systems

The main failing is that there are fundamental differences between the categories that different systems propose. For example, Gaver's taxonomy (1986, 1989), contains a category called metonymic icons where a feature is used to indicate a whole. Rogers' taxonomy describes a category of icons called symbols, which are used to convey an underlying referent that is at a higher level of abstraction that the image itself. Gaver's system does not contain anything resembling Rogers' symbol category, and Rogers' taxonomy does not include Gaver's classification of metonymic icons. Similar inconsistencies can be found for all taxonomies cited. By failing to contain categories that are present in other taxonomies, and vice versa, none of the taxonomies can be described as comprehensive.

If we ignore the inconsistent naming strategies, what taxonomies do appear to agree on is that there is one class of signs that share inherent properties with the object that they represent at one extreme and another class of sign at the other end of the scale that have an entirely arbitrary relationship with the object they represent.

Another inconsistency is that some taxonomies confuse **how** a thing is represented with **what** is represented. For example, Maissel's classification system identifies differences in the physical or pictorial structure of signs (how). Other taxonomies, such as Gaver's, focus entirely on the way in which the sign represents the underlying concept or referent (what). Others show a combination of the two (e.g. Blattner et al.).

Rogers (1989a) is aware of this distinction. She defines the type of representation used in the depiction as the form. Generally the icon form consists of either concrete objects (e.g. files, trash can), abstract symbols (e.g. arrows, circles, lines, mathematical formulae), or a combination of the two. It is the mapping between the form and the underlying object or function (generally known in the field of semiotics as the referent) that determines the category within which the icon can be classified.

2.3.5.4 Recent advances

The idea of separating referent from representation and concentrating on the mappings between the two is developed further by Familant and Detweiler (1991). Rather than attempting to develop a taxonomy they developed what they describe as a 'generative framework of semiotic reference'. The aim of designing this framework is to guide subsequent research, by demonstrating how different categories of signs are related to one another, thus ultimately leading to the creation of new categories.

The basis of this framework is to use predicate calculus to draw the distinction between individuals and attributes. Individuals are defined as "any discernible aspect of a world of representational interest", for example, a function, or state of the machine. Individuals have attributes that can be either physical (e.g. size, colour) or functional (e.g. the ability to move). For any individual it is assumed that there is an essentially infinite set of attributes that are true of it. The relationship between the individual and the attributes is therefore one to many. Attributes are, however, also discernible aspects of the world, and therefore can also be described as individuals with their own set of attributes.

Signals (the form of the icon) are individuals and thus have their own attributes. Similarly the referent is also an individual and has an attribute set. Therefore iconic mappings are those where the attribute sets of the signal and the referent purposely intersect (i.e. they both contain common attributes). By contrast, for mappings that can be defined as symbolic, any similarities between the signal and the referent are accidental.

Familant and Detweiler further develop their theory by introducing the concepts of denotative and sign referents. Denotative referents are the specific aspects to which reference is made by the sign. Thus a sign is successful if it can unambiguously refer to the sign's denotative referent. The example provided is the phrase "The Big Apple" which has the denotative referent New York. The sign referent is the original referent of the sign. Therefore for "The Big Apple" the sign referent would be a large fleshy fruit.

These two ideas - the use of predicate calculus and the theory of denotative and sign referents - are combined. Therefore the sign referent's attribute set can be related to the denotative referent's attribute set in five logical ways, so developing the framework of semiotic reference. These five distinct relationships are defined as follows:

1. The two sets are disjoint, having no attributes in common.
2. The denotative referent's attribute set can be a proper subset of the sign referent's attribute set.
3. The attribute set of the denotative referent and the attribute set of the sign referent are identical.

4. The sign referent's attribute set intersects with the denotative referent's attribute set, but both sets also have attributes not found in the other set.
5. The sign referent's attribute set is a proper subset of the attribute set of the denotative referent.

According to Familant and Detweiler relationships 1 and 2 do not exist in the real world, whereas relationships 3, 4, and 5 have been found to exist.

2.3.5.5 Limitations

Although the framework appears more complex than previous taxonomies, Familant and Detweiler claim that it is only an initial attempt to understand semiotic reference in a computing environment. Consequently, the framework fails to make any direction towards the classification of more complex signs that exist on current graphical user interfaces. Typically icons tend to contain both symbolic and iconic signs (see Figure 2.27) and are therefore described by Familant and Detweiler as form composites.

Figure 2.27: An example of a form composite

It is likely that Greenlee (1973) was referring to composite icons when he suggested that it is possible for a sign to function either as an icon or a symbol depending on how it happens to be interpreted. Equally conceivably, it may signify in both ways if the interpreter in question interprets it both as an icon and an index. Therefore a computer-based sign "is not only a sign when it functions significatively, but is, strictly speaking as many signs as there are ways in which it signifies. Objects are identified as significative according to their significative functions."

The question is whether there will ever be a completely comprehensive taxonomy which can be used by designers as a basis of icon design and evaluation. This may be unlikely. The American Institute of Graphic Arts (1981) reports that "It is often difficult to classify icons as clearly being of one type or another. It is best to consider representational and abstract as the endpoints on a continuous scale of abstraction." This corresponds with Peirce's view. According to Fisch (1986) Peirce believed that:

> There are not three mutually exclusive kinds of signs: icons, indices and symbols. These are rather elements or aspects of semeiosis that vary greatly in relative prominence or importance from semeiosis to semeiosis. We may therefore call a sign, for short, by the name of that element or aspect which is most prominent in it, or to which we wish to direct attention, without thereby implying that it has no element or aspect of the other two kinds. (pp. 332)

Therefore the representation or form of the icon can be thought of at different levels of abstraction depending on the mapping between it and the referent. As a simple example if Figure 2.27(iii)were used to represent a disk, it would be classified as highly representational, but if it represented the function ‘insert page break’ in a word processing package then it would be considered to be a completely arbitrary representation. It is possible however, that classifying the representation determines to some extent its potential to convey particular mappings. For example, Figure 2.28(iii) has the potential to show a direct relationship between it and an underlying referent. Therefore it follows that Figure 2.28(iii) has the potential to be categorised as an icon, index or symbol. Figure 2.28(ii), shows the representation used to show the metaphorical mapping for the command 'Print'. On the computer interface, it is unlikely that the representation of a foot-print could be used for anything more than a metaphor, however, it could be used to show an abstract or symbolic relationship. Figure 2.28(i) is an abstract representation for the command 'store to disk'. It is unlikely that this form could ever be used for anything other than a symbol.

i ii iii

Figure 2.28: Signs used to represent store to disk, print and disk, respectively

The fact that different researchers managed to find a number of possible sign categorisations suggests that objects such as signs can be categorised in many ways depending on what the person considers to be important. Therefore users, designers and researchers may have different categories.

The important question for designers is what beneficial information can be taken from these taxonomies to be of practical use in the design of icons. The conclusion to be made is that attempting to design suitable taxonomies is not particularly of any practical use to designers. By suggesting that an icon is metaphorical or metonymic or symbolic or abstract does not make the design process any easier than merely suggesting that one icon design is more or less abstract than another. The only practical advice derived from the taxonomies presented is that designers should select icons that appear to be closest to the iconic end of the spectrum. The series of practical guidelines derived from the number of empirical studies performed in this area are of far more use to designers.

2.3.5.6 Practical design issues

Looking from a more practical perspective, what can designers use to analyse and design icons? The only clear assumption that remains unchallenged by empirical investigation is that the more direct the mapping between the surface form of the icon and its underlying referent, the more efficient the icon; in other words, the easier it will be to understand, learn and use (Hemenway, 1982; Gaver, 1986; Rogers, 1988). The implication therefore is that designers should attempt, wherever possible, to design icons that are as close as possible to the real thing thus ensuring that the sign-referent mapping is nearest the representational endpoint on the continuous scale of abstraction. According to Arnheim (1969) however, it is unlikely that designers of graphical icons would ever be able to design a truly representational sign. Kohlers (1969) suggested that they can, but they will only be representational to people who know the iconic code (e.g., the mapping between how something is represented and what it represents). This was the point made earlier in the chapter in the section on user characteristics. However, Arnheim, goes further by saying that "any picture other than a photograph, must be considered partly as an abstraction". Ultimately, the most representational sign is the object itself.

As Arnheim suggests perhaps icons, whenever possible, should be photographs of the object. However as previously mentioned, many computer operations are particularly abstract and do not have obvious pictorial representations. In such cases the relationship between referent and representation is likely to be difficult to convey directly. Consequently, metonomical and metaphorical mappings have often been used. If the analogy or mnemonic is a particularly bizarre one (e.g. the use of a spinning top form to represent "go to the top") then the link may be particularly learnable simply because of its bizarre nature. On the other hand, if the link is rather tenuous and not made explicit to the user it can be counter-productive.

According to Gittins (1986), and Familant and Detweiler (1991), if a sign is abstract users will at least not run the risk of confusing it with something else. If the designer uses an entirely abstract icon set however, the interface is unlikely to be guessable, and users will be required to learn associations between representation and referent through interaction.

This is a rather unsatisfactory situation, giving designers no clue as to how to represent the numerous abstract concepts and objects present. Turning to the psychological literature, it appears that little is known about how visual symbols representing abstract concepts are processed and, moreover, how well they can be understood. The evidence, however, suggests that users generally prefer depictions that attempt to concretise the concept in some way (e.g. Jones, 1983).

Paivio (1971) suggests that this process may occur by means of "associative chaining". For example, the word 'religion' may evoke an image of a church as an associative reaction. Associative chaining can be related to the concept of population stereotypy (Fitts, 1951). The principle of population stereotypy is that a given stimulus evokes consistently the same or similar responses in a large section of the communicating population. In the case of abstract concepts, this response can be achieved through some form of associative chaining. Therefore by asking subjects which images come to mind when they are presented with a certain concept, the responses that occur most frequently (the modal responses) are considered indicative for the response stereotypy for that function.

The immediacy and directness of associative chaining are however affected by the "idiosyncratic associative frequency in the perceiver's experience" (Paivio, 1971). Thus the word 'liberty' may immediately evoke an image of the Statue of Liberty, especially for people living in New York. Furthermore, Paivio suggests that such an imaginal reaction may not in principle be any different from that occurring to a more concrete word such as 'dog'.

From this work, Rogers and Oborne (1987) predicted that the extent to which abstract concepts involved in the functioning of computer systems are able to evoke what she terms "imaginal referents" may depend on how well they can be concretised. However, this process may be affected by the degree of abstractness underlying the concept and the frequency with which it is experienced.

Previous research has shown strong preference for using concrete objects as referents for abstract nouns (Jones, 1993). These concrete objects may be accompanied by conventional abstract symbols such as ticks, crosses and various forms of lines, typically used to represent ongoing actions (Howell and Fuchs, 1968).

Results of several experiments by Rogers (1987) and Rogers and Oborne (1987) show that when asked to generate drawings for a set of abstract verbs similar to the type of command operations used at the computer interface, subjects found it much easier to produce drawings for high imagery, high-frequency verbs (e.g. open) than for low imagery, low frequency verbs (e.g. substitute). Furthermore, the drawings produced for the high imagery verbs were considered to be the most representative and had the highest stereotypic strength. Conversely a diversity of drawings were produced for the low imagery verbs, none of which really stood out as being most meaningful.

An analysis of the form used for the high and low-imagery verbs showed that for high-imagery (invoking an imaginal reaction as described by Paivio, 1971) , high-frequency verbs, subjects produced mainly concrete pictures. For low-imagery, low-frequency verbs subjects used a mixture of concrete and abstract pictures. For changes of states subjects used supplements and universal symbols, for movement they used concrete objects (i.e. action at the moment before its climax), and for conceptual actions they produced abstract shapes and symbols.

For more abstract concepts, where there is no obvious resemblance between surface form and referent, it appears that the most direct form of relationship is one that depicts concrete objects being operated on in conjunction with abstract symbols (Rogers, 1988). The function of the abstract symbols is to provide an indication of the state of the action. For example, Rogers found that the command operations "go to the bottom" and "go to the top" of a page are most effectively conveyed through the depiction of a piece of paper with writing on, together with an arrow outside pointing downwards and upwards respectively. The least direct mappings have been found to consist purely of abstract symbols in which there is an arbitrary relationship between the icon and referent.

Similar findings have been presented by Rohr and Keppel (1984), who found that operations concerned with processing functions (e.g. print) were rated as being more meaningful when the representation consisted of a prototypical object performing this process. Alternatively, commands requiring manipulative control actions (e.g. insert) were best represented by a prototypical action sequence comprising abstract visual symbols (i.e. arrows, lines).

2.4 Limitations of previous research

The main limitation of work done previously, both experimental and theoretical, is that it fails to consider any attribute other that the icon's form as the attribute used to communicate the relationship between the icon and its referent. The ability to represent this relationship is not exclusive to form. Very little work has considered the possibility of other attributes performing this role; exceptions being Arend et al. (1987) and Lansdale, Simpson and Stroud (1990), who have looked at the merits of using attributes such as colour, size and shape. The general finding from these studies has supported the idea that users can use these attributes to relate icons with their referents both through recognition of these attributes and even through recall.

An important feature of icons, identified when defining a computer-based icon at the beginning of this chapter, is that they are rarely displayed in isolation from other icons. If researchers placed more importance on this feature when considering icon design, then it would become immediately apparent that by placing icons within an appropriate context the number of potential relationships between icon-referent, icon-icon, referent-referent and icon-other referent increases dramatically. From this it could be assumed that with more information present, the chances of the icon-referent relationship becoming apparent to the user is increased.

Research applying this principle is almost non-existent. However, Garcia, Badre and Statsko (1994), performing a very basic analysis of icons in isolation compared with icons in a suitable interface context, concluded that:

> Our conjecture here is that icons should be used contextually. People seem to recognize and discriminate between icons better when in a meaningful situation (pp. 206)

The only other apparent empirical work which relates to this idea is the work described by Arend et al. (1987) which suggest that icons should be evaluated in context in order to test for maximum discriminability.

The most relevant document suggesting a possible theory of how information is conveyed by an iconic interface is a technical paper by Steve Draper (1991). He suggests that there are four possible dimensions to consider in icon design:

1. The type of link underlying an informational link.
2. Comparison icons.
3. The contrast set.
4. The amount of learning required.

The first dimension is very similar to the taxonomies described in the previous section. This dimension considers how the representation in isolation relates to the referent. The other three dimensions, however, are virtually non-existent in other theoretical papers in this area. Dimensions two and three both relate to contextual issues and therefore deserve some discussion.

The fourth dimension is also important to discuss since it addresses the second distinguishing characteristic of icons, as defined by the Oxford English dictionary; that people interact with icons over long periods of time.

2.4.1 Comparison icons

Comparison icons contain attributes which are non-existent when the icon is presented in isolation. A typical example would be an icon which varied in size in relation to the size of the file it represented. If looked at in isolation, this potential piece of information would be lost since there is no context in which to compare the icon. By placing the icon in a context with other icons, which contain attributes that also follow this rule, it is then possible to see how big the file is relative to the others in this context. The important point to note here is that comparisons will only work if the interface is consistent in the analogy being drawn upon by the individual icon attribute. This perhaps accounts for the success of the desk-top metaphor.

According to O'Malley (1988) this type of contextual feature is advantageous since by communicating relations at the interface, the user can gain a collective model from which they can infer the structure of the system.

An important attribute which would not exist if the icon were presented in isolation is the icon's relative position. Blankenberger and Hahn (1991) have suggested that position is an important attribute commonly used to discriminate one icon from another. However in order to pinpoint an icon's location one must be aware of various features relative to that icon (i.e. other icons, scroll bars, screen corners etc.).

2.4.2 The contrast set

This is in a sense the opposite to the previous dimension. In isolation all attributes in a contrast icon will be apparent to the user. However, when placed in an interface, Draper claims that the success of the attribute at communicating its message will depend on how greatly it contrasts with similar attributes present in the interface.

An important aspect of the computer interface is that it can only display a limited number of icons at any particular time. This suggests that when the user has a particular task in mind he or she is faced with a forced choice of which icon to select. This forced choice is what Draper (1991) calls the contrast set. The size of the set from which the user has to make a choice will be determined by how many icons they can safely discard from their selection. Therefore if the user wishes to perform the command 'Print' and there is only one icon displayed whose form resembles a printer, then the user can safely discard all but this icon. At the opposite extreme, if all icons are highly abstract and situated in some random circular shape, then the user would perhaps not feel happy about discarding any of the icons and would probably resort to selecting icons randomly.

Therefore the two main points to note about the contrast set are:

- How well an attribute communicates its meaning depends on how well it can be discriminated from the attributes surrounding it
- The number of options there are (the number of icons it has to contrast with) will also affect how well the attribute, and ultimately its meaning, can be discriminated.

How quickly the icon selection can be made from the contrast set will obviously be influenced by previous knowledge about what should be present in specific situations. The major advantage of having a set of beliefs attained from previous experience is that feeble cues can then be sufficient to convey the correct information, therefore exactly how discriminable an attribute appears to the user will not only be context-dependent but also user-dependent.

The main difference between contrast sets and comparison icons is that it is possible to have an interface where there are no comparison icons (e.g. if, for example, position were randomised).

However, as soon as there is more than one icon present there will always be a contrast set which has the potential to guide the user to select the correct icon for the command or object.

2.4.3 The amount of learning required

The ultimate aim of design should be to ensure that the interface is usable. Many researchers, however, have associated usability with how easily the icon's meaning can be interpreted at first sight. However in a realistic setting users interact with icons within an interface over extended periods of time. Therefore, it could be argued that a usable interface needs not just to be guessable but also easy to learn, and ultimately it should be able to support high levels of performance once the user has become experienced. However, as Draper (1991) points out, the speed with which a user learns an interface will be influenced by factors such as the frequency with which they use the interface; whether they are required simply to recognise the appropriate icon, or whether they have to recall it (recall requiring far greater learning than recognition); the degree to which the icon attributes are able to latch onto prior knowledge, and the degree to which the interface corresponds to social conventions that the user expects.

It is likely that there are many potential icon attributes that only become apparent through learning, for example, learning to associate abstract form, location, or colour with the icon referent. Due to the intangible nature of many computing concepts it is unlikely that the designer will be able to create an interface consisting of only icons with highly representational attributes. Therefore, the issue of how well users learn to associate ambiguous attributes with the referent is of obvious importance to the designer. This is discussed in greater detail in the next chapter.

Therefore, the important aspects to be taken from Draper's theory are that the surrounding icons can greatly influence the degree to which the iconic attributes are distinguishable to the user. As a result it can therefore be assumed that by placing icons within an interface the user will find the icon more usable. Also, his theory stresses that many icon attributes are not necessarily easy to guess, but are incredibly easy to learn. Since most users will interact with and interface for a relatively long period of time it is valid to study the influence of these attributes on usability.

2.5 Chapter summary (thesis justification)

This extensive review of the literature has allowed an insight into the current state of icon design, as well as emphasising areas in which research has been rather sparse. Of particular importance is the area concerning the influence of the interface on the ability of an icon to convey its meaning (though its attributes). The influence of the interface can be in two forms; it can either create the presence of additional attributes (creating comparison icons) or it can affect the degree to which the icon attributes are distinguishable from those surrounding it (contrast icons).

This kind of contextual influence shall be defined throughout the rest of this thesis as interface context. It is given this name to distinguish it from the many other forms of contextual influence which this thesis is not concerned with. It is, however, essential that we are aware of the other forms of context so that they may be controlled for when designing a suitable experimental paradigm.

One form of contextual information is the expectations that the users will have relating to the task that they are currently performing. Therefore if typing a letter on a word processor, users will expect to be able to set margins, tabs, print etc. Icons are also influenced by the cultural environment in which they are displayed (cultural context). There is also the issue of previous experience - to what degree does the user's own experiences affect the icon's ability to communicate its intended meaning. These issues, all of which it could be argued, could be loosely termed contextual issues are comparatively well analysed.

Interface context, or the information available though the icon's immediate environment, is predicted to have a strong influence on the ability of any icon attribute/referent mapping to communicate its relationship to the user. Therefore, it may be possible to suggest that any icon which may be classed as having a representational form may, in some situations, be difficult to recognise (for example, if several other icons in the immediate environment contained similar forms). The aim of this thesis is to support the idea that an icon's success should not be estimated by how it performs in isolation, but by how it performs when situated within a context containing a number of other possible pieces of information available to reinforce the association between the icon and its referent.

This would suggest that design principles previously encouraged to be applied generally should not be applied 'mindlessly' but should only be used as 'rules-of-thumb' when making design decisions.

Another, important but as yet unexamined issue raised by the literature review, relates to the usability of the icons and their attributes over time. In other words, how quickly can a user learn to use non-guessable attributes and do they eventually reach the same level of performance using these attributes than if they had been able to rely upon guessable attributes? The second main aim of this thesis would therefore be to show that users continue to learn new attributes (other than form) and may even shift from using one attribute to another if it offers a better performance advantage.

A multi-component definition of usability, created by Jordan et al. (1991), is used as a starting point to begin to analyse the second thesis aim; to explore the issue of icon design in relation to extended user performance and the many questions which arise from it. A discussion of the multi-component definition and related work is presented in greater depth in the next chapter.

3. Usability

The aim of icon design should be to make the interface easy to use. This chapter therefore considers in some depth what the term usability actually means. Jordan, Draper, MacFarlane, and McNulty (1991) have illustrated, through their multi-component definition of usability, that when the entire learning curve is considered in design potential benefits to both the design and the evaluation of interfaces become apparent. Two main points are asserted in this chapter. Firstly, by applying a multi-component definition of usability it becomes apparent that the commonly held assumption, that a 'good' icon is a guessable icon, is far too restrictive to be regarded as a valid design principle. The second assertion is that a design applying to usability throughout the learning curve may result in a more usable interface overall.

3.1 A definition of usability

An important goal of HCI evaluation has been to establish a framework for taking measurements which will effectively characterise the usability of an interface. Such a framework would aid comparisons of an interface against other interfaces or 'benchmarks', and could support design improvements. The work of contemporary standards committees, such as the International Standards Organisation (ISO; Brooke, Beven, Brigham, Harker, and Youmans, 1990) and a number of earlier studies (e.g. Eason 1984, Shackel 1986) are steps towards this.

The ISO (Brooke et al 1990) define the usability of an interface as "the effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in a particular environment." This definition recognises that usability will differ between tasks and between users. However, as Jordan (1993) points out, the definition does not discuss the idea that usability can also vary within tasks or within users. For example, the task performed may differ when the user has an extended period to complete a task, compared with when they have a short deadline to complete the same task. Similarly, the usability may appear to differ between a user who has had one hour's sleep and one who is fully awake. Such variations are particularly transient and are therefore difficult to monitor. However one factor, which can be monitored, and which does vary both within users and between users, is the level of interface experience.

This gradual change in the degree of user experience has rarely been considered by researchers in usability. However Jordan, Draper, MacFarlane, and McNulty (1991) have developed a three component framework, designed to account for changes in an interface's usability in relation to user experience. The three components are outlined in detail in the next section. Although derived from the very general concept of a learning curve, they are oriented towards empirical measurement and are not concerned with any particular psychological theory of learning. The components are also not specifically contextual in orientation.

3.2 A multi-component definition of usability

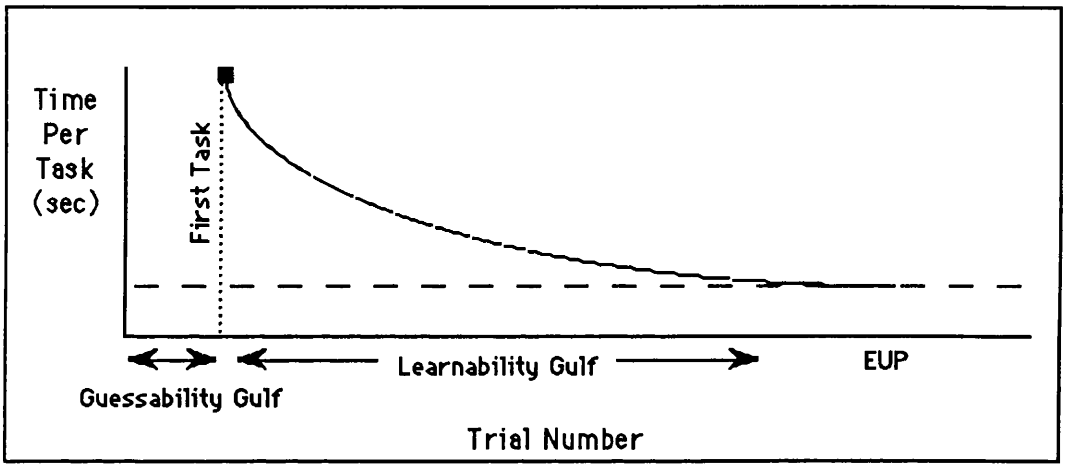


Figure 3.1: The learning curve with the three stages of usability illustrated

One of thc central aims of this thesis is that users will learn different attributes throughout the learnability curve. It is therefore important to identify at which stages this learning occurs. A suitable model to express this information is the multi-component definition of usability proposed by Jordan et al. (1991). It is this model that will be referred to when describing the stages of learning in experiments 1 to 6. The diagram of the three stages of usability and their correspondence with the learnability curve, as described by Jordan et al. is illustrated in Figure 3.1. The three stages: guessability, learnability and experienced user performance are defined below.

3.2.1 Guessability

Guessability is defined as the measure of the cost to the user involved in using an interface to perform a new task for the first time. The lower the cost, the higher the guessability (cost can be measured either in terms of time, errors, or effort). Most of the work on icon design has focused exclusively on how guessable the interface is. In cases where interfaces are going to be used by a high proportion of first-time users then guessability is important, for example on public information systems. Also, a lack of guessability could put users off an interface which in the long run might be comparatively easy to use.

Guessability is of less importance in situations where procedures are initially demonstrated to the user, or for interfaces which will only be used by experts after long training. This might include, for example, military equipment, and aircraft controls. However, it is important that emergency diagnosis and recovery procedures should still be guessable for pilots. Even though users may be expert generally, there may still be rare tasks which they have never seen before.

3.2.2 Learnability

Learnability is the measure of the cost to the user in reaching some reasonable level of performance on a task, but excluding the special difficulties of completing the task for the first time. A highly learnable interface would be one where a task was instantly memorable once the method had been shown to the user. Conversely, interfaces which cause 'interference' with user expectations are likely to be un-learnable.

Learnability may be particularly important where a user is to be 'self-taught' with an interface, or where training time is short. For example, a temporary secretary may be introduced to new word processing packages on a fairly regular basis. Clearly, significant amounts of working time will be wasted if he or she does not reach a reasonable standard of performance fairly quickly. Learnability is of less importance where training time and resources are more plentiful, again a pilot learning to fly an aircraft would be an example of this.

The term learnability has already appeared widely in the usability literature. However, its meaning has not always been clearly expressed, and different researchers have assumed different definitions for it. For example, Payne and Green (e.g. 1986) use the term with reference to completing new tasks for the first time (defined by Jordan et al. as guessability), whilst others (e.g. Shackel 1986 and 1991) use the term more loosely, to refer to the performance by any user who could not be deemed experienced. Indeed, the idea that there is a distinction between usability for an experienced user, and usability for one with less experience is not uncommon. However, the distinction is usually couched in rather vague terms, with little attention to the guessability / learnability distinction proposed by Jordan et al. It is however this guessability / learnability distinction that is of particular interest in icon design. So far this distinction has determined the boundary between what people have considered to be good icon design and bad design. However it would be interesting, and of significant importance, to see whether over an extended period of time icons defined as learnable can result in similar performance levels to those icons defined as guessable.

3.2.3 Experienced user performance (EUP)

Experienced user performance is a measure of the ability of the user to perform a task when he or she has reached a relatively steady level of performance. Again, the lower the cost, the higher EUP. Although this may vary over long time scales, it may be expected to remain comparatively steady when compared to performance during the learnability phase. This component is often what is meant by practitioners when they use the term 'usability' (e.g. Bennett 1984, Eason 1984, Shackel 1986).

EUP will be comparatively important in situations where constraints on training time and resources are few, but where it is important that the experienced user makes few errors. Suitable examples would be driving a car, or perhaps using a specialist software package, for example for computer aided design (CAD).

3.3 Benefits of using a multi-component definition of usability

There are two main benefits to employing the multi-component definition to usability. The first benefit is that, being a more structured definition, it allows both the evaluation and design of interfaces to become more tailored to particular stages of usability.

In terms of interface evaluation, the paradigm employed can be altered to test the specific level of usability. Therefore if interested in guessability, subjects should not be given the opportunity to practice the experimental tasks beforehand, whereas if EUP were of interest subjects should be well practised in the experimental tasks. This may cast doubt on some of the 'traditional' approaches to usability evaluation, where measures of performance are taken after apparently arbitrary practice opportunities. If, for example, a user were given a practice session, during which they were free to explore an interface, including, possibly, the experimental tasks, it might be difficult to know what their level of experience on a particular task was when the experimental session commenced. Without knowing this it would not be clear whether guessability, learnability, or EUP were being measured.

In terms of design the multi-component definition can allow designers the ability to classify design options as being beneficial for particular components of usability. Therefore if designing a piece of specialist software to be used by highly trained users on a regular basis, it may not be worth the effort and expense of designing a highly guessable or learnable interface, provided high EUP is supported. In practical terms this might mean, for example, designing a command key based interface, rather than one with menus. Conversely, an interface for a walk-up-and-use health information system should support high guessability, but need not support high learnability or EUP.

The second benefit of using this framework is that it allows the possibility of examining iconic attributes that may influence usability beyond the guessability stage. By defining usability in three stages it becomes apparent that most of the work on icon design has only focused on achieving icons that are usable at the first stage; guessability (e.g. Rogers, 1989b, Maissel, 1990). It is therefore important to extend the research and to focus on the later stages of usability. Although representational form will increase the likelihood of the icon being considered guessable, it is possible that in later stages of the usability curve different attributes have increasing influence on determining the usability of the interface.

Therefore, by using the multi-component definition of usability it may be possible to account for which attributes are important at each stage and to note that not all successful attributes are required to be guessable. Of the many properties of an icon that influence usability including, for example, form, size, position, shape, hue, and consistency, some may have greater effects at any given stage than others. Even if the designer cannot cater for all potential levels of experience at least the framework creates a frame of reference for any trade-offs which might have to be made.

3.4 Limitations of the multi-component framework

There are two main drawbacks to this framework, which become apparent when it is applied to any non-experimental situation. Firstly, in real life situations, users are unlikely to be experienced in all features within the interface. Some features will rarely be used, and others never. Therefore there will always be a situation where users are experienced in some features, novice in others and naive in a few more parts of the interface. This is particularly true when the interface has any degree of complexity. For example, many users of Microsoft Word may be experienced in using the command 'Save', but be described as novice when faced with the command "Insert TOC entry". There are however examples, such as a public-information system, where tasks throughout the interface would be performed in the same manner.

A second limitation is that, again in real life situations, it is likely that there will be a degree of inter-user variability. The major difference between users is that they bring to the interface greatly differing amounts of previous experience. Baxter and Oatley (1991) identified three types of user experience which may affect performance when using an interface.

Task domain experience is gained from experience of performing a task, irrespective of the environment in which it is performed. So, for example, experience of writing the word 'computer' will be helpful whether the word is being written by hand, typed, or word processed.

Operating environment experience can be gained through familiarity with a particular 'type' of interface. For example, the Apple Macintosh has conventions that are used in many different types of application program. Commands for saving work and printing can usually be found on a menu headed 'FILE', irrespective of whether the application is, for example, a word processor, or a statistics package.

Application program experience is gained from experience of using a particular application program, perhaps in a different operating environment. For example, a particular graphics package may employ similar icons regardless of the operating environment in which it is run.

Differences between users with respect to each of these experience types can cause marked differences in performance. For example, in a comparison of two different Macintosh-based spreadsheet packages, Baxter and Oatley found that the level of task domain experience which first time users brought with them far outweighed any differences between the packages as a factor affecting performance. Thus, in this case, the usability of the machine was very much user-dependent. Presumably the other types of experience could have equally dramatic effects on the usability of interfaces for some tasks.

Another example is where the task can either be performed in perfect environmental conditions or done under stress. The fact that work and attention will suffer under stress suggests that interfaces should be tested under such conditions. This way, when in perfect conditions, results can only be better. Does this mean that an interface that performs the best in perfect conditions will still come out on top in hectic conditions? Don Norman (1991) believes that under stress, it will be the highly representational icons that will do best. Is this still the case when the representation is poor but all other cues are maximised? Although experiments in this thesis do not focus on these question they remain interesting issues for future research.

It may be argued that there is a possible "Pandora's box" of exceptional cases to this framework. It can also be argued that the drawbacks of both inter-task and inter-user variation are common to all situations where results from experimental analysis are put into practice. However, provided investigations are designed in such a way that subjects with matching levels of previous task domain experience are equally divided between conditions and that there is an equal distribution of realistic tasks, meaningful overall comparisons between interfaces, and also meaningful statements about an interface's overall usability can be justified.

3.5 Usability and iconic interface design

The point that is continually returned to is that the main measure of good interface design is *increased* usability. Increased usability is exhibited by faster learning, indicated by quick, stable reaction times and low error rates. In the previous chapter is was suggested that a truer measure of an icon's usability can be discovered when the icon is not tested in isolation, but placed within a realistic interface, amongst icons and other interface features. The suggestion for this increased usability was that by placing the icon within a suitable context the number of attributes and pieces of potential information being conveyed to the user dramatically increases. Increased information should allow the user to make a more informed judgement as to what any particular icon may mean.

Similarly, an even more accurate measure of an icon's usability could be gained by evaluating user performance throughout the entire learnability curve. Many abstract attributes (e.g. form, location, colour, shape) will not be guessable, but it is possible that their meaning may be extremely learnable. So much emphasis has been placed on icons being usable at the guessability stage. However, the graph in figure 3.1 suggests that guessability only appears to apply to a small percentage of the total learnability curve By ignoring any non-guessable attributes, designers are potentially ignoring design issues that may have greater lasting consequences over the majority of the learning curve.

By considering usability in the longer term many questions relating to icon design become apparent. For example:

• Can attributes that are not guessable, but are learnable, result in equivalent performance levels to guessable attributes at the EUP stage?

• Does a user's performance improve over time because they:

a. build up a more complex picture of the icon, as the number of attributes they are aware of increases, thus reducing the chances of error?

b. learn more attributes and swap from the guessable one to one which offers a better performance advantage?

c. learn one significantly distinguishing attribute and ignore any others that may become apparent over time with performance improving simply due to practice at recognising this attribute?

To expand upon what was stated in the previous chapter, there are two parts to the thesis aim:

1. To analyse icon usability within the context of a richer interface, as was discussed in Chapter 2.
2. To assess icon usability over an extended period of time. The aim would be to measure user performance until EUP level.

The main motivation for this would be to assess how well users would perform throughout the learnability curve using either guessable or learnable attributes. From this is would be hoped that some answers to the type of question raised above could begin to be supplied. By answering these questions it is possible that the scope of icon design could be widened to include more practical guidelines on how to create guessable and learnable icons that perform well from within a suitable context.

3.6 Thesis Hypothesis

It was decided that icon position would be a suitable attribute to look at. This is important since it is an attribute which is only present when the icon is placed within a context, (defined by Draper, 1991, as a comparison attribute of the icon). It is an especially useful attribute to study since it is one present in most current interfaces, yet its importance is perhaps under-credited. To look at more than one attribute it would become confusing as to which attribute was being used when. It would be hoped that results from looking at the inter-relation between two attributes could be extended and applied to more complex interfaces.

For the purposes of this thesis, position will be used as a learnable attribute. Icon form will be presented as the representational, and therefore guessable attribute. Based on suggestions made in the previous chapter and the questions raised above, five general propositions are suggested for research:

1. Representational icon attributes are easier to learn than abstract attributes.

It is expected that attributes that are considered to be highly representational are also the most guessable. This reinforces the idea that if icons are to required to be guessable then the designer should attempt to use the most representational attributes to convey the icon's meaning.

2. Users presented with representational icon form do not rely on icon position.

This proposition attempts to answer question c) above. Do users learn one significantly distinguishing attribute and ignore any others that may become apparent over time? This may be a good strategy for the user since it requires little cognitive effort.

3. Users presented with abstract icon form will rely on icon position.

This suggests that people will use and learn a non-guessable attribute, thus suggesting to designers that if an icon contains no guessable attributes, it does not mean that it is unusable. Of course, abstract form should also be potentially learnable, thus to specify what this hypothesis is suggesting, if users learn position as opposed to form, it should be because, in this scenario, recognising location requires less cognitive effort and greater performance advantage than learning form.

4. Guessable and learnable attributes will achieve equivalent performance measures at EUP level

This further adds support to the idea that non-guessable icons are still valid in design.

5. Users are aware of both form and position, but rely on one more heavily than another.

Whether users always rely on the most guessable attribute, or change to rely on the learnable attribute may depend on whether there is a performance advantage achieved by making such a swap. Alternatively users may continue to rely on only one but be aware of the other attribute in case the dominant attribute is disrupted or changed. This is a rather more complex proposition which attempts to answer questions a and b above. Therefore it requires a greater detail of analysis than any of the other propositions. This aim challenges the belief held by researchers such as Nickerson and Adams (1979) that users economise on learning and only learn the minimum amount of information required in order to identify the appropriate icon. This proposition, if supported, would suggest instead that people learn everything but may rely on one attribute more than others. The advantage of this is that if the attribute which is being relied upon becomes inconsistent it should take a limited amount of effort to switch to using an alternative attribute to identify the icon.

4. Experimental Paradigm

4.1 The paradigm employed

4.1.1 Motivation

Six experiments were performed in total. The initial hypothesis contained five, basic propositions:

1. Representational icon attributes are easier to learn than abstract attributes.

2. Users presented with representational icon form don't rely on icon position.

3. Users presented with abstract icon form will rely on icon position.

4. Attributes that are learnable will achieve equivalent performance times and error rates to those achieved using guessable attributes by EUP level.

5. Users are aware of both attributes, but rely on one more heavily than another.

Experiment 1 launched the empirical research in this area. However as the research continued, there was increasing awareness that the processes involved in determining which attributes users rely upon are far more complex than one would initially assume.

Experiments performed by Blankenberger and Hahn (1991) were used as a starting point for the development of the design paradigm. They reported several "search and select" experiments designed to evaluate the reaction times of subjects for three icon sets of "different articulatory distances" (different types of icon form along the "continuum of abstractness"). The first of the experiments randomly placed the icons each time. The second experiment held the stimulus positions constant, since this was considered to be a more natural setting for an experienced user. Their results showed that in the random condition, which Blankenberger and Hahn considered to represent the type of task a novice user would face, there was significant difference between the three sets of icons for both reaction time and error rate. However, in the condition where the sign position was fixed, there were no differences between icon types. An error analysis was done showing a significant effect only for symbol position ($p = 0.016$). Their findings were as follows:

> There used to be an intuitive agreement stating that icons should be graphically simple, should differ in global features and should have some kind of "articulatory directness". We failed to find

> clear support for that. Using fixed screen positions our subjects seemed to establish kind of "local bindings", and connected functions with positions.
>
> (p. 373-374)

Using the icons created by Blankenberger and Hahn (1991; see Appendix 1), a similar experimental design was employed for this thesis. As the series of thesis experiments progressed the design paradigm became increasingly refined. Refinements made for specific experiments will be discussed in depth as the experiments are reported. What follows is a summary of the general paradigm used by all experiments.

4.1.2 Apparatus

The interfaces were created using Hypercard (version 2.02). All experiments, apart from experiment 6, were run on a Macintosh LCII using system 7. Experiment 6 was performed on a Macintosh Quadra, using system 7.1.2.

4.1.3 Design

For all 6 experiments the dependent variables were:

- Time taken in each trial
- Number of errors
- User opinion about the tactics they employed

For experiments 4 and 6 the number of pre-emptive moves was also used as a dependent variable.

The independent variables that remained constant for all 6 experiments were:

Icon form	**2 levels**
	Representational
	Abstract
Position	**2 levels**
	Consistent
	Random

In most experiments subjects were presented with an interface until they had reached EUP. Users were considered to have reached EUP by the time performance had reached an optimal, steady level. The initial interface would either contain all representational or all abstract icon forms. Before users had reached EUP the location of all icons remained constant. Once reaching EUP level the interface would change unexpectedly as shown in Table 4.1.

	Trials to EUP	Experimental Trials
Representational-to-abstract	Representational form Constant position	Abstract form Constant position
Abstract-to-representational	Abstract form Constant position	Representational form Constant position
Representational-to-random	Representational form Constant position	Representational form Inconsistent position
Abstract-to-random	Abstract form Constant position	Abstract form Inconsistent position

Figure 4.1: Icon attributes present in each of the four experimental conditions. (The name of each condition refers to the state of form and position after the changeover)

For the representational-to-abstract and abstract-to-representational conditions, the initial hypothesis was that performance in the latter condition would be less affected by the change-over in representation since it would be predicted that subjects had not been relying on the representation, but rather the position. If they had not learned position, then the increase in performance would be significant at this level.

For the representational-to-random and abstract-to-random conditions the hypothesis was that the latter condition would be significantly affected by the change in position since, in this case, it was predicted that the subjects were relying entirely on position as a cue. However for the representational-to-random condition there ought not be such a significant difference since users are expected to rely on the representation.

Therefore the experimental design for all experiments should be:

One Within-group factor	Training v's Experimental Trials	(2 levels)
One Between-group factor	Between Conditions	(4 levels)

Experiments 1 to 6 did however differ in the size of the icon sets displayed, the size of the penalty for an incorrect response (explained in section 4.2.1), and the number of trials. The exact value for each of these variables for each experiment are recorded in the table in Figure 4.2.

	Size of Icon Set	Size of Penalty	Number of Trials (per block)
Experiment 1	17	Large	5 before and 3 after
Experiment 2	2	Small	20 before and 20 after
Experiment 3	2	Small	5 before and 5 after
Experiment 4	2	Small	20 before and 20 after
Experiment 5	2	Small	20 before and 20 after
Experiment 6	16	Small	20 before and 20 after

Figure 4.2: Comparison of the size of icon set, penalty, and number of trials for experiments 1 to 6

The terms 'before' and 'after' refer to the number of trials that occur prior to the changeover and preceding it. Furthermore, a block of trials is completed when all icons in the set have been displayed once. Therefore for experiment 1 there would be 5 x 17 trial before the changeover and 3 x 17 trials after.

4.1.4 Subjects

Subjects were required to be experienced users of word processing packages to ensure that they had some previous knowledge of the commands that they were being presented with. It was also important that subjects had some experience with iconic interfaces, in order to be familiar with the physical task of moving a pointer round the screen and clicking on icons to select them. All subjects had either normal or corrected-to-normal vision. Both male and female subjects were used, although the majority of subjects for all experiments were male. Subjects were mainly from undergraduate and postgraduate computing courses at the University of Glasgow. Only recruiting subjects from the student population was not thought to effect the results since the visual appearance of the interfaces constructed were not based on any previous interfaces, and therefore all subjects would be equally inexperienced, and have relatively little task domain knowledge to apply to their performance.

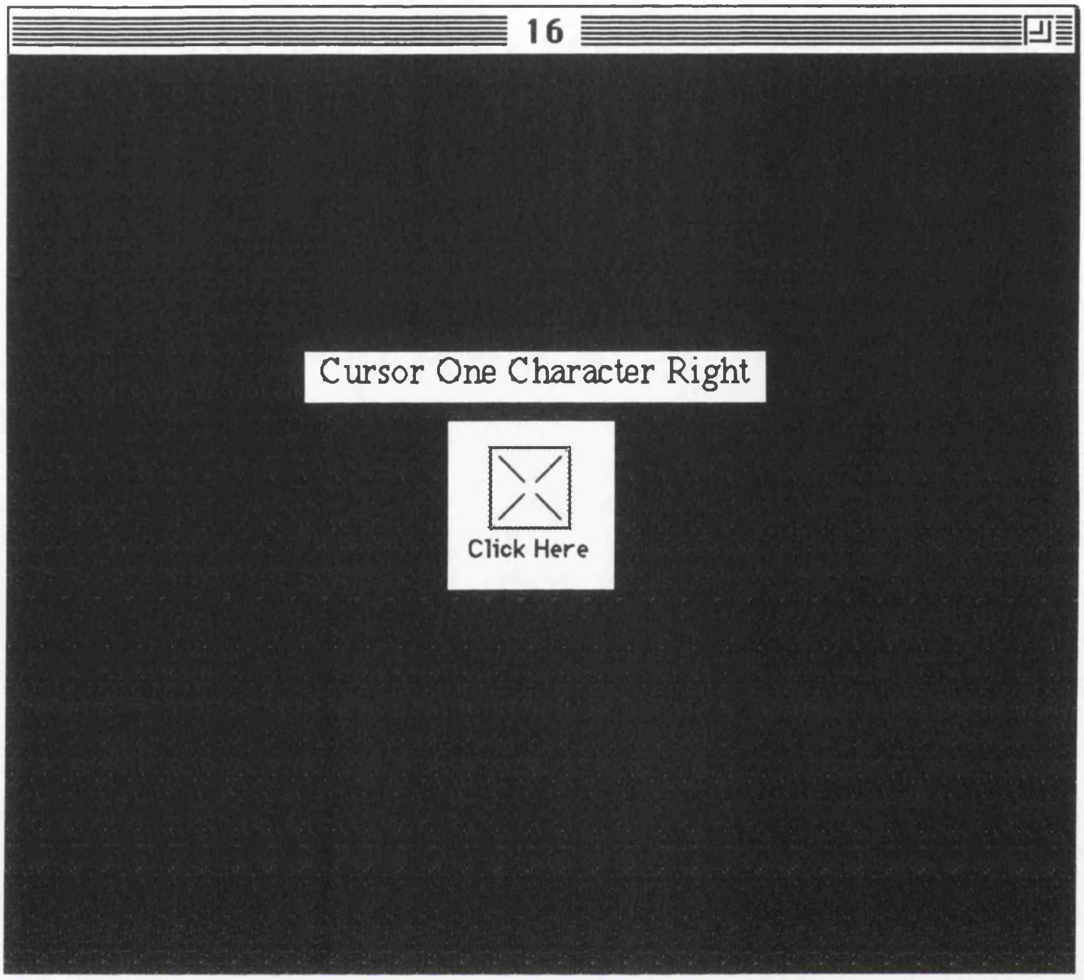


Figure 4.3: Card containing the command and centre button

4.1.5 Procedure

Subjects were initially presented with a block of practice trials to familiarise themselves with the experimental task. Although the task itself was identical to that performed in the experiment a mock experimental interface was created where the icons in each of the positions were replaced by numbers (e.g. 1), and the commands were in textual format (e.g. Number One).

For all experiments, with the exception of experiment one, the trial commenced when the command and centre button appeared (see figure 4.3). Once the user clicked on the button the interface card opened, displaying an icon set. The icons sets for both the representational form and abstract form are presented in Figures 4.4. and 4.5. The icons are positioned in the circular format used in experiment 6. The subject was required to respond by clicking the mouse over the icon he or she considered to be appropriate for the command. If a correct response were made the set disappeared and approximately 0.5 seconds later the next trial commenced. An erroneous response was signalled by a high tone and recorded. This tone informed the user that he or she has selected an incorrect icon, and to select again. The exception to this is the 2 icon set, where the subject obviously knows what the correct icon should be after selecting the incorrect one. Therefore in this case subjects were only allowed one chance to select the appropriate icon before the penalty was induced. This penalty is described in more detail in section 4.2.1.

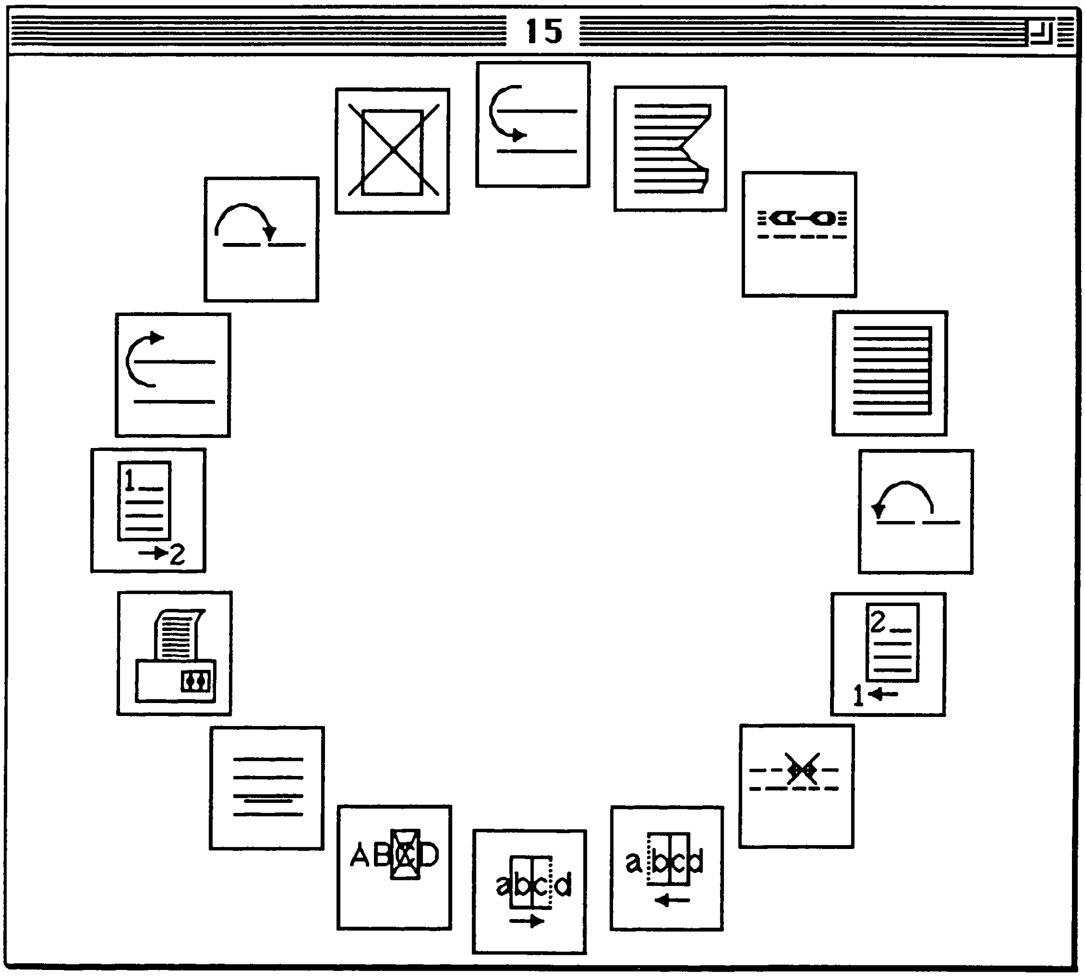


Figure 4.4: Set of 16 representational form icons positioned in the circular formation used in Experiment 6

Subjects performed a pre-determined number of blocks of training trials and a similar number of blocks of experimental trials. This number varied between experiments.

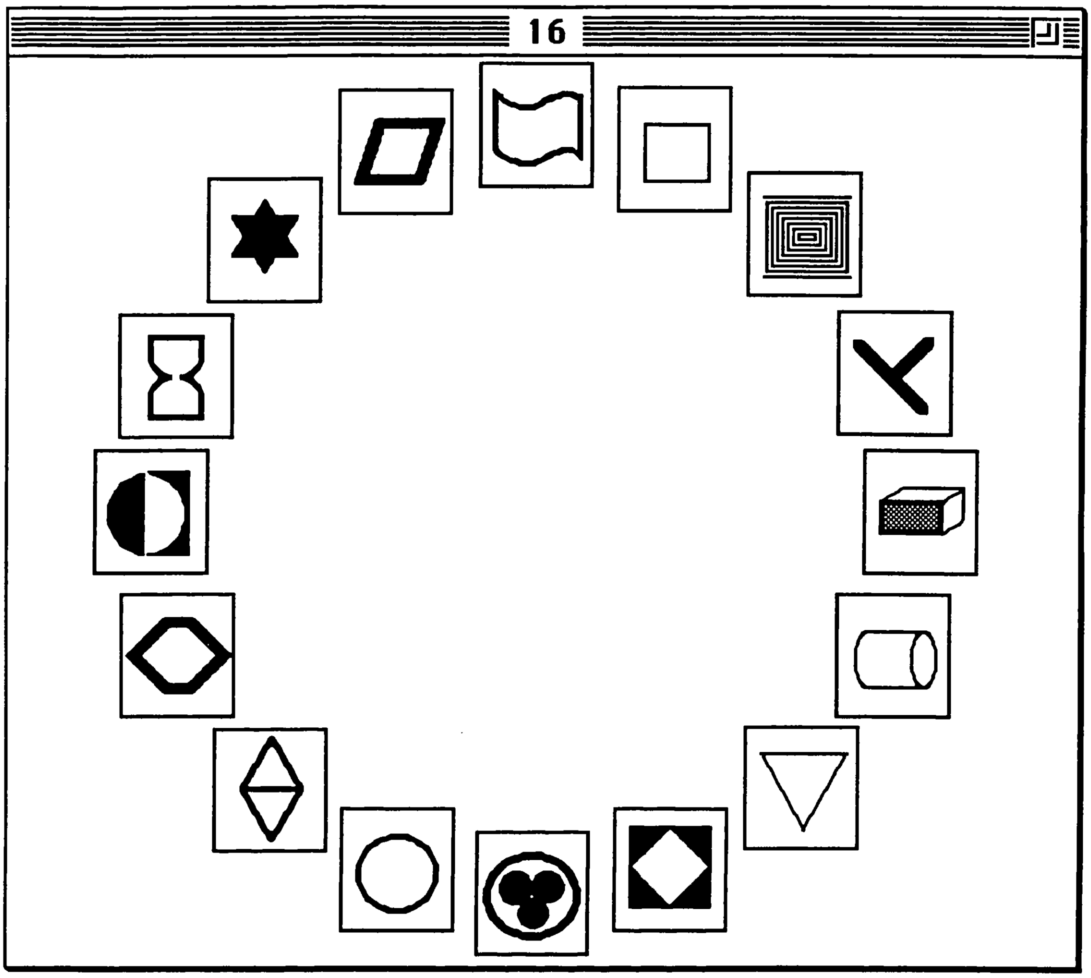


Figure 4.5: Set of 16 abstract form icons positioned in the circular formation used in Experiment 6

At the end of the experiment subjects were presented with a questionnaire (see Appendix 1). The aim of the questionnaire was to allow subjects to comment on the techniques they used and the attributes they may have felt they relied upon to identify which icon represented which command.

4.2 Refinements to the paradigm

The experiments described in the next three chapters are presented in the order in which they were performed. This is done to emphasise the development of the argument from which the revised hypothesis as discussed in Chapter 8 derived. Over time several modifications were made to the paradigm. These modifications are discussed in greater depth in the next four sections.

4.2.1 Penalty induced

For experiment 1 the penalty or punishment for making an error was large. Users were given no help if they were unable to select the appropriate icon. Instead they were required to continue to search and select icons until the correct one was found. Therefore subjects could spend several minutes trying to select just one icon correctly, resulting in higher error rates, longer reaction times, and also increased stress levels. Experiments 2 to 5, however, only use sets of two icons. Therefore if the user selects the incorrect icon then the correct icon must be the other one. Therefore there is no cost associated with making an error. To discourage people from making errors some sort of penalty was required. Therefore, in the 2 icon experiments, a delay mechanism was installed whereby the subject would be shown the correct icon before being allowed to move on (see Appendix 1). For the last experiment, using sets of 16 icons, this mechanism was again used, this time to reduce the amount of time users may spend looking for any icon, thus allowing then to perform a large number of trials in a far shorter space of time than had been accomplished in experiment 1. However, in this case the penalty as perceived in experiments 2 to 5 was more likely to be seen as a help device. These difference must be kept in mind when directly comparing experiments 1 and 6 with any of the other experiments.

4.2.2 Introducing a user controlled button

The first experiment remained more faithful to the experimental design proposed by Blankenberger and Hahn in that the trial commenced with a tone, followed 0.5 seconds later by one of the 17 commands displayed textually. After 1.5 seconds the command disappeared and was replaced by the appropriate set of icons. However on performing this experiment it became obvious that there were frequent cases where the subject's concentration had lapsed at the point where the next textual command was being shown. As a result subjects failed to perceive the textual command and therefore were unable to perform the task.

The solution adopted in all later experiments, was to introduce a button (referred to as the centre button) which appeared on the same screen as the textual command which the user was instructed to select when they were sure of the command (as in Figure 4.3). An additional advantage to the use of this button was that it ensured that the pointer was at an equivalent starting position at the start of every trial, thus reducing any bias in reaction times as a consequence of variation in the distance travelled by the mouse.

4.2.3 Icon position

Corresponding to the experimental design reported by Blankenberger and Hahn (1991), icons in experiment 1 were positioned in a rectangular formation around the screen edges (see Appendix 1).

However it was thought that less skill may be required to point the mouse at icons situated in these areas since users would be aware that they could drag the pointer as fast as they wished in the knowledge that it would 'bounce' at the screen edge. It is perhaps even easier to direct the mouse to a corner, since the pointer then has two sides from which to bounce. Furthermore, because the card window is rectangular, rather than square, icons at the left and right are further away from the centre of the card than those at the top and bottom. Therefore, for the two icon sets (in experiments 2 to 5), icons were situated away from the sides, but equidistant from the centre button (see Appendix 1). For the 16 icon sets (experiment 6), the icons were placed in a circular format, again equidistant from the centre button (as shown in Figure 4.4).

4.2.4 Details Recorded

For every trial eight pieces of data were captured. These were:

1. The errors. Each icon had a number, and the number of the correct icon was stored, followed by the number(s) of any incorrect icons that were selected.

2. The time that the command and the centre button appeared.

3. The time that the subject clicked on the button to move to the interface.

4. The (x,y) position of the pointer on the screen as the card closed. This showed whether or not the mouse was outside the button area, and in what direction it was travelling. This data was collected from experiment 2 onwards.

5. The time that the card closed.

6. The (x,y) position of the cursor as the interface card opened (in experiment 6 only).

7. The time that the first icon was selected (and whether it was correct or incorrect).

8. The time that the interface card closed.

A typical reaction time was measured from the time that the command appeared until the user had selected the correct icon, or been informed of its position, and the interface card had closed. In experiment 6 however a 'pure' measure of reaction time was recorded, taken from the time that the user clicked the centre button until he or she had selected the first icon (irrespective of whether the icon selected was correct or not).

4.3 Requirements for a valid paradigm

Kantowitz (1992) suggests that although empirical research should "not be done by rote" there are several important points which must be considered when designing an empirical evaluation. If these points are used in the form of a checklist, then any results gained from the resulting paradigm should be valid. Therefore, in order to check the validity of the experimental paradigm just discussed, the paradigm should be considered in relation to an adapted list of the points suggested by Kantowitz. The points are as follows:

1. The paradigm should be both valid and reliable.

The independent variables (the variables that are manipulated by the experimenter; the manpulation expected to cause a result) considered over 6 experiments were:

- Icon form
- Icon location
- Number of icons present on the interface
- Penalty induced
- Number of trial blocks

The first two variables are of central concern to this thesis. A measurement of the effects of icon form and icon position on user performance can be attained by comparing the error rates and reaction times (the dependent variables - names so, since their values are dependent on the manipulation of the independent variables) between users in the four experimental conditions. The other three variables almost certainly have a pervasive influence on results. Therefore, by actively designing experiments to contain these extraneous variables it is possible to analyse their influence by comparing performance across experiments.

The fact that results were often replicated for several experiments does suggest that the paradigm was indeed reliable.

The validity of testing these dependent variables should be supported by the volume of previous work which has also considered form and position as important factors in icon design. A criticism could be to ask why look at "yesterday's icons" when now there are animated icons and 3D icons available. However Dan Venolia (1993), talking about 3D interfaces at the INTERACT'93 conference, suggested that in order to design these more complex icons we must use the design principles for more traditional icons as the basis. This means being aware of how to design 2D, static icons. As the literature review has shown, this has not been fully achieved yet.

2. The empirical study should also contain a high degree of subject and setting representativeness.

Suitable subjects were considered to be people who had previous knowledge of word processing packages and also of iconic interfaces. These were essential requirements since the subjects should have some concept of what the word processing commands mean to be capable of selecting appropriate icons. Also, they must have had some experience of interacting with iconic interfaces and using a mouse so that results were not affected simply by user inexperience with the physical task of moving the mouse around an interface.

All subject were University students, and therefore it could be argued that they were unrepresentative of the entire population. However there experiences were perhaps representative of typical computer users (with a continuum of experience ranging from first year undergraduates who had used computers only for a matter of months up to Ph.D. students in the Department of Computing Science, who had used computers extensively over a number of years). Since the interface was constructed only for this experimental paradigm no subject would have had prior use of it. Therefore, in this respect, all subjects began at the same level of experience of the interface. The spread of possible ages and intelligence is reduced by using just students. However, many of the tasks involved in the experiments are not intellectually challenging (e.g. requiring people to learn only 2 icons) and rely more on visual ability rather than intellectual. Visual acuity does decrease with age, but this can be corrected. It was a requirement that only subjects with normal or corrected-to-normal vision could take part in the experiments.

The setting itself was obviously not a natural one. Subjects were aware that they had to perform the set number of tasks quickly and efficiently. The experimenter was also present throughout the entire session. This however is a common problem with experimental design and is not easily solved. The interface being tested was, however, not designed for any particular user group, or specific situation, therefore the fact that the setting for the experiment was not so naturalistic is perhaps not a large design flaw.

3. Are the dependent variables fundamentally relevant to users' behaviour?

The essence of the experimental paradigm is to measure the usability of icons. As discussed in chapter 3 usability means ease of use. The easier an interface is to use, the faster users should be able to perform, and the fewer errors should occur. In real-life situations people may not necessarily be concerned with being able to perform quickly, but they do wish to use the interface with as little effort as possible. Therefore if reaction time and error rate are indicators of ease of use the dependent variables must be relevant.

A problem which could arise to affect the relevance may be whether subjects stress accuracy of performance, or speed of performance, or both as important to achieve. In the experimental overview ·(see Appendix 1) users were asked to be both fast and accurate. The penalty incurred from an error may have an influence on whether users considered it worthwhile to be accurate or not. Also, having an experimenter present to watch their performance encouraged users to react promptly. Therefore in cases where the penalty was seen as a definite penalty (experiments 1 to 5 perhaps) the balance between reaction time/ error rate trade-off was probably acceptable.

4. Have all control variables been considered and held constant?

As already discussed, many of the variables likely to affect results were controlled by manipulating them across experiments and analysing the effect. Other extraneous variables include user differences; e.g. degree of motivation, degree of wakefulness. Hopefully these differences have been reduced by the sample sizes of subject used.

Practice effects were prevented by randomising the order in which subjects were presented with commands. Also the icon position was randomised between users so that the chances of any icon performing better simply because it was in a more distinct position was eliminated.

5. Is the paradigm valid? Does it try to mirror real-life conditions and requirements as closely as possible?

In a real-life, iconic interface it has been suggested that people use recognition, rather than recall, in order to select the appropriate icon. The "search and select" paradigm in the proposed paradigm is actually testing recognition or icon comprehension. This paradigm presents users with a task and then asks them to select the appropriate icon. Many paradigms used previously in icon research (e.g. the comprehension test; Howell and Fuchs, 1968) do the opposite; presenting subjects with an icon and asking them to suggest what it represents. This is completely unrepresentative of real -life performance and therefore the "search and select" paradigm used in this thesis should be the preferred paradigm.

It could be argued that the tasks themselves are not particularly realistic. Although the commands are present in existing word processing packages, users would often interact with the icons by deciding that they had a task to perform and then look for the icon representing that command. Therefore, in the experimental condition, users were performing the commands out of their "context of use". This was essential, since in order for subjects to achieve EUP level many trials had to be performed. To control any variations in the frequency of use and to encourage users to activate commands so many times would be extremely time consuming. Furthermore, this may not be a negative feature since it increases the likelihood that all subjects had the same task in mind (e.g. select icon the refers to the command for printing a word-processing document) and therefore may have actually controlled for any variation as a result of user differences.

It is also un-realistic to expect users to perform all tasks with equal frequency. However it was considered important to control this feature since it may have made comprehension of the results more difficult - deciding whether the results were due to the attributes present or simply the frequency of use.

Unlike most icon studies, the icons are tested within a context of other icons therefore, in this respect, the paradigm employed is far more realistic than many other methodologies documented.

6. When looking at guessability, learnability and EUP, have some kind of criteria quantifying each of the usability stages been considered?

This point was not present on the original list by Kantowitz (1992). However it is essential to establish at which points in the experimental session guessability, learnability and experienced user performance (EUP) are being measured. Jordan and O'Donnell (1992) developed a 'semi-formal' analysis with the aim of quantifying guessability, learnability and EUP. Jordan however argues that this analysis, whilst introducing a formality, is far too costly to apply in terms of experimenter time and effort. He therefore suggests a far quicker analysis using a criteria-based analysis technique.

This technique is based on the approach of Gilb (1981), and advocates setting usability specifications for an interface. In order to measure the particular components of usability criteria must be set which define the levels of performance a user must achieve before they can be said to be at a particular level of usability. These criteria are, however, vary from experiment to experiment, and therefore it is essential that a pilot study is carried out before the actual experiment in order to obtain these criteria. If an interface is being tested, rather than some experimental hypothesis, then it is more appropriate to set target criteria, which the interface must achieve if it is to be accepted.

Obviously, when measuring recall, there is not going to be a guessability stage, since users are required to learn before they are able to recall. Recognition is going to be important for all stages of usability. Discriminability is also relevant throughout the whole of the learning curve but it is more likely to be important in the learnability and EUP stages. No work has however been done to test this to date.

Data collected from the first block of experimental trials was considered to represent the data for the guessability stage. The learnability and EUP stages were identified by performing a series of pilot studies and examining the curves created by the data as performance gradually improved and then levelled off to a steady state. For almost all experiments, learning appeared to occur between blocks 2 and 6, after which users achieved a relatively steady state of experienced performance.

4.4 Conclusion

Therefore, based on the 6 point checklist adapted from Kantowitz (1992), it appears that the experimental design is valid. The following 3 chapters describe experiments 1 to 6 re-emphasising where the method differs. The experiments are presented in the order in which they were performed thus giving some idea of the development of the revised hypothesis which is discussed in Chapter 8.

5. Experiment 1

5.1 Introduction

We often assume, both from introspecting on our experience as users and as researchers, that users learn command names; and that if the interface uses icons rather than textual names, then we learn the icons. More precisely, we learn to associate commands with particular icons through recognition of their visual features and particularly their shape. On the other hand, it is widely suspected (from introspection and other sources) that experienced users have often learned the position of items (e.g. command names on menus) as well. However it is quite difficult to demonstrate this, partly because learning position does not seem greatly to improve measured reaction times. One of the few to demonstrate a clear effect is Kaptelinin (1993), who showed that after training on textual menus, randomising position had a disruptive affect on selection times, which failed to diminish over time. If however position remained consistent, and instead the letters of the textual commands were masked while retaining word length subjects performed as well as they had with the familiar textual commands.

Blankenberger and Hahn (1991) similarly remarked in an experiment designed to compare performance using textual commands, abstract icons and representational icons, that when position was fixed, subjects appeared to learn position and no significant difference in performance between conditions was found. Only when position was randomised did they find a superiority effect for the representational icons.

The results from both pieces of work suggest that icon position has an important influence on usability. However, they fail to supply answers to any of the thesis propositions suggested in Chapter 3. For example, do people learn representational shape **and** position, or do they ignore representational shape once they have learned position? Another hypothesis would be that people who are presented with representational form ignore position, while those people who do not have representational form rely on position. The achievement of equivalent performance measures for both scenarios would then be attributed to the experienced use of either of these attributes.

Experiment one is therefore an initial attempt to answer these questions. The experimental hypothesis is that subjects will not be significantly disrupted by randomising icon position when the icon set they use contains guessable, representational icons. By contrast, those subjects using an abstract icon set will sustain a detrimental effect on performance when position is randomised. In addition, if subjects using the representational icons did rely on associating the command with the shape rather than position, then manipulating the shape of the icon while keeping position constant should result in a decrease in performance. Subjects trained on the abstract set of icons should rely more on position and therefore performance should not be so significantly affected by changes in the icon form.

5.2 Method

The general paradigm applied is described in detail in Chapter 4. This section discusses exceptions to the general paradigm. All materials for this experiment are found in Appendix 1.

5.2.1 Subjects

Twenty-four subjects, with a mean age of 24 participated in the experiment. All had normal, or corrected-to-normal vision, and were experienced users of word processing packages. Four of the subjects were female, the other twenty were male.

5.2.2 Design

The design of this experiment replicated the design by Blankenberger and Hahn more faithfully than any of the other experiments. The differences in design from that described in the paradigm section are as follows:

1. Sets of 17 icons were used for each condition.
2. There were 8 blocks of trials in total; five blocks before the changeover and three after.
3. The icons were displayed in a rectangular format (replicating the formation in Blankenberger and Hahn's experiment; see the screen-shot in Appendix 1).

5.2.3 Procedure

Again there were several major differences from the general paradigm employed. The start of each trial was signalled by a tone. After 0.5 seconds one of the 17 commands was displayed textually for 1.5 seconds. The command disappeared and was replaced by the appropriate set of icons. After the subject responded, the set disappeared, and approximately 0.5 seconds later the next trial started. An erroneous response was signalled by a high tone and recorded. This tone informed the user that they had selected an incorrect icon, and to select again. The user was required to continue selecting until the appropriate response was made.

5.3 Results

Looking at the overall performances graphed for both reaction times and error rates (see Appendix 2) the main distinction appears to be between performance using representational form and performance of subjects using abstract form. Subjects using the representational form icons very quickly settle down to a steady state of performance (around the end of the first block of trials) whereas performance in the abstract form conditions remains significantly slower, more error prone and more erratic. However, at the block of trials preceding the changeover (block 5) it appears that the performance levels of subjects in the abstract conditions are beginning to converge with those of users in the representational conditions.

Figure 5.1 shows a graph of the mean reaction times between blocks 4 and 7. The interesting blocks to consider are blocks 5 and 6. Block 5 represents the level of performance users had attained through learning the consistent interface, while block 6 represents performance levels once the appropriate manipulation had occurred.

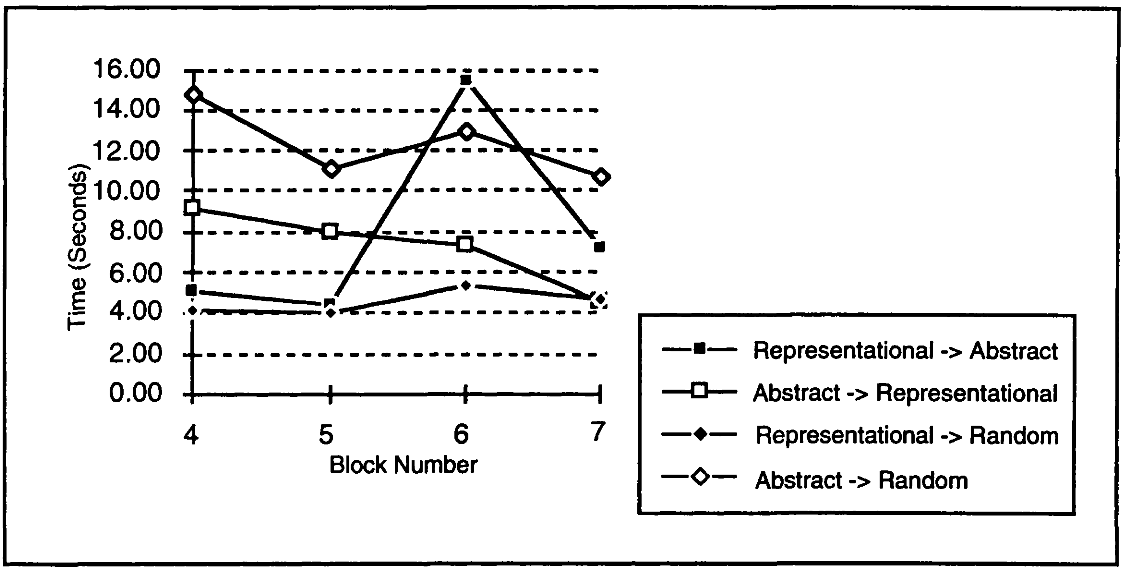


Figure 5.1: Graph of the mean reaction times for the four conditions across blocks 4-7

A two way analysis of variance (ANOVA), with repeated measures on one factor (Block) was carried out to compare performance results over blocks 5 and 6, for the representational-to-abstract and abstract-to-representational conditions, and also the representational-to-random and abstract-to-random conditions. Significant interaction effects between block and conditions were found for both comparisons (comparing the representational-to-abstract and the abstract-to-representational conditions, $p < 0.0001$; comparing the representational-to-random condition with the abstract-to-random condition, $p < 0.01$). Figure 5.1 gives a clearer description of the results, and can be described as follows:

Representational-to-abstract Condition

Between blocks 5 and 6 there is a four-fold worsening of performance as a direct result of the representational icon set being switched to the abstract set. This disruption appears to suggest that subjects relied on shape and failed to use position when recognising the appropriate icon.

Abstract-to-representational Condition

Between blocks 5 and 6 there is no significant improvement in performance caused by the introduction of representational icons ($p = 0.1081$). This is a highly significant result, however, since it suggests that subjects were not relying on form, but were using position. In this condition, one subject's spontaneous comment showed that he had not noticed that the shapes had been changed until after he had made his response (correct icon selection).

Representational-to-random Condition

Between blocks 5 and 6 the icon shape is maintained and position is randomised. The graph suggests that no significant change in performance occurred ($p < 0.08$). This is again a significant result, reflecting the idea that subjects in this condition had not come to rely on position, but must have used shape.

Abstract-to-random Condition

Again shape is maintained, but, in block 6 position is altered. The graph suggests that performance significantly worsened ($p < 0.01$), suggesting that subjects must have been relying on position in order to recognise the appropriate icon.

5.4 Discussion

Has this experiment increased our progress towards either supporting or rejecting the thesis propositions? The hypothesis that representational attributes are easier to learn than abstract ones is supported by the huge gap in the difference between performance using the representational icon form and the abstract icon form sets. However although pilot studies suggested that users' performance reached what appeared to be EUP before changeover the actual data appears to suggest otherwise. It may be that more trials are required before a reliably steady state is reached by users in the abstract conditions.

The changeover data does appear to support the idea that subjects trained on representational icons (i.e. icons whose shape was easy to guess and to learn) relied on that shape. If the shape were changed their performance was disrupted, if it were maintained then randomising position had little effect. On the other hand subjects trained on abstract icons (i.e. those whose meaning and association with function was hard to guess and to learn) relied on position to identify the icon rather than shape; if position were maintained then changing the shape had little effect, while if position were randomised their performance was disrupted.

5.5 Conclusion

In this experimental situation, at least, subjects appear to select either position or shape to associate an icon with a command but not both. Which one they select appears to depend on whether the icons' forms are easy or hard to learn.

The most pertinent question to arise from this work would appear to be whether this effect can be replicated with a far smaller set of icons, the extreme case being only two icons in the set. If it is possible to replicate with such a small number then this would suggest that the distinction between relying on form or position depending on whether the form is representational or not is not simply an artefact of the extreme difficulty in attempting to learn such a large number of abstract icon forms. It would instead suggest that the results achieved here are a reliable indicator of which attributes users would rely on when faced with similar choices in real life interfaces.

Another issue raised is whether people naturally do learn more than one attribute if given longer to experience the interface. In this experiment training was over five trials per icon only. However this experiment has already demonstrated that the attribute used to identify the appropriate icon depends on which is easier to learn, and is not a fixed property, such as form.

6. Experiments 2 to 5

6.1 Introduction

The conclusions from experiment one recommended that the scope of the effect demonstrated should be tested by employing two variations in the design to modify the number of icons within the icon set, and to vary the amount of training time before changeover. These issues are considered over the next series of experiments. As a stringent test all experiments in this chapter use sets of two icons. By doing this some degree of surrounding contextual information is available on the interface, allowing comparisons and contrasts to be made by the user. Even with just one other icon, position becomes a potential distinguishing attribute. There is also the potential for the user to decide that an icon represents a particular command on the basis of their belief that the form of the other icon suggests that it cannot be the icon. By using sets of two icons it is possible to perform a larger number of trials thus achieving more reliable data giving some insight as to what happens when users reach EUP. As a result it was possible to vary the number of trials between the four experiments, ranging from a total of 10 blocks to 40, allowing comparisons to be made between experiments. However, the interesting thread linking these experiments together is the emerging idea that what may actually be happening at changeover is perhaps far more complex than first predicted. Users are not simply selecting an attribute and ignoring all others. What really may be happening is more similar to the fifth proposition suggested in Chapter 3; that users are aware of both attributes but rely on one more heavily than another. Furthermore, which attribute they rely on may not be as predictable as initially thought. This finding is important since it suggests that designers should not assume that users will automatically rely upon the icon's form.

6.2 Experiment 2

The main question studied by experiment 2 is whether the size of the icon set may have an effect on which attribute subjects use to associate icons with commands. The set of seventeen icons used in experiment 1 was perhaps too large to learn in such a short space of time. An alternative explanation may be that if the set is so large then, irrespective of the amount of experience, subjects will use only the most distinguishing attribute, the attribute that allows the greatest contrast between icons in the set, to associate icons with commands. It is interesting to see whether users become more or less selective in the attributes they learn as the icon set decreases.

In experiment 1 users trained to use the abstract icons found the tasks arduous, often taking around an hour and a quarter to complete the experiment. This is in marked contrast to subjects using the representational icons. These subjects were able to complete the experiment easily within 20 minutes. It was for this reason that in experiment 1 only five training trials were presented (5 x 17 tasks) before the experimental trials began. As has been suggested, the main motivation for reducing the icon set to two is that it makes it possible to increase the number of trials to a level that ensures that all users have reached EUP. This design has two advantages. Firstly, it gives a clearer idea of which attribute subjects rely on (position or form or both) when there are only 2 icons. Secondly, it should also give some idea about which attributes subjects favour at EUP level. The results from experiment 1 could simply have been achieved because users were taking longer to learn the abstract icon shapes. Perhaps, had subjects been allowed to reach EUP in experiment 1 they would have learned abstract form, and as a result no significant differences would have been found between the conditions.

The hypothesis for experiment 2 (a replication of experiment 1 but with only two icons) is that given an easily discriminable and consistent attribute by which users can recognise the appropriate icon (form in the representational conditions, and position in the abstract conditions), subjects will ignore all other attributes. Consequently, when the learned attribute is made inconsistent or removed, performance levels would be predicted to deteriorate. Manipulation of the other attribute should have no significant effects on performance. This result would support propositions 2 and 3 in Chapter 3. It is predicted that this hypothesis supported in experiment 1 with sets of 17 icons will also be supported when there are only 2 icons in a set.

6.2.1 Method

The method for experiment 2 is essentially the paradigm described in Chapter 4.

6.2.1.1 Subjects

A total of 16 subjects were used for experiment 2. All subjects were either postgraduates or fourth year undergraduates in the Computing Science Department, Glasgow University. They were all familiar with both graphical interfaces and word processing packages. The mean age was 24. All had either normal, or corrected-to-normal vision.

6.2.1.2 Design

Sets of two icons were used in this experiment. Combinations of icons were randomised between subjects in order to prevent any extraneous factors affecting results.

An experimental session consisted of forty blocks of trials. Icon manipulation (changeover) occurred at block 21.

6.2.2 Results

Figure 6.1 shows a graph of the mean reaction times per trial block (1 block equals 2 trials). It suggests that from block 5 of the training trials onwards subjects in all conditions settle into an optimal level of performance. There appears to be no significant difference between the conditions. The changeover occurs at block 21, and is reflected in the spike on the graph. Recovery however is very rapid, and by block 23 the conditions are again at equivalent performance levels.

An analysis of variance (ANOVA) was performed comparing times of the four conditions of subjects over blocks 20, 21, and 22 (see Appendix 3). As can be seen there are significant main effects for both conditions ($p < 0.05$), and blocks of trials ($p< 0.005$). No significant interaction effect was supported.

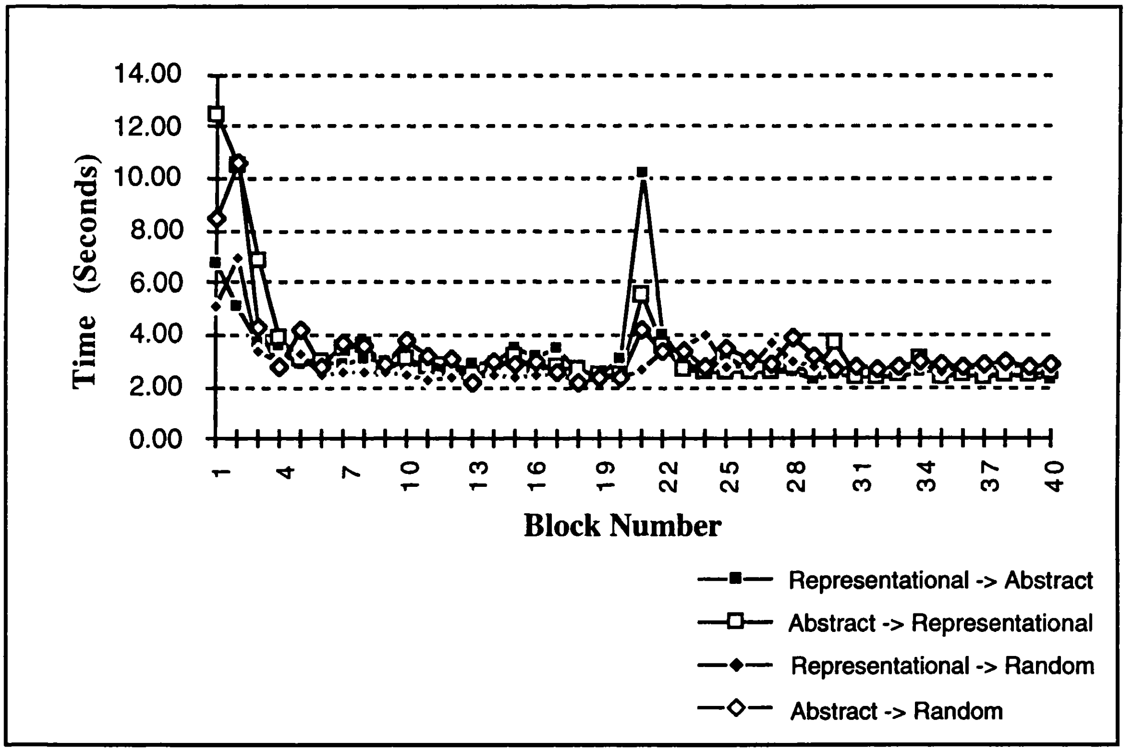


Figure 6.1: Mean times to complete each trial (across blocks 1 to 40)

Looking at the simple effects (Appendix 3), the only significant difference between the conditions appears to occur at block 21 (when the changeover occurs). This reflects the findings already described by the graph. Before the changeover there is very little variation between the conditions, and after the changeover all conditions rapidly recover to their original performance levels. Results from a Tukey (hsd) test suggest that there is a significant difference in times at the 0.05 level between conditions 1 (representational-to-abstract) and condition 3 (representational-to-random) after changeover. Since both conditions were trained on the representational icon set the difference must be attributed to the changeover. For condition 1 the change is a change of form. Since these users were predicted to be relying on form then their performance should deteriorate most dramatically. However, icon form remains the same for the representational-to-random condition and only position alters. Since subjects in condition 3 are expected to rely on form rather than position then there should be no significant difference in performance, accounting for the significant difference between the 2 conditions.

The representational-to-abstract condition also shows a significant difference in times between blocks. This again is reflected by the huge spike displayed on the graph at block number 21.

6.2.3 Discussion

Experiment 2 therefore suggests that there may be no difference between how users learn icons when there are 17 in a set compared with when there are only 2. The results in experiment 1 were essentially replicated by experiment 2, even though users had twenty training trials before the experimental trials began. Looking at the graph in Figure 6.1 it is possible to see a learning curve develop during the training trials. From block 5 onwards all 4 conditions appear to settle down at an equal and constant level of performance, thus suggesting that, for the 2 icon set at least, EUP does occur quickly. It was interesting to note however that in experiment 1, users were similarly converging to a settled pace of performance by block 5. By trial 20 in experiment 2 performance levels are equivalent in all conditions suggesting that users have become particularly experienced in whatever attribute they use to recognise and select the appropriate icon.

6.3 Experiment 3

It is interesting to consider whether the disruptive effects achieved in experiment 2 are greater or lesser when the changeover occurs after the fifth block of trials. If all users learned, or at least attempted to learn, form first and position later then perhaps the effects shown in the previous experiments wouldn't occur. In other words, when do users begin to learn position? Performance measures do not make this clear. Perhaps by causing the disruption earlier, and comparing results with results from experiment 2, we can start to get some idea.

The hypothesis for experiment 3 is therefore that those subjects exposed to representational icon forms learn form first, and those with abstract icon forms learn position. If those exposed to the abstract icons learn form, then at changeover for the abstract-to-representational condition performance should be expected to become faster as a response to being presented with easier icon forms to recognise. For the abstract-to-random condition, however, there should be no change. Similarly, for those trained to use the representational icons the representational-to-abstract condition should show a significant decrease in performance only if users have not learned position. The representational-to-random condition should maintain similar performance levels throughout.

If position has not been learned by the changeover point, then it would be expected that the conditions in which form is altered should show significant differences in performance, whereas conditions that randomise position after changeover should not show any significant performance differences at all.

6.3.1 Method

The paradigm for experiment 3 is identical to that used in experiment 2 (using sets of 2 icons). The only difference is that there are only 5 blocks of trial tasks and 5 blocks of experimental trials. The subjects used for experiment 3 were also similar to those used in experiment 2. Again the mean age was 24. All 16 subjects were either postgraduates, or fourth year undergraduates from the Computing Science Department, and all had either normal or corrected-to-normal vision.

6.3.2 Results

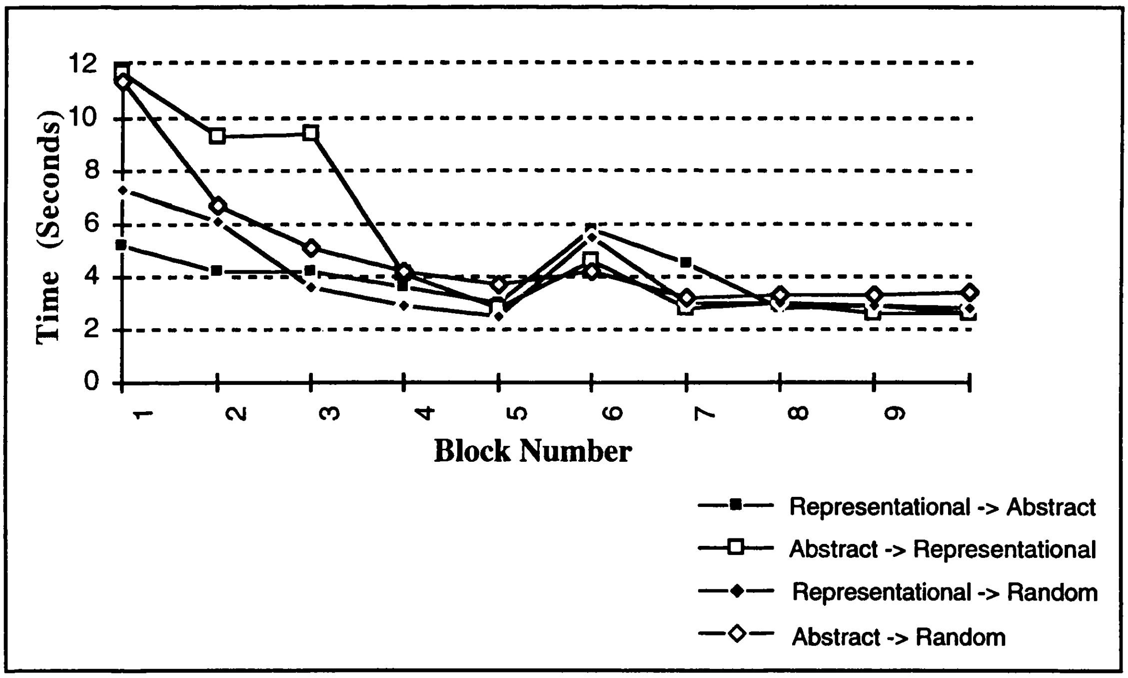


Figure 6.2: Mean times per trial over across 10 blocks of experiment 3

Figure 6.2 shows the mean reaction times for experiment 3. Again a gradual learning curve can be seen for all conditions levelling off at a common optimal level of performance by block 5. Similar to experiment 2, there is a spike in performance as direct reaction to the changeover at block 6. Unlike experiment 2, however, the spike is much smaller, and there is also less difference between the conditions. This is reflected in the ANOVA data for trial blocks 5, 6 and 7 (see Appendix 4).

Only variance between blocks shows a significant main effect ($p>0.001$), suggesting that the changeover causes enough disruption, but that the disruption is of equal proportion for all conditions. A Tukey test comparing blocks shows a significant difference between blocks 5 and 6, and 6 and 7.

Both differences are significant at the 0.01 level. This again suggests a rapid recovery rate after the changeover.

Looking at the table of simple effects in Appendix 4, it appears that there are significant differences between blocks for both the representational-to-abstract condition, and the representational-to-random condition. A significant difference between blocks for the representational-to-random condition would suggest that users are relying on position. This contradicts any previous argument. However the standard deviations (Appendix 4) appear far greater than those for the data in experiment 2 so perhaps these results are not as reliable.

6.3.3 Discussion

Are users attempting to learn form before learning position? It is impossible to tell from these results since the disruption is of particularly equal degrees in both abstract conditions irrespective of whether it is form that has been manipulated or position. An alternative angle is to consider whether subjects in the representational conditions learn position before the changeover. This would give some indication of whether attributes were relied on singularly or whether users are aware of more than one.

One possible measure to estimate whether or not position is being learned is to look at how often subjects are moving towards the icon before the screen displaying the icons has appeared. Figure 6.3 shows a table of the total numbers of correct and incorrect moves for all 16 subjects in both experiments (split into conditions). There were four subjects in each condition and a total of 10 trials (5 blocks) in both the training and experimental sections of the experiment. For experiment 2 there were 40 trials (20 blocks) for each subject.

		Experiment 2		Experiment 3	
		Training	Experiment	Training	Experiment
Condition 1 Rep -> Abs	Correct	3	20	3	2
	Wrong	0	1	0	1
Condition 2 Abs -> Rep	Correct	41	62	3	1
	Wrong	1	0	0	0
Condition 3 Rep -> Rand	Correct	65	14	7	4
	Wrong	0	15	2	2
Condition 4 Abs -> Rand	Correct	33	14	3	6
	Wrong	0	10	0	4

Figure 6.3: Table of numbers of pre-emptive correct and incorrect cursor movements (Totals for experiment 2 are out of 160. Totals for experiment 3 are out of a possible 40)

The data for the table were collected by simply working out the co-ordinates of the central button and from that inferring from the x,y co-ordinates collected as the command card closed, whether the user had moved the cursor to the left or right of the button before the interface had appeared. Even with this rather crude method of analysis it is apparent that subjects using icons with representational form are making pre-emptive moves. Even in experiment 3, with the short series of training trials, evidence of position learning in all conditions is apparent. Users were recorded making pre-emptive moves around blocks 3 and 4. This was not dependent on condition. This contradicts the argument that people learn only one attribute. These results seem to suggest that users do learn form and then once they have learned position swap to using this attribute instead. However, if users do gradually shift strategy to rely on position rather than representational form, then how does this account for the large disruption effects found consistently in the representational-to-abstract condition, which are not so apparent in other conditions?

6.4 Experiment 4

After the results gained in experiment 3 it was suggested that perhaps the large spike found at changeover in the representational-to-abstract condition for experiment 2 was either caused by:

a. Disruption due to learning, as the attribute users' had been relying on to recognise the associated icon for the command had been taken away and they were now required to learn a new attribute. This would suggest that a shift in strategy does not occur for this condition. Perhaps the switch only occurs in abstract form conditions?

b. Surprise at the change in the interface. Subjects had already learned position and this was therefore only temporary disruption caused by surprise rather than re-learning.

c. An experimental error. Some other external variable had been the cause (i.e. user differences).

To ensure that the cause of the disruption was not case c, experiment 4 was a re-run of experiment 2 using a larger subject sample to see if results could be replicated.

6.4.1 Method

The paradigm is identical to that of experiment 2. Experimental sessions contained 40 blocks of trials, with the changeover occurring at block 21. A total of 28 subjects were tested. Twenty-one subjects were male and seven were female. The mean age was 27. All subjects had normal or corrected-to-normal vision.

6.4.2 Results

The large spike apparent for the representational-to-abstract condition in experiment 2 is replicated in experiment 4 and is apparent on the graph of the reaction time data in Figure 6.4. Unlike experiment 2 however, the second biggest disruption appears to be for the abstract-to-random condition. The other two conditions show equivalent amounts of disruption at changeover. This result is more similar to the original result extracted from experiment 1.

ANOVA data for blocks 20, 21 and 22 (see Appendix 5) shows significant main effects for both block ($p < 0.0001$) and condition ($p < 0.05$). As for experiment 2, there was no significant interaction effect.

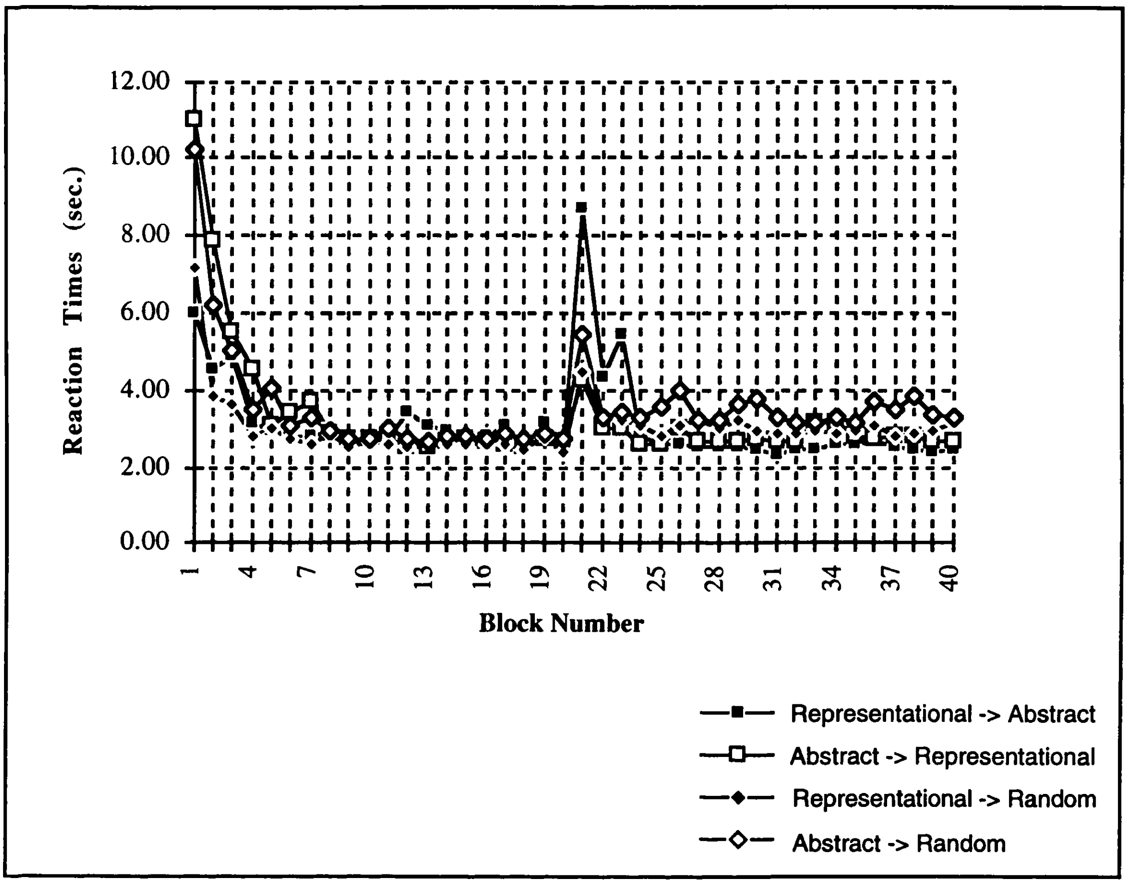


Figure 6.4: Mean reaction times per block for each condition in experiment 4

Looking at the simple effects analysis there is again a significant difference between the conditions at block 21 (when changeover occurs) and also between blocks 19, 20 and 21 for the representational-to-abstract condition ($p < 0.0001$). This latter result is expected as it reflects the large spike on the graph.

6.4.3 Discussion

It is therefore possible to say that the chances of the large disruption being due to experimental or subject error are minimal. This suggests that the consistent disruption at changeover must either be due to surprise or relearning. In a two icon set it is impossible to differentiate between measurements that may indicate surprise and measurements that would indicate learning. Even if users had not previously learned position before the changeover, the fact that there are only two icons makes learning very easy, therefore possibly resulting in a recovery to previous performance levels in the equivalent time it may take to recover from the effects of surprise. The only way to get a firmer grasp on what is actually happening, and ultimately understand whether people have or have not learned both attributes before changeover, is to repeat the experiment using a larger number of icons. If learning does occur at changeover then it is going to take longer when there are more icon positions/forms to learn.

6.5 Experiment 5

Previous experiments have presented subjects in the abstract-to-representational condition with representational icons at the changeover. Results for this condition have, until now, shown that subjects performance is not significantly disrupted at changeover. This was assumed to be because users were relying on position rather than on icon form to associate the icon with the command. However, it could also have been as validly argued that subjects in this condition actually benefited from the changeover, by being presented with a set of guessable attributes. Perhaps users actually did rely on abstract icon form, but at changeover the effort to learn the representational icon form was so small that disruption caused by the removal of the abstract form was obscured. Experiment 5 therefore re-ran just this condition, and compared it with results gained for the other three conditions in experiment 4. The condition varied from previous conditions by presenting users with a set of different abstract forms at changeover instead of representational forms.

6.5.1 Method

The paradigm is identical to experiment 4, except only one of the four conditions is being run; the abstract-to-representational condition. Furthermore, this has been altered so that at changeover the icon form is replaced by another abstract icon form. Thus, on this occasion, the condition will be renamed abstract-to-abstract to prevent any confusion.

Seven subjects were used in this experiment. All subjects were male postgraduates or undergraduates at the University of Glasgow with a mean age of 24.

6.5.2 Results

A similar result was discovered for the replaced abstract-to-abstract condition (see Figure 6.5) as was found for the original abstract-to-representational condition in experiment 5. Using an ANOVA, significant differences were found between both blocks and conditions, however no interaction effects were discovered. This was expected since the data is identical to the data analysed in experiment 4. The simple effects analysis suggested that altering the abstract-to-representational condition resulted in making the abstract-to-random condition significantly different between conditions ($p < 0.05$). Perhaps this is because the new condition resulted in even less of a disruption at changeover than the previous manipulation had.

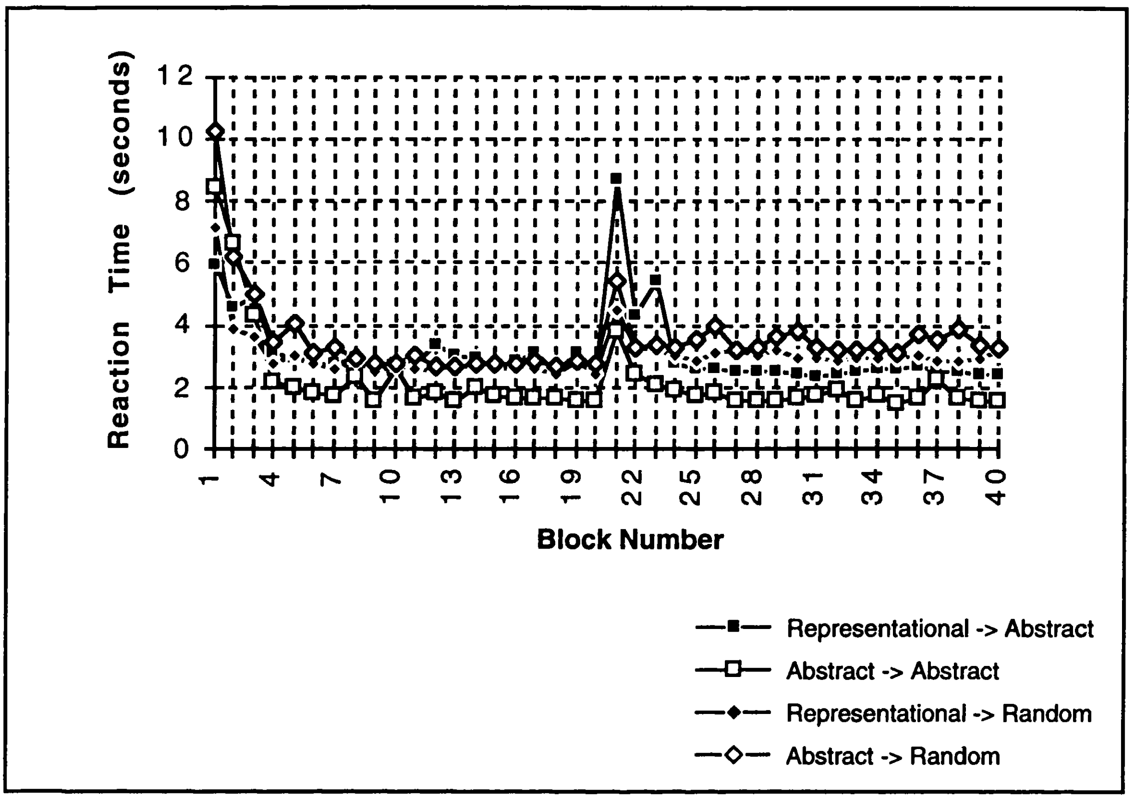


Figure 6.5: Mean reaction times for 3 conditions in experiment 4 and 1 condition from experiment 5.

6.5.3 Discussion

The conclusion suggested by this experimental result is that subjects are relying on position when presented with abstract icon form and therefore were not gaining any benefit from the addition of representational icons at changeover. This would suggest that subjects were already using a suitable attribute to associate icons with commands - position. Results could, however, be dependent on set size. Perhaps if there were more icons then there would be a sharp increase in performance by the introduction of the representational icon. This was indeed suggested to be the case in experiment 1 where performance in the abstract-to-representational condition quickly improved after the changeover.

Comparing the second condition in figures 6.4 (abstract-to-representational) and 6.5 (abstract-to-abstract) it is clear that after the changeover subjects in the abstract-to-abstract condition appear to perform better. Could this be accounted for by suggesting that when presented with representational form subjects feel almost obliged to rely on it, even though, as was shown in the abstract-to-abstract condition, they have an adequate attribute which they have already learned to rely upon. It would appear therefore that people are conditioned to believe that a representational form must mean something and therefore is an important attribute to learn.

However, one strange anomaly in the data is that subjects in the abstract-to-abstract condition (shown in figure 6.5) appear to perform better before changeover than subjects in the abstract-to-representational condition. The reason for this cannot be accounted for since subjects in both conditions were presented with identical icons, they used the same machines, and came from the same user population. The time of day in which subjects were tested was randomised for both conditions and all subjects were presented with the same instructions. The only apparent difference is that subjects were tested in different months. Such an anomaly was not found in any other experiment, with pre-changeover performances being similar for both representational conditions and abstract icon conditions. Does this place question on the experiment's findings? The important finding from this experiment is that subjects using icons with abstract form rely, to a greater degree, on the position of the icon. The pre-changeover data suggests that for the abstract-to-abstract condition, the users appear to be relying on position sooner. This would perhaps account for the lesser disruption, and faster recovery at changeover than was found for the abstract-to-representational condition. However, the important to point to note, irrespective of general performance speed, is that the fast recovery after changeover is not due to the appearance of the representational form, but because subjects appear to have an attribute they have previously learned to rely upon.

6.6 Conclusion

From these experiments five points can be made:

1. The results described in experiment 1 can apply equally to sets of two icons.

2. Results, however, are more complex than originally assumed. It appears that pre-emptive moves are made in all conditions, suggesting that icon position has been learned. The analysis of pre-emptive moves in these experiments was rather crude. There is no indication from the data of whether the move was made with the intention of moving towards the appropriate icon, or simply as a random movement made by the user. Similarly, the user may not make any pre-emptive moves, but if asked to locate an icon would be able to do so. The number of pre-emptive moves made is perhaps dependent on whether the user perceives speed or accuracy to be more important when performing the task well. However, the fact that there are equal proportions of pre-emptive moves made in each condition does tend to suggest that learning is occurring irrespective of the type of icon form.

3. If it is the case that people learn position then why do we get significant disruption effects that appear to fit our hypothesis? Can it be attributed to some degree of learning after all, or is it caused by something else such as surprise or experimental error? Experiment 4 suggests that the massive disruption for the representational-to-abstract condition was not due to experimental error.

4. It was originally suggested that people were not relying on form at changeover in the abstract-to-representational condition. Experiment 5 was designed to test this hypothesis it appears to have been supported. Whether or not users had initially attempted to use form and then switched to using position is not evident from the data. Many users did, however, claim in their questionnaire responses to have employed such a strategy.

5. Although subjects in the abstract-to-representational condition appeared to be relying upon position, by comparison with the results in experiment 5 it appears that the introduction of representational form icons may actually have hindered performance at changeover. It was suggested that this may be because people have come to expect that representational form must mean something and therefore is a useful attribute to learn. This refutes the idea that users economise on the amount of information that they learn in order to identify the appropriate icon. This result would instead support the idea that users learn a multitude of attributes, although they may only rely upon one to recognise the icon.

To gain a clearer idea of whether or not the disruption at changeover is due to surprise or learning it was decided that the experiment should be re-run using larger sets of icons. If the disruption were due to the effect of learning rather than surprise, then the presence of additional icons should have the effect of increasing learning times. If the disruption were due to surprise then the performance measures should return quickly to a steady level irrespective of the number of icons present.

Furthermore, larger icon sets should allow a more rigorous investigation of whether people are making purposeful pre-emptive moves. With more icons present on the interface a pre-emptive move can no longer be signalled by a simple move anywhere to the right or left of the centre button, but should be more direction specific.

7. Experiment 6

7.1 Introduction

The conclusion of the previous chapter was that it was still not clear whether disruption at changeover could be the result of surprise rather than learning. This is reinforced by the evidence from pre-emptive move data suggesting that subjects are relying on position, even in conditions where representational icon form is present.

This idea is also supported by the relatively quick recovery rates after changeover. The disruption caused in all conditions is far smaller at changeover than in the first few trials of the training section of the experiment when users are initially learning to associate icon attributes with commands. However with only two icons in a set it is possible that the speed difference in learning at the start of the session and at changeover are purely due to an increase in performance practice.

Experiment 6 uses larger icon sets (16 icons per set) in an attempt to increase the time needed to learn non-guessable attributes. The hypothesis for this experiment is that if disruption is caused by the user learning a new attribute, then performance should take a long time to recover after changeover. If the disruption is entirely due to surprise at the change in the interface then the disruption should last one or two trials at maximum.

A second aim of this experiment is to analyse the extent to which icon position is used in all four conditions before changeover. Pre-emptive moves were captured in the form of two sets of x,y co-ordinate positions representing the cursor position before the command card closed and the cursor position as the interface card opened. A HyperCard stack was developed to plot out these movements allowing a visual representation of the mouse movements to be made apparent (see Appendix 9). In correspondence with the initial hypothesis made, it is predicted that only those conditions using abstract form should show evidence of icon position being learned. The null hypothesis, supported by the previous analyses of pre-emptive moves is that position will be learned irrespective of the type of icon form present.

7.2 Method

The method is again similar to that described in Chapter 4 for a 16 icon 40 block design. However, several important changes were made to the design.

7.2.1 Apparatus

The experimental sessions were performed using a Macintosh Quadra (system 7.1.2). The machine change was made since the Quadra offered faster performance and larger memory for data storage.

7.2.2 Design

Several alterations were made to the design of the experiment. The icon set size for this experiment was 16. If 17 icons had been used then a more direct comparison with experiment 1 would have been possible. However, the number of icons used was determined by their layout on the interface. In experiment 1 the icons formed a rectangular shape matching the window outline (see Appendix 1). The criticism of this layout was that not all icons would be equi-distant to the central button and therefore performance times may be affected by the differences in the distance the mouse had to travel. For experiment 6 icons were placed in an equi-distant circle round the position of the centre button (see Figure 4.4). Circle size was limited by the size of the window, and thus only 16 icons were able to fit within the biggest possible circle.

Another additional feature of the design of experiment 6 was that a second set of mouse co-ordinates was captured for every trial. Co-ordinates of the mouse position as the interface card opened combined with the co-ordinates collected as the command card shut allowed a visual representation of the mouse movement to be created. This representation could then be inspected for evidence of intentional pre-emptive moves.

Utilising the recommendations made as a conclusion to experiment 5, the first two conditions (representational-to abstract and abstract-to-representational) were altered. Now at changeover subjects were not presented with a new set of icon forms. Instead icon form was completely removed and the icon frame remained. These two conditions were therefore renamed representational-to-blank and abstract-to-blank.

Forty blocks of trials were carried out by each subject (40 X 17 trials). Changeover occurred at block 21.

Finally, one potential confounding variable to consider when interpreting the results for experiment 6 may be tedium. Users in this experiment are performing rather basic tasks for a far greater number of trials than have been performed in any previous experiment.

7.2.3 Subjects

Sixteen subjects were used (four per condition). Thirteen subjects were male and three were female. The mean age was 26.

7.3 Results

Statistical results for this experiment can be found in Appendix 7. A more in-depth analysis of the data was performed due to the unexpected results found. An analysis of the reaction time data suggested that a significant disruption did occur (supported by significant between condition variation; $p < 0.001$). However this disruption did not occur as predicted since all conditions were equally disrupted at the changeover. This is indicated by significant differences between blocks ($p < 0.001$) for all four conditions.

Analysis of what can be described as 'pure' reaction times was performed. This reaction time measure is taken from the time users click on the centre button until the first icon is selected. It therefore attempts to look at user confidence in their selection of a command, irrespective of whether the selection is correct or not. Results from this data contradicts the original hypothesis. Analysis of variance suggests that main effects for both block and condition are again significant, however the smallest amount of disruption is apparent in the representational-to-blank condition; the opposite result to that found in every other experiment performed previously.

Analysis of the error data however appeared to represent the pattern of results achieved throughout this thesis; the biggest disruption appearing at the changeover for condition 1 (called representational-to-blank in this experiment). The established pattern was not so evident for other conditions.

7.4 Discussion

An examination of the graphs of the data may make it clearer to postulate what might have been happening and why.

7.4.1 Looking at the graph of the mean RTs over 40 blocks of trials

Reaction time for one trial is recorded from the time the user clicks on the centre button until the interface card closes. There are both benefits and drawbacks to this as a reliable measure of usability.

The benefit is that it allows the effects of error to be recorded into the time taken. In a natural scenario, without the penalty, time would continue until the user managed to select the correct icon. Thus by including the penalty time (or perhaps for this experiment it should be called tutoring time) into the time data we should get a better idea of the degree of learning required to select accurately the appropriate icon.

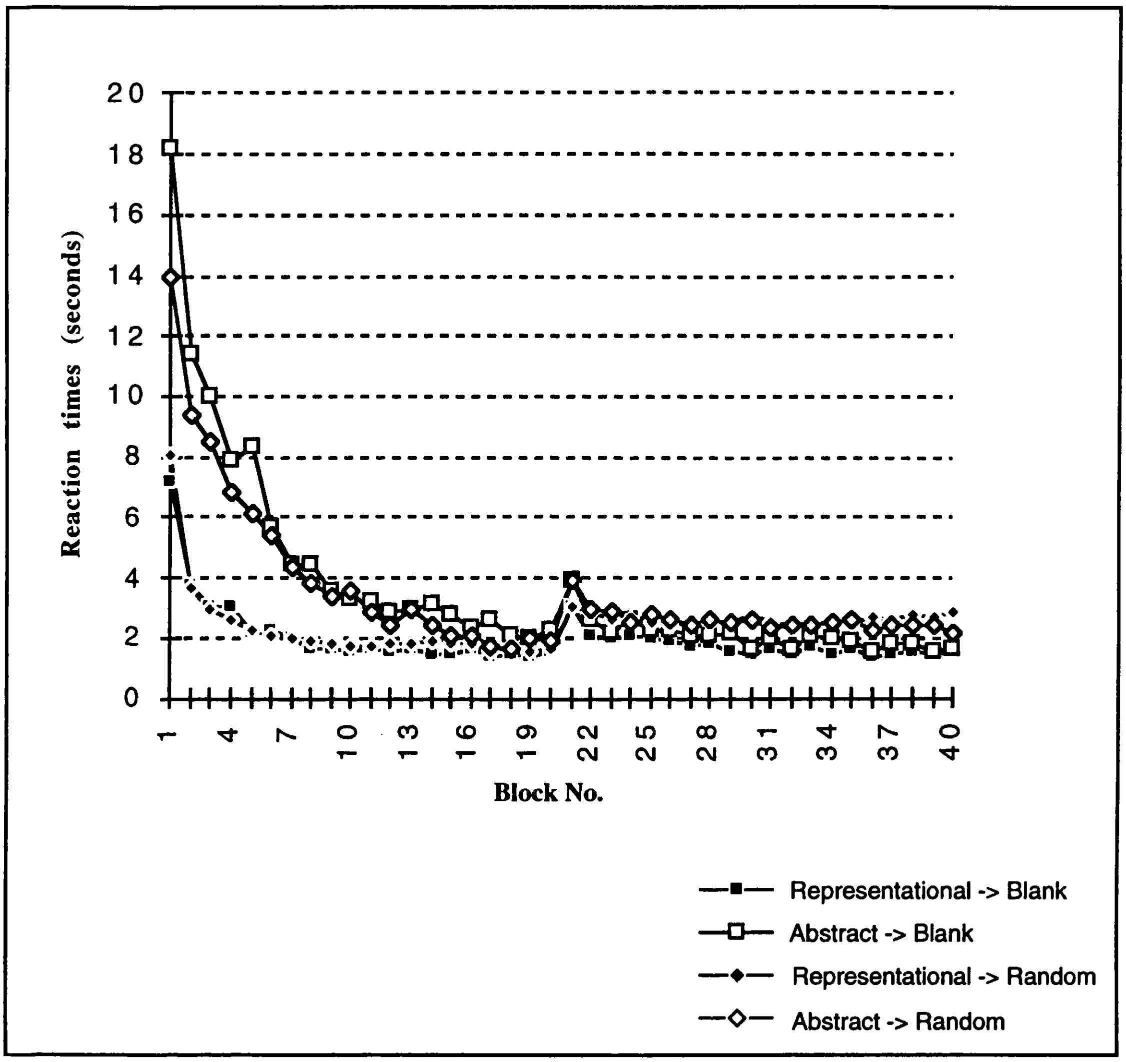


Figure 7.1: Graph of the mean reaction times across blocks 1 to 40

The drawback of this measure is that it does not distinguish between occasional slips and genuine "don't knows". Subjects who know which icon is the correct one but, due to a momentary lapse of attention, select the wrong icon are penalised identically to people unable to make the correct response. In experiment 1 the former type of person would have immediately recognised his mistake and selected the correct icon. The latter person performing experiment 1 would most likely have spent far longer and incurred far more errors before selecting the right icon. This distinction is sacrificed in order to speed up the number of trials the subject is able to get through in the experimental session. A possible solution would have been to alter the penalty so that the dialogue box appears after the user has made two incorrect selections, thus allowing the person making the slip a second chance to correct it without incurring the penalty.

7.4.1.1 Before changeover:

The graph shows that subjects presented with the abstract icons take longer to learn (are four times slower) than the subjects in the conditions using representational icons.

The representational-to-blank and representational-to-random conditions follow an almost identical pattern to each other, levelling off by block number 8 to a steady level of performance. This suggests that abstract icons are harder to learn than representational icons. It also suggests that learning position information from 16 icons is not easy to do either. It does not tell us how much people are relying on position as opposed to form.

7.4.1.2 At changeover:

- The abstract-to-random condition and the abstract-to-blank conditions appear to reach equivalent performance levels. The abstract-to-random condition however has a slightly bigger disruption to attain this level.

- The representational-to-blank and representational-to-random conditions also both reach the same performance level as a result of the disruption.

This is contrary to what was expected. In previous experiments it was suggested that the representational-to-blank and abstract-to-random conditions suffer the biggest disruptions, whereas the other two conditions were hardly affected. Also, in previous experiments, disruption did not last long with a return to previous performance levels within two blocks of the changeover. For two icon sets the recovery speed was attributed to the ease with which people could learn a new attribute with so few icons to compare. However for sixteen icons it was thought that if people were learning a new attribute after the changeover, this learning curve would be far more gradual and therefore obvious from the experimental data. Such a learning curve is not evident in the graph shown n figure 7.1, thus failing to support the idea that disruption is as a result of learning.

7.4.1.3 After changeover:

- The representational-to-blank condition quickly recovers from the disruption, however it does not reach the same performance level as it had previously attained. This would suggest that although relying on position, the icon form must have been instrumental in allowing subjects to perform so quickly.

- Subjects in the abstract-to-blank condition recover from the disruption by block 22 and eventually are performing at the same level as subjects in the representational-to-blank condition. This result could be because, after changeover, users in both conditions are using identical interfaces.

- The subjects in the representational-to-random condition never recover to the same performance level as was achieved before the changeover. By block 27 subjects are actually starting to show a decrease in performance, becoming progressively slower as they reach the end of the experiment.

- The abstract-to-random condition recovers from the disruption, but as expected does not reach the same level of performance as attained before the changeover. The times gained by subjects in this condition match those for subjects in the representational-to-random condition, until block 27 where performance of subjects in the latter condition deteriorates.

- There is no evidence of tedium confounding results, since reaction times for all conditions level off at an optimal level and remain there until the end of the experiment. If tedium had effected results there should have been an increase in reaction times as user attention wandered from the task.

7.4.2 Looking at the graph of 'pure' reaction times over 40 blocks

Pure reaction time is the measure of time taken between the user clicking on the centre button until the first icon is selected, irrespective of whether the icon is correct or not. The measure is therefore intended to give some indication of how confident a user is that he is selecting the appropriate icon. As the user learns, confidence should grow and therefore the reaction time measurements should speed up. A look at the error rates should support this. If users' performance increases simply because they do not care any more then there should also be a significant increase in the error rate. This is not, however, the case.

7.4.2.1 Before changeover:

The graph shows that subjects presented with the abstract icons take longer to learn (are slower) than the subjects in the conditions using representational icons. Results are similar to those discussed in the previous section.

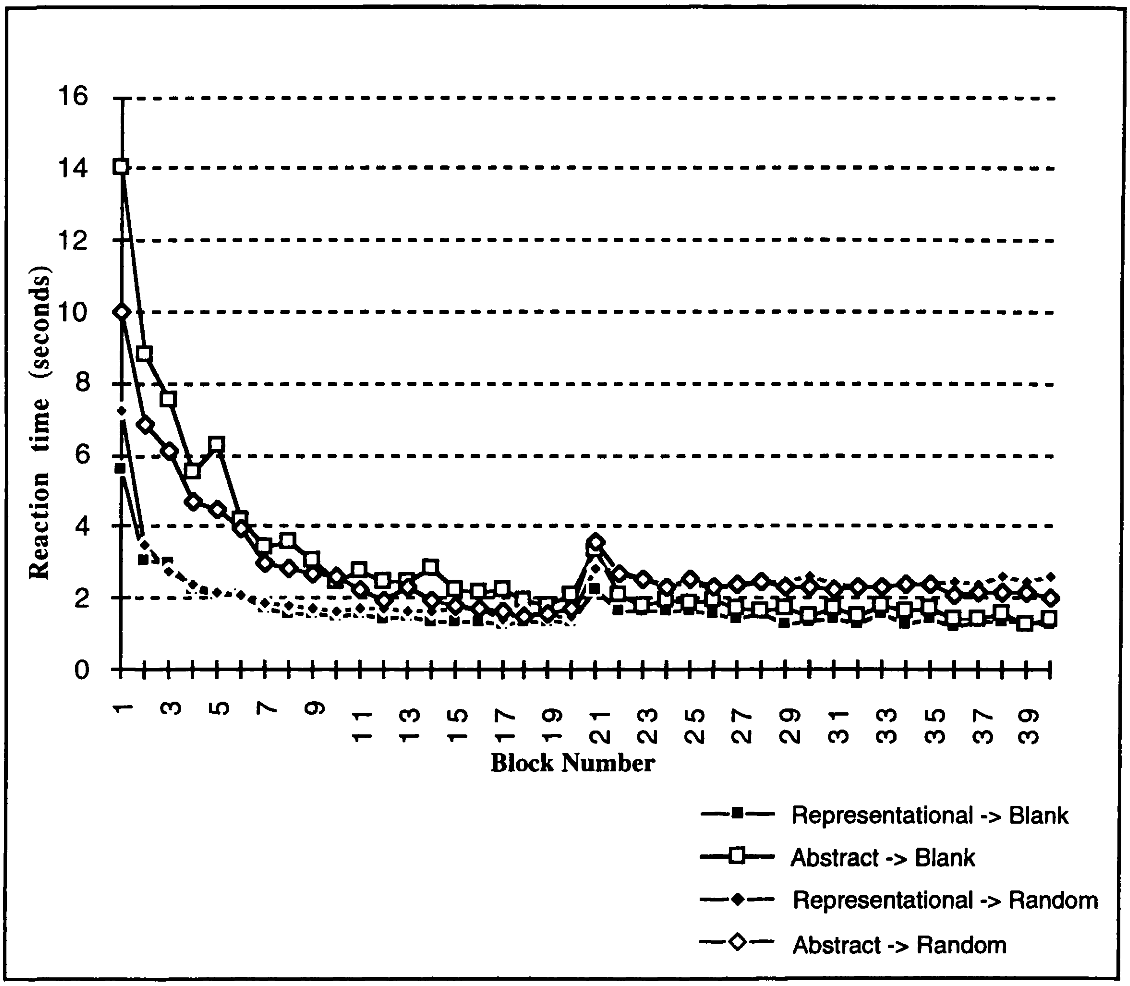


Figure 7.2: Graph of mean 'pure' reaction times across 40 blocks of trials

7.4.2.2 At changeover

- The abstract-to-random condition has the greatest disruption. This is good since we want to show that removing the position has a negative influence here. The original hypothesis would also expect the representational-to-random condition to be less disrupted in comparison since users in that condition would have been expected to be relying on representational form rather than position.

- The disruption in the abstract-to-blank condition is as dramatic as the disruption for the abstract-to-random condition. If we were trying to support the original hypothesis then this would not be a good result. This result instead suggests that subjects were relying upon the abstract form of the icon and therefore suffered as a consequence of its removal.

- The most contradictory result however is that the representational-to-blank condition maintains a comparatively faster performance with the least disruption.

It is possible that the surprising disruption found for the abstract-to-blank condition may be due to the introduction of blank icons at changeover. Several subjects in both this condition and the equivalent representational condition presumed that the blank icon meant that the experiment must either be over or that something was wrong. Therefore it could be argued that performance would be affected as a result of this confusion. However since subjects in the representational-to-blank condition encountered a similar confusion this could not be considered an appropriate explanation.

7.4.2.3 After changeover:

• The representational-to-blank condition takes until block 29 to recover to a level of performance equivalent to that achieved prior to changeover. This would suggest that subjects in this condition are not merely surprised by the change, but are taking some time to adjust; perhaps as a result of having to learn something they had not had to learn before changeover. However, the fact that the disruption was not so dramatic suggests that subjects must have, to some extent, already learned position. However, from this data it is impossible to predict whether subjects had previously learned some positions and relied on form for others, or whether they had been relying on form for all icons and positions learned were just incidental.

• Subjects in the abstract-to-blank condition return to an equivalent level of performance by block 22. This may suggest that the large disruption found in block 21 was a result of surprise rather due to some lack of knowledge This would be a favourable result, since even if users in conditions containing representational icons did use form to some extent, we would hope to predict that users of abstract icons would have been relying on position more. This is an important finding since it suggests that over time users are likely to rely on more than form, especially when icon form is particularly abstract. This issue is discussed in much greater depth in Chapter 8.

• As was suggested by the standard reaction time results, the performance of subjects in the representational-to-random condition actually deteriorates near the end of the experimental session. The reason for this decrease in performance can perhaps be attributed to fatigue. Subjects in this condition tended to find the experiment particularly tedious. The icons were easy to learn and therefore the majority of subjects had mastered the interface by the fifth block of trials. Randomising position did not appear to make the task significantly more taxing, therefore by the thirtieth block of trials subjects were extremely bored and often asked when the experiment would end.

• Again, users in the abstract-to-random condition never return to the level of performance achieved before the changeover. It is significant that neither of the conditions where position is randomised ever return to similar performance levels. By learning icon position users can gain a performance advantage by making pre-emptive moves, before the interface appears, towards the appropriate icon.

Once the interface is randomised pre-emptive moves may actually be a performance disadvantage. By suggesting that users in the representational-to-random condition may not be able to perform as quickly after changeover because pre-emptive moves are no longer a performance advantage suggests that subjects must have been using position prior to the changeover.

7.4.3 Looking at the graph of average errors over 40 blocks

7.4.3.1 Before changeover:

The significantly higher error rates illustrated for both conditions containing the abstract icon sets support the undisputed hypothesis that abstract attributes are harder to learn than representational ones. As was found in other assessments of this data, the conditions with representational icon forms have corresponding error rates, as do the two conditions with abstract icon forms. If subjects in each condition are equally matched then this result should be expected since, until changeover, each pair are using identical interfaces.

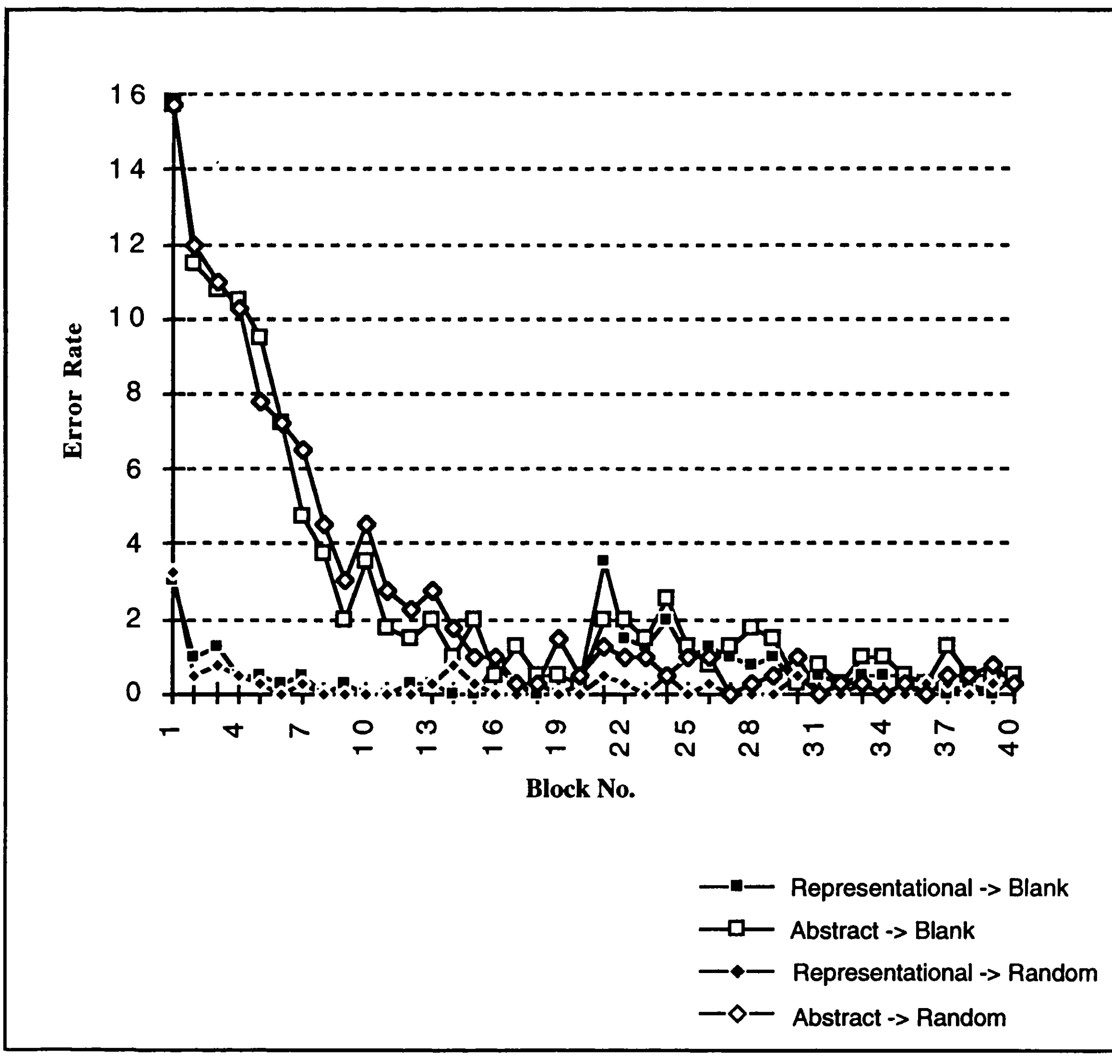


Figure 7.3: Graph of mean errors across 40 blocks of trials

7.4.3.2 At changeover:

The representational-to-blank condition appears to have the biggest disruption of all conditions, going from 0 errors in block 20 to an average of just over 3 in block 21. This is more consistent with the results gained in previous experiments. However, an average of 3 errors out of a possible 16 is still relatively low. If users had not learned position then it would have been expected that the removal of any form to identify the icons would have had an even more exaggerated effect on performance. This however is most certainly not the case.

7.4.3.3 After changeover:

- Subjects in the representational-to-blank condition do not appear to return to an optimal level of performance immediately. In fact, it may be possible to suggest some evidence for a gradual learning curve, flattening out by block 30. This curve is, however, insignificant when compared with the learning curves achieved at the start of the experimental sessions by subjects using the abstract icon sets. Therefore, perhaps, subjects in the representational condition were not learning position from scratch, but had learnt it to some extent prior to the changeover.

- Performance in the abstract-to-blank condition also appears to remain disrupted for some time. However, prior to the disruption, performance had at no point reached zero errors, therefore it could be argued that the disruption was a result of subjects still trying to learn to associate some icon attribute with a command. Removing an attribute must have been particularly disruptive if the subject had not yet established which attribute to use. This raises the issue that perhaps trying to learn the position of 16 icons is actually quite a complex task, and therefore it may actually be easier to learn the form of the icon (no matter how abstract).

- The representational-to random condition remains at zero errors suggesting that subjects really aren't disrupted by position being randomised. This supports the initial hypothesis.

- The abstract-to-random condition shows a smallish disruption which has stabilised by block 25.

- Again, there is no evidence of tedium confounding results. With the exception of the abstract-to-random condition, error rates appear to reach an optimal level and remain there.

7.5 Pre-emptive moves

The graph in Figure 7.4 is extremely enlightening. If pre-emptive moves are indeed taken as evidence for position knowledge, then the graph suggests that position is learned by subjects in all 4 conditions. There appears to be a gradual learning curve that commences at block 1 and increases steadily until the disruption at block 21. As to be expected, after the changeover, there is a decrease in the amount of pre-emptive moves made by subjects in the conditions where position has been randomised. The other two conditions appear to return to the learning curve established before the disruption. By the end of 40 blocks of trials subjects in these conditions appear to be making pre-emptive moves for an average of 10 out of every 16 trials. This result is extremely important since it suggests that learning has not ended, even though the traditional measures of reaction time and error rate have reached optimal levels. The question which then arises is when does learning ever end? Is experienced user performance level defined as being the point where classical performance measures reach asymptote, or should users reach EUP only when learning ends (if it ever does)?

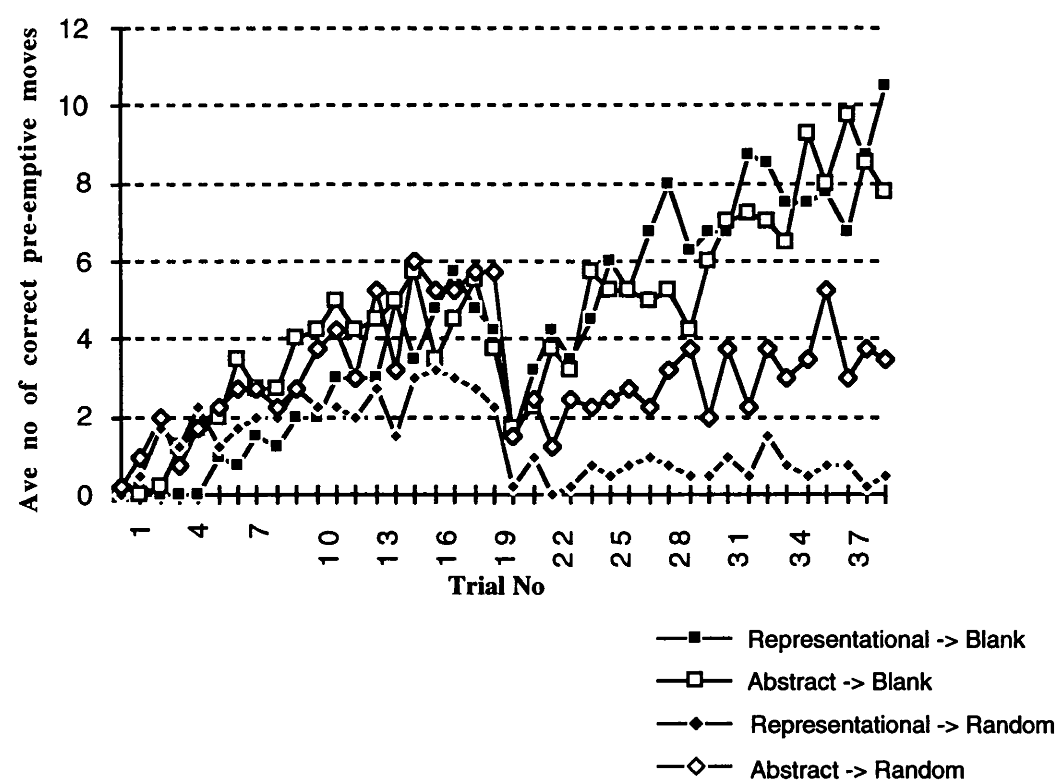


Figure 7.4 : Graph of the mean number of correct pre-emptive moves made in each condition across 40 blocks of trials (16 being the maximum number of moves possible).

This is a highly significant finding, but is it valid? A small application was created on Hypercard, called the "Plotter Stack"(see Appendix 9) . This accepted the co-ordinate positions collected from the experiments and visually plotted the co-ordinate positions and the trajectories of mouse movements relative to the icons present on the interface. Using the plotter stack it was possible to plot out the recorded direction the mouse had been traveling in before the interface appeared. From this visual representation of mouse movements it was possible to distinguish visually between intentional (and correct) mouse movements and any incidental movements that may have occurred. Only intentional movements indicate learning and therefore any movements not considered to be intentional were removed before the average numbers were calculated for the graph in Figure 7.4. Intentional movements are defined as movements that were recorded outside the area of the centre button before the interface appeared. Any movements falling within the area of the centre button were disregarded since, in such a small area, these movements could be as likely to be due to slight hand movements rather than intentional gestures.

7.6 Conclusion

The main aims of this experiment were:

- To evaluate whether the recorded disruption was due to surprise or learning.
- To discover when subjects began to learn position and whether subjects in all conditions learned it or just those using icons with abstract icon forms.

Results appear to suggest that it was not surprise that caused the disruption, since the disruption lasted longer than one or two trials. However, the disruption in almost all cases tended to be shorter than was predicted if the user was learning a completely new attribute. This suggests that subjects had learnt the attribute, considered in the initial hypothesis to be the recessive attribute, and so were able to return to a similar level of performance in a short period of time.

The pre-emptive move data convincingly supports the idea that position is learned by subjects in all conditions. The graph in Figure 7.4 suggests that there is a gradual learning curve commencing from block number one for all conditions. This confirms the assumption that people learn form first and then move on to learn position. However, the speed of learning position as opposed to form in this experiment could be influenced by the obvious performance advantage offered by knowing icon position. This is an extremely important point in interface design since it suggests that the interface context (e.g. the amount of information present on the interface) is going to have a strong influence on which attributes are used. This is discussed in greater depth in Chapter 8.

The graph also suggests that by block 40 users have not yet reached an optimal state in the number of pre-emptive moves they will make. It is possible that the curve will plateau only when the user is able to make pre-emptive moves for all icons in the set. Again this result is probably highly influenced by the high performance advantage offered by learning position. The interesting point to note, however, is that both reaction times and error rates have stabilised around block 30, yet the pre-emptive move data suggests that learning is still occurring. So, although learning appears to continue, reaction times and error rates fail to detect it. Therefore, other measures need to be considered if learning beyond what Jordan et. al (1992) define as EUP is to be detected and analysed.

The question that remains, however, is that if people in all conditions are simply learning icon position why is there a difference in the amount of disruption between the conditions at the changeover point? This question is considered in more depth in the general discussion in Chapter 8, where results from all six experiments are compared.

Although the main aims of this experiment have been achieved, it is interesting to consider why the results were so inconsistent with previous results. Why does the error rate data give us something like the results found in previous experiments, but the reaction time data does not? This could be a result of how subjects using the 16-icon sets perceive the imposed penalty compared with how subjects using the 2 icon sets perceive it. With 2-icon sets if the user selects the wrong icon then it is obvious that the correct icon must have been the other one. The penalty imposed is supplying the user with no useful information, and also slows performance. Therefore the penalty is perceived as it was intended - a penalty. However, for the sets of 16 icons, if the user genuinely doesn't know which icon is the correct one, incurring a penalty may actually be a benefit. Being informed of which icon should have been selected by incurring a penalty is perhaps seen by subjects as a way of increasing performance. Therefore reaction times may remain low at changeover because subjects genuinely do not know which icon to select and turn to the 'penalty' for help. This however means that the number of errors will increase, thus resulting in an pattern of error rates more representative of the pattern of reaction times found in other experiments at the changeover.

The final and extremely important finding from this experiment appears to be that learning continues beyond the asymptotes recorded by the classical performance measures of reaction time and error rates. This suggests that the learning curve continues beyond that originally specified. Does EUP commence when reaction time and error rates reach an optimal level of performance, or should it be defined as occurring at some later point in time?

8. General Discussion

8.1 Original Hypothesis

The original aim was to investigate several propositions:

8.1.1 Representational icon attributes are easier to learn than abstract attributes

This is a point that has been studied in some depth by previous work (i.e. Maissel, 1990; Rogers, 1986). The error and reaction time results from all 6 experiments do support this proposition. However, we need to look at the deeper consequences of this. Most work has looked at whether the form of the icon itself is the attribute required to be meaningful to the user. Various researchers have placed form on a scale ranging from highly representational (i.e. the object itself) to abstract. According to Lansdale (1988) "Geometric shapes can be assumed to lie toward the bottom of any putative scale of intrinsic meaningfulness".

However, when referring to 'representational' icons, we should not be simply referring to the form. We could equally be referring to the position or the colour. An example may be where the referent is the command "Move the cursor one character right". The interface only contains 2 icons (as in experiments 3 to 6); one to the left of centre and one to the right. Both icons are in the form of identical blank rectangles. It would be hoped that if the icon that represented the command to move the cursor right were positioned to the right, then this icon would (in this particular interface context) be representational. Therefore this example would still support the above hypothesis while both icons displayed equally meaningless forms.

Therefore the experiments suggest the commonly-accepted hypothesis that representational icons are easier to learn than abstract icons is supported. However, the suggestion made is that the hypothesis should be widened to include the idea that any representational attribute (not just form) is going to increase the speed of learning. Whether or not form is more influential than other attributes at increasing the speed of learning has not been shown and is likely to be entirely dependent on context. For example, for the command "move the cursor one character right" it is unclear whether the icon shown in Figure 8.1 described by Blankenberger and Hahn (1991) as 'representational' is more or less representational than having the physical position of "the icon on the right".

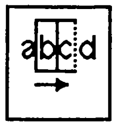


Figure 8.1: Representational icon used in the set of experiments to represent the command "move the cursor one character right"

What has been suggested by the results is that people may have come to expect that representational form should mean something and is therefore important to learn. This proposition is supported by comparing the reaction time data for the abstract-to-representational condition in experiment 4 with the data for the abstract-to-abstract condition in experiment 5 which shows that subjects who were presented with representational forms after the changeover performed slower than subjects who were given blank forms.

8.1.2 People given representational icon forms do not rely on position

The starting point for this research was that it was advantageous for users to perform the task with the smallest cognitive load possible. Therefore the most logical method would be to select the attribute that was considered to be the most representational, and use it to discriminate icons as representing particular commands. In the case of experimental conditions where the subjects were exposed to interfaces containing icons with a representational form, it was predicted that the subjects would rely on the form and ignore the position of the icon as a discriminating feature.

This statement is supported by the results found in experiments 2 and 4. In both experiments there is a massive disruption at the changeover for the representational-to-abstract condition which is almost as large as the initial learning of abstract icons as shown in trial block 1 of the abstract-to-representational condition. In both cases disruption lasts for 3 blocks (6 trials). There is a similar result in experiment 1, where 17-icon sets are used. The representational-to-abstract condition shows a massive disruption at changeover and only settles down 17 trials after the changeover.

The main evidence against this proposition however is the fact that subjects in both experiment 4 and experiment 6, in the representational-to abstract and representational-to-random conditions do use pre-emptive moves in the pre-changeover trials. The very fact that they do make a cursor movement towards the appropriate icon before the icon is available for visual inspection suggests that subjects **must** have learnt its position. The total number of pre-emptive moves for experiment 4 are shown in Figure 8.2. The last 2 conditions in the table have significantly reduced totals as a result of the dramatic decrease in pre-emptive moves after the changeover when position is randomised.

For experiment 4 the subject merely has to remember whether the icon is to the right or to the left of the central button on the command card, and then move in that direction in order to be recorded as having made a pre-emptive move. In experiment 6 there were 16 icons in a circular formation. The plots of pre-emptive moves suggested that subjects in all conditions were relatively accurate in moving the cursor to the area of the screen where they believed the icon to be positioned, thus giving stronger support to the idea that subjects using icons with a representational form learn not only that form, but also the position of the icon.

This support may have been further enhanced by considering the trajectories of pre-emptive moves. It could be assumed that the faster a subject moves the cursor to the icon position, the more confident they are likely to be about their knowledge of its location.

Conditions	**Correct Pre-emptive Moves**	**Incorrect pre-emptive moves**
rep-to-abstract	114	4
abstract-to-rep	202	2
rep-to-random	50	16
abstract-to-random	84	12

Figure 8.2: The total numbers of correct (moving towards the correct icon) and incorrect pre-emptive moves for each condition in experiment 4 (each total is out of a possible 560 trials).

However, it could be argued that purely the task itself motivated subjects in all conditions to learn position. Subjects in all experiments were asked to perform the experiment as quickly as possible (although accuracy was also emphasised). By moving the cursor to the position in which the icon is expected to be situated the user can potentially increase the speed of his or her performance over trials. Therefore it can only be speculated whether in day-to-day situations, when there is not such an emphasis to perform as quickly as possible, people will not be as motivated to learn both form and position.

Although the argument here is that people do learn both form and position the data suggests that subjects couldn't have learnt all they needed to know about what was originally thought to be the redundant attribute. The fact that there is an identifiable learning curve after the changeover particularly in experiment 6 (where there are more icon forms and more icon positions to learn) suggests that they were not only surprised, but are learning something. In the case of experiment 6 subjects in the representational-to-abstract condition are not learning the new form (which in previous experiments had been the abstract icon sets), since all icons after changeover in the representational-to-abstract condition are actually blank in this experiment. Therefore we can only conclude that the disruption is a direct result of subjects learning position.

In other words the situation is far more complex than was originally hypothesised. Although there is evidence that people do have to learn position after the changeover, there is the suggestion from the pre-emptive moves data that the position of some icons has already been learned **before** the changeover. Whether this learned spatial information is approximate (e.g. the icon is located in the top right quadrant of the circle) or absolute (e.g. it is the icon in the 12 o'clock position) cannot be specified.

8.1.3 People given abstract icon forms will rely on position

Defining this hypothesis more strictly, what is being stated is that in the conditions where there is no representational form it is predicted that users will not automatically use form in order to associate icons with commands, but will rely instead on position. For the 2-icon case position is likely to be easier to learn that form. Whether it is as easy for larger icon sets is not so obvious.

If this proposition were true then the abstract-to-representational condition should show no significant disruption at changeover, whereas the results for the abstract-to-random condition would be predicted to show a high degree of disruption at changeover as a direct result of the continuous randomisation of the positions forcing learning of the abstract shapes; a task which had been avoided in the first half of the experiment. The results in experiments 1, 2 and 4 do mirror this pattern, with reaction times in the abstract-to-representational condition returning to asymptote in less than one block.

The point has been raised, however, that these results could simply be an effect of the experimental design. In the abstract-to-representational condition subjects are actually receiving an additional guessable cue at changeover as opposed to one being taken away. Therefore in experiment 6, this condition was replicated, but at changeover subjects were presented with a second set of icons with an abstract form, instead of the representational set. The results were found to be identical.

A pre-emptive move is indicated when the user moves the cursor in the correct direction (as indicated by the x,y co-ordinates) before the interface card opens. Looking at the data of pre-emptive moves collected in experiments 4 and 6 (Figures 8.2 and 8.3) there appear to be far more pre-emptive moves made for the conditions where abstract icons are presented compared with conditions where the icon form is representational. This does suggest that learning position is more important in the conditions where form is abstract, even though the attribute is present as consistently and as prominently in all conditions.

Conditions	Pre-emptive Moves
rep-to-abstract	698
abstract-to-rep	732
rep-to-random	213
abstract-to-random	497

Figure 8.3: The total numbers pre-emptive moves for each condition in experiment 6. Maximum number possible would be 2720 (17 trials x 40 blocks x 4 subjects in each condition)

As for the previous proposition however, it can be argued that recovery after the changeover is faster than initial learning and therefore it suggests that users in the abstract-to-random condition are not having to try to learn something that is completely new to them. The fact that there is a learning curve suggests that they must be learning something, but the extent to which they have to learn the abstract form is unknown. This is partly because the comparison between the initial learning curve and the learning curve at the changeover will be confounded by the performance increase as a result of the subject becoming familiar with other aspects of the task as well.

8.1.4 Guessable and learnable attributes will achieve equivalent performance measures at EUP level

The graphs for both reaction times and error rates for all six experiments substantially support this proposition. By EUP level in almost all experiments each of the four experimental conditions have levelled to a steady and equivalent state. Since usability is such an important topic a discussion on this result is presented in a separate section at the end of this chapter.

8.1.5 Subjects are aware of both form and position, but rely on one more heavily than another

This proposition is supported by the results attained for experiments 2 to 5, originally developed to examine hypotheses 8.1.1 to 8.1.3. It appears that it is incorrect to assume that users uniformly rely on one attribute or another, but that they rely on a combination of both form and position. In other words subjects discriminate some icons by using form and others by their position. This would suggest therefore that there are no entirely 'redundant features', and that it is impossible to predict whether the user is likely to rely on one or other for a particular icon. To confuse the issue further it is also conceivable that subjects may use rough position e.g. to reduce the contrast set from 16 to 4 and then rely on form to identify the specific icon.

The situation is complicated further, since it appears that whether the appropriate icon is identified by either its shape or its position is not uniform throughout subjects. Therefore results suggest that some subjects will rely on the form of the icon in order to identify it, whereas other subjects in the same condition may rely on the position. For example, looking at the graphed plots for pre-emptive moves it is clear that one user in the representational-to-abstract condition in experiment 6 showed a bias for learning the position of icons situated on the right hand half of the circle. Another subject commented on the fact that he tried to associate the abstract icons with the commands, but there was only one that he could not create a viable association for. He could still select that icon correctly, however, because he remembered it as being "the one that didn't apply to any of the other commands".

Perhaps subjects use both shape and position for all icons. If so, then why is there this consistent difference in the degree of disruption between conditions? Instead it may be likely that they rely on only one attribute for each icon, but the attribute that they rely on may vary for each individual command.

This raises the question of whether or not we can predict the attribute that subjects were likely to depend upon. If subjects did rely on both attributes for all conditions why did the results consistently show that people using icons with representational form were more disrupted than those using abstract form when form was removed? Similarly why were people trained to use the abstract icon set more disrupted when position was randomised?

The answer may be that subjects in these conditions are more likely to depend on the attribute considered to be dominant (e.g. form in the representational form conditions, position in the abstract form conditions) as the distinguishing attribute for more icons, as opposed to all. Therefore, although subjects do show signs of relying on the attribute predicted to be of lesser importance for discriminatory purposes, they rely to a higher degree on the attribute predicted to be of importance. This would account for the results attained for experiments 1 to 6. Subjects, for example in the conditions using representational form, would use form as the distinguishing attribute for the majority of the icons, and position for the others. Therefore at changeover they are only required to learn the position of the icons on which they had previously relied on form for. Thus there would be a disruption and a degree of learning would be necessary, but not to the same extent as had originally been recorded during initial learning.

Out of these core propositions come many other issues:

8.1.6 Is the disruption is a result of surprise rather than of learning?

This proposition was raised to account for why there appeared to be a sharp spike suggesting disruption immediately at changeover, followed by a drastic return to asymptote soon after. Is this disruption the result of learning, or simply a sign of derailment as a result of monitoring effects? In other words is the delay due to surprise at the change, rather than the users' inability to output a fast response?

In the 2 icon set experiments this was impossible to tell. Initial learning at the start of the experiment was almost immediate, therefore once performance effects had been added, even if learning were to occur at the changeover learning would happen so quickly that it is unlikely that it would show up in the results.

However, looking at the results for experiment 6 it is possible to graph a gradual learning curve after the changeover. The learning curves for the representational-to-blank condition and the abstract-to-blank conditions return to asymptotic performance levels by block 23; 3 blocks of 16 trials after the changeover. The other 2 conditions never return to a similar performance level achieved before the changeover. This is to be expected for the abstract-to-random condition since it was predicted that subjects would be relying on position. However it was not expected and had not been displayed in previous experiments for the representational-to-random condition. This anomaly adds support to the idea that subjects presented with icons displaying representational forms do rely on position as well.

8.1.7 Is some information always harder (slower) to use even at asymptote?

Although it may be easier to discriminate between icons by relying on their form (particularly when the form is representational) the experimental design suggests that if users can learn the icon positions then they can benefit by being able to make pre-emptive moves and perform faster. Therefore there is some motivation for subjects to learn position even though it may involve a slightly greater cognitive load.

This point is emphasised by the fact that the reaction times for the abstract-to-random condition, in all experiments, never return to a performance level equivalent to the one achieved before changeover. This is as a direct result of the removal of constant icon positions and therefore a reduction in the success of the pre-emptive move strategy.

8.1.8 Abstract icon form takes longer to learn (reach asymptote) but is the asymptote the same as for representational form?

The premise that the asymptotes are the same is supported by the reaction time results for experiments 2, 4 and 6. Experiment 2 and 4 suggest that the asymptote for the representational icons begins at block 5, whereas in experiment 6 (with sets of 16 icons as opposed to 2), it commences at block 8. However, it is not until block 8 in experiment 4, block 9 in experiment 2 and block 15 in experiment 6 that the abstract icon sets reach the same level of performance.

These results could suggest that if given abstract icons people learn both position and shape and so learn twice as much and therefore learn much more slowly. Alternatively it could be that users find it hard to learn form, and therefore it is only by block 8 that they have learnt position and thus can perform more efficiently. The graph of pre-emptive moves made in experiment 6 supports this latter suggestion indicating that there is a gradual increase over trials in the amount of moves until reaching an asymptote at block 15 (for all conditions) which corresponds to the achievement of an asymptotic level of performance in the reaction time data for the conditions using abstract form.

Experiment 4 used a switch at block 5 by which time, according to the data from other experiments, only users trained to use representational icons sets should have reached asymptote. The results suggest that the biggest disruption occurred for the 2 conditions trained to use the representational icon sets; the least disruption was for the abstract sets. This could possibly support the idea that users in the abstract conditions do rely on position. Position takes longer to learn and therefore by the changeover the subjects are not yet relying on position, thus are not so greatly disrupted.

8.1.9 Results will be the same whether there are only 2 icons or 17?

Experiment 1 used sets of 17 icons. Experiment 4 used sets of 2 icons. The reaction time results for each are similar: the representational-to-abstract condition has the biggest disruption, followed by the abstract-to-random condition, the representational-to-random condition and lastly, the abstract-to-representational condition. This is quite surprising, considering that it would seem logical that it would not be particularly hard to learn the abstract icon form when there are only two to learn and therefore a more uniform performance across conditions would be expected.

Experiment 6 was essentially performed to validate the results from experiment 1. Since this earlier experiment, the experimental design had been improved, and therefore it was important to see whether the same results could be achieved. The same results were not achieved, however. Instead for experiment 6 the abstract-to-random condition suffered the biggest disruption, followed by the abstract-to-blank, representational-to-blank and representational-to-random respectively. Whether this change was a direct consequence of the increase in the number of icons/commands or their positions (in a circular format as opposed to a rectangular one in experiment 1), or even as a result of the degree of penalty imposed, is unclear.

However, looking at the error rates for experiment 6, a more similar pattern of results to those found in previous experiments appears. From this it could be argued, therefore, that the severity of the penalty is different for experiment 6 and as a result what was seen as a penalty in other experiments was instead seen as a help option for this more complex interface. Therefore at changeover, users in all conditions were thrown by the change to an equal proportion as equivalent users in other experiments. However instead of trying to select the appropriate icon they felt that it was in their interests, in respect to the time spent, to select any icon and be informed of the correct icon (thus resulting in a similar pattern for error rates as opposed to reaction times).

8.1.10 Are people more likely to rely on icons when the icons are in specific positions?

Blankenberger and Hahn (1991) reported that the highest reaction time speeds were reached for icons that were positioned at the screen corners. Thus they advised that the more frequently used icons should be situated at the edges of the screen. Results from experiment 1 however were unable to support this suggestion. In experiment 6 the icons were displayed in circular formation. If people learned specific positions then it would be predicted that the data collected to plot cursor movements would show subjects making more pre-emptive moves to the positions that they recalled best. It was thought that the icons situated in the clock positions 12, 3, 6 and 9 would be best remembered using position. However the plot data did not support this. Perhaps had there been twelve icons as opposed to sixteen, set out in exact clock formation, subjects may have relied on position more. As a result of all icons in this context following the same layout (e.g. representing a clock face) each icon position could be described as more representational, assuming that by representing something that the user can identify from their previous experience (e.g. a clock face) classifies the position as being more representational than when the positions form a concentric circle, but without the same degree of association to external knowledge. It may be possible to suggest that although performance is not faster for certain angular positions, chances of a user relying on position as opposed to some other attribute may be increased when the icon is situated within a distinctive position. Therefore icons in corners may be remembered by their position since a corner is a distinctive position on the screen relative to other positions. For experiment 6 it could be argued that the 16 icon circular formation prevented any one position from being considered more distinct than another, resulting in no particular pattern in which icons users tended to identify using position.

8.1.11 Abstract forms are still easier to learn than (abstract) position?

Data for experiment 6 suggests that users' reaction times in the abstract conditions decrease proportionally with the number of pre-emptive moves made. This would be expected, since moving the cursor to the icon position before the icon appears, as opposed to waiting until the icon is visible before moving the cursor is logically going to reduce the time spent per trial. Therefore are people actually learning only position here or are they learning the shape as well? The data doesn't really give a clear answer to this. Although subjects may be using position for some of the icons there is no guarantee that they are not relying on the form for other icons. Therefore it could be argued that during the initial part of the leaning curve they are learning the form of some icons, and over time learning the position of others.

8.2 Discrimination and interface context cues

8.2.1 The revised hypothesis

Until experiment 6 it was thought that the most likely explanation for the results was that users were selecting the distinguishing attribute that involved the smallest cognitive load to learn and relying on this characteristic in order to identify the appropriate icon for the command.

There was an issue of how much people actually learned about the 'redundant' attribute. It was suggested that the fact that changeover did not result in the user reverting to reaction times or levels of error similar to those recorded at the start of the experiment was because users forced to rely on the redundant characteristic had learnt something about the characteristic before the changeover.

Results from experiment 6 complicate matters by suggesting that the question of what people either learn **or** rely on is not a simple one. It appears that reliance on a particular property when a standard set of characteristics are present is not an absolute (and therefore predictable) phenomenon.

To re-state, the original hypothesis was that people rely predominantly on one characteristic. For the experiments using form and position it was predicted that when the icon form is representational subjects rely on form, but when abstract they rely on the position. Even if, due to some experimental design flaw, this were not supported through reaction time and error data, it would be assumed that it could be supported by data collected to assess the extent of pre-emptive moves for each condition.

Results from experiment 6 suggest that there is no significant difference in the number of pre-emptive moves made by subjects in each condition. Instead it appears to be more subjective. In some cases (independent of condition) subjects show evidence of pre-emptive moves for all commands, while for other subjects pre-emptive moves are only made for certain commands. For all conditions the pattern suggests that subjects gradually use position more and more until the changeover at block 21 where there is a dramatic reduction, followed by a second steady increase in the use of position towards the end of the experiment. It would be expected that conditions which have randomised positions in the second half of the experiment would show a reduction in the use of position after changeover when position is randomised. This is certainly true for the representational-to-random condition. However, subjects in the abstract-to-random condition still show a reduced, yet active use of pre-emptive moves, although these moves often tend to be in the wrong direction.

It could be argued that these results are purely an artifact of the experimental design. Since subjects are required to select the appropriate icon by moving the mouse to the correct position and clicking on it, there is obviously going to be a natural motivation, whatever the condition, to learn position thus allowing the user to make pre-emptive moves and therefore speeding up performance. However, it could be argued that the identical motivation was also present in all the other experiments, so why did people in all conditions not use it, for example, in experiment 4?

An alternative explanation is that subjects in the conditions using abstract icons were not motivated **more** than subjects in other conditions to rely on position, because the icons were positioned to form a concentric circle. It could be suggested that placing the icons in a circle makes each position less distinct from others. Plots were made to analyse whether there was any certain pattern to the pre-emptive moves (e.g. subjects relied more on position when icons were situated at the 12, 3, 6 and 9 positions as would be found on a clock face). This was not found to be the case. Therefore the suggestion that results are affected by the lack of any distinct positions may actually be valid.

This suggests that people faced with a set of abstract icons will only rely on position when it appears to involve less of a cognitive load. It could be argued that the more icons that are present then the more difficult it becomes to learn position, and therefore it becomes harder to predict that users will use position as a distinguishing characteristic. The trouble with this argument, however, is that this contradicts the results for experiment 1 which used 17 icons, the results of which initiated this set of experiments. It must surely be argued that it was not the number of icons that diminished the likelihood of users relying on position, but the relative positioning of the icons which may have appeared rather indistinct. In experiment 1 the icons formed a rectangle. The placing of icons at the corners appeared not to result in any increase in reaction times or reduction in errors for these advantageously placed icons, but they may have acted as suitable reference points from which to remember the positions of other icons.

This would suggest that people will rely on the characteristic that is most easy to discriminate from a set of characteristics. In other words when faced with a forced choice, users rely on the attribute contained in the icon which has the greatest contrast to other icons within that set. Logically, therefore, it is impossible to say that any particular property is going to be **the** property that people rely on until you know the degree to which the property value varies from the value of that property for other commands. Some contrasts may be particularly obscure. For example, a subject in experiment 6 said that he could remember a particular abstract icon since it was the only one that he could not find any contrived association to in order to discriminate a particular icon for a command.

One further interesting point is that the commands that most often resulted in pre-emptive moves (irrespective of condition) were the commands 'print' and 'underline'. These are distinguishable from the rest of the commands in that they are the only commands which are not part of a pair (e.g. cursor one line up, cursor one line down), or a set (e.g. delete one line, delete one character, delete the whole text). If this is the case, then the converse should also be true; commands with meanings which are particularly confusing or indistinguishable from one another should have the smallest number of pre-emptive moves. The experimental results, however, do not support this theory. Perhaps this was simply because there were no commands significantly more indiscriminable than the rest.

Again it may logically follow that the attribute will be more discriminable if there is greater distinction between icons sharing that attribute. Therefore if the attribute was hue, having a huge range of colours as opposed to different shades of the same colour is going to reduce the probability of error in selecting the appropriate icon by relying on colour, and thus increasing the chance of that attribute being relied upon to distinguish that particular icon. The greater the contrast between the attribute value as it is represented on the target icon, compared with the other icons on the interface, the further the possibility of error is reduced, and therefore the greater probability that the user will rely on this attribute to identify that icon. So for example, if the target icon were white and all others were black, then we could be almost certain that the user would rely on hue in order to identify this particular icon. This suggests that using multiple properties in an interface, with sets of icons sharing similar visual characteristics for that property and a few having significantly different features of the property will result in subjects relying on different properties for different icons, but managing to perform accurately and quickly.

It could even be suggested that in some cases people may use different cues for different objects in the set, and in effect associate not values but attribute types with a given object. This would however depend on there being equivalent numbers of unique attributes to icons. For example, it may be that users recognise the print icon by its outline, the save icon by the colour etc.

It may be worth considering that there may be a 'natural order' for learning and using attributes. Therefore users may attempt to learn attributes that can supply then with some degree of meaning (most commonly form) and then move on to attributes that they know will increase their performance speed or reduce cognitive load (e.g. position). After this they may learn other attributes, perhaps to be used as redundant cues which will be used if the attribute they are relying upon becomes unreliable.

Previous work (e.g. Norman, 1991) has argued whether too many "semantic markers" may cause confusion, whereas others argue that the more semantic markers that are present then the quicker you will process the icon and respond to it. It could be argued that the more semantic markers present the more likely that there will be one that is particularly distinguishable for that command. The work done so far suggests that this may indeed be true, but is merely suggestive and provides no strong evidence for it.

It is unclear from the results however, whether there is some order in which people try to discriminate icons for particular commands. Do they, for example, try to associate form of the icon with the command first and then position? No suitable data is available from the experiments done so far to answer this question.

8.3 Usability, learning and interface context

It was apparent that previous research on iconic interface design has rarely considered design beyond the guessability stage of usability. Therefore one of the main thesis aims was to consider the effects of contextual information on user performance throughout all stages of usability. Although guessability is important, the main thrust of this work has been on learnability: what determines what is learned and how quickly. Issues raised by the experimental results are discussed in greater depth in the next three subsections.

8.3.1 Guessability

Previous research has supported the idea that representational attributes are more likely to increase the probability that an icon is guessable. This finding was again strongly supported by the experiments reported here. For all experiments the conditions in which form is representational show faster reaction times results and fewer performance errors in the guessability stage than when form is abstract. It could be suggested that in the guessability stage users are attempting to attribute meaning to new icons; looking for relationships that will signify that this icon is the correct icon for the command. Representational icon form supports this task and therefore performance is faster.

However, even at the guessability stage it is possible that context is an important influence. The degree to which the physical properties of the icon differ from those surrounding it will perhaps affect the likelihood of those properties communicating their meaning to the user. In other words the contrast between the icon's attributes and those attributes of the surrounding icons (the contrast set) will determine whether the meaning people attempt to pick up and to associate the icon with the command can be obscured by other icons that contain attributes that appear to the user to communicate roughly the same message. This is a problem which should occur irrespective of how representational icon form may be. Conversely, the contrast set may have a positive influence, and actually aid an icon containing non-guessable attributes to be easily identified, due to its distinct appearance.

Another point to note is that there may not necessarily be a uniform attribute by which all icons in the interface are distinguished at the guessability stage. Where there is more than one representational attribute then it is possible that users will choose either of these attributes. Looking at the commands which contain left and right in their instructions, it is possible that in the 2 icon cases left commands with icons positioned to the left and right commands with icons positioned to the right will be considered guessable. Although it is impossible in the representational condition to say whether users are relying on position or form in the first trial, the fact that in the abstract condition the fast, error free response in these conditions suggests that they did rely on position, thus implying that there is a high possibility that subjects could have relied on either form or position in the representational conditions.

Although there is no experimental proof to support it, it may not necessarily be the case that where there is more than one representational attribute present that users will automatically rely on form, as is assumed by most of the previous research done on icons so far.

8.3.2 Learnability

The Gibsonian view of perception (and possibly of visual contextual cues) is that there is always information there, but whether you perceive it or not depends on whether or not you have the appropriate visual apparatus. Gibson (1979) believes that information pickup occurs over different times and therefore people are constantly picking up new information about an object they are familiar with. This corresponds with Kaptelinin's work (1993) which concludes that there is a shift in attention over time. Users may start off by noticing the local features, i.e. the icon form, and in the learnability stage gradually pick up and shift focus onto the more global features of the icon (i.e. its outline, or its spatial position). The experimental results for this thesis suggest that users do rely on form first, at least when the form is representational. This is supported by the superior performance by subjects in the representational icon conditions compared to subject using abstract icon forms at the start of all experimental sessions. However, results from the pre-emptive moves data in experiment six suggest that it is possible for users to switch the attribute that they rely on, after some learning has occurred, if the new attribute offers better performance levels or less cognitive effort. Thus in experiment six there is evidence that users in the conditions using the representational icon form switch from relying on form to discriminate many of the icons to position because position offers them better performance advantage. Results for the abstract conditions however fail to highlight whether people are attempting to learn form, or are instead simply using trial and error until they are able to learn the position. Therefore, for the representational form conditions the explanation provided by Kaptelinin is that there is a shift away from reliance on the local feature (form) to reliance on the more global attribute of spatial position (although it is unclear from the results whether position is being used exclusively or still with form).

Retrospectively it appears quite naive to assume that once a person has learned one attribute (e.g. form) he or she will rely on this and ignore all other attributes from then onwards. However it is not possible to suggest, as Kaptelinin does, that there is a uniform change from local features (e.g. the form of the icon) to global features (e.g. the position of the icon on the interface). If this had occurred then it would have been the case in the representational-to-abstract condition that there would have been no disruption at changeover, since everyone would have switched to using position by this time. Similarly, people in the abstract-to-random condition are still able to perform to a certain extent after changeover suggesting that they must have learnt to some degree the abstract icon form

Subjects using the experimental paradigm established for this thesis (and also most iconic user interfaces) are motivated to use position because of the type of task they are required to do. Using the pointer to click on an icon situated at a consistent screen position makes this task highly position-oriented, giving users a performance advantage if they learn position. Therefore it can only be assumed that this is why there is such a shift in strategy. This would suggest that it may be possible to predict that wherever there appears to be some performance advantage gained from using a particular attribute, then it is likely that users will shift attention from the guessable attribute to the attribute that offers the smallest cognitive load or the best performance. Logically it should follow (although this is not tested by the thesis experiments) that where the global attribute might not offer such an advantage, users will continue to rely on the local feature, suggesting that the shift in attention that Kaptelinin predicts may be more dependent on context than he suggests.

It appears that there is a possible shift in the learnability stage away from attributes which intrinsically contain meaning to those attributes that allow easier discriminability (and therefore faster performance). This is an extremely important point, and essential for designers.

8.3.3 Experienced User Performance

The important point to note from the results of all experiments is that by the time users have reached the level of experienced user performance (indicated by the performance levels reaching an asymptote), the asymptotes for representational-to-abstract and abstract-to-representational conditions are identical. In other words, conditions with icons which have no guessable attributes show similar levels of performance to the icons with the representational attributes. This may be because users at this stage of experience have shifted strategy to rely upon those attributes that are perceived to show the greatest amount of discriminability for each icon. Results, however, suggest that with guessable attributes present, users reach asymptotic level much faster than those with no guessable attributes. The question still to be answered is that once users reach asymptote in all conditions are they using identical attributes (i.e. position) or different ones? The fact that in all experiments different degrees of disruption occurred at changeover in each of the four experimental conditions suggests that people faced with different contexts will not rely uniformly upon the same attributes.

Mayes et al. (1988) suggested that experienced users can be defined as people who are able to use the interface using the smallest possible cognitive load. As an example they refer to touch typists, who must have originally relied on visual search of the keyboard; but this has now become ‘incidental’ as position has taken over. Ask any touch typist where a certain letter is on the keyboard, and it is highly likely that they will have to recall which finger they use in order to answer it. Certain commonly used words (i.e. ‘the’) will be typed without thought. In other words, experienced users have developed sequences of skilled task executions, thus are no longer concerned with the inherent meaning of the attribute, but are more concerned by how easily it allows discrimination and performance advantage in relation to the task being performed. By learning key position typing speed can increase. Therefore experienced typists may have forgotten what the key looks like (e.g. the font used on the key, its colour etc.) but instead rely on its position. Therefore it would be predicted from this that by the EUP stage users have settled down to identifying particular icons by set attributes, and therefore although they have learnt other attributes, these tend to be discarded, thus reducing the cognitive load to an optimal minimum. This is, to an extent, supported by experiment 3 which was designed to investigate the degree of learning before EUP was reached. The results suggested that there was a disruption when both position and form were altered, but to a smaller degree, and not following the same pattern of disruption that was found with the previous experiments. This would suggest that by the fifth block of trials users were aware of the learnable attribute, icon position, but had not settled into a routine of relying on either one or the other of the cues.

Which attribute offers the smallest cognitive load is entirely context dependent. The set of icons within which the icon is situated will determine which attributes of the icon are particularly distinct in comparison with the surrounding attributes, thus determining how efficient that attribute will be to use within this set of icons. This may suggest that what is meaningful and what is discriminable are actually two different things, and that perhaps, over an extended period of time, meaning is not so important. What is important is that there is some feature of the icon that is discriminable enough within an interface to reduce the amount of effort required by the user to identify it.

Results have also suggested that an asymptotic performance rate does not necessarily indicate that users have actually reached full experienced user performance level. If Gibson is correct then it is highly likely that people are consistently picking up new information about objects they are familiar with. It is possible that the attributes that users become aware of in the guessability and learnability stages will allow the user to perform optimally. Therefore with the paradigm used, even though users may have continued to learn new pieces of information no resulting improvements in performance would have been recorded. This was highlighted by the pre-emptive move data in experiment 6. Although standard performance measured (i.e. error rates, reaction time) had reached asymptote the pre-emptive move data suggested that user were continuing to learn icon position and hadn't reached asymptote in this learning curve even by the end of the experiment (40 blocks of 16 trials). Therefore it is possible that although reaction time and error rate performance could not be improved on, the strategy behind this continued learning was a reduction in the cognitive load involved in performing this task.

8.4 Chapter Summary

To sum up, what the original hypothesis suggested was that users guessed the representational attributes and for cases where there were no guessable attributes they had to learn one. Once one reliable attribute was available to distinguish icons by, this was used consistently from the point of learning onwards. What the results have shown is that users in the guessability stage will attempt to associate icon meaning with the underlying command, thus relying on representational attributes at first, but once in the learnability stage, as other less guessable attributes become apparent may switch to rely on another attribute if it appears to be more visually discriminable than the attribute the user was originally relying upon.

Therefore, as suggested by the experimental results, depending on interface context, where icon position is particularly distinctive it offers a performance advantage and perhaps lower cognitive effort (as a result of moving from local to global feature processing as suggested by Kaptelinin). As a consequence users will very often shift to using position. This, however, is not a uniform transformation. It appears that users decide which attribute to use on an icon-by-icon basis, thus making predictions about which attributes subjects will use difficult.

By apparent EUP level it is predicted that users will have become proficient in performing on the interface using an optimally minimum cognitive load. No more adjustments to their strategy should be made, although it is possible that some fine tuning will occur as users become aware of other possible distinguishing attributes. However, if any tactics are changed they are likely to be changed to reduce cognitive load rather than to improve performance measures, therefore making it impossible to pick up this learning in the experiments performed, should it have occurred.

Any changes to the interface will cause user reaction (as shown by the changeover in all six experiments). Users are likely to re-evaluate the interface and make adjustments to any strategies if contrasts are no longer as distinct as a consequence of the changes made.

9. Applying new findings to an old experiment

If the findings achieved for experiments 1 to 6 are indeed valid, then it should be possible to use the conclusions to explain results from experiments looking at other icon attributes in similar terms. An experiment performed by Moyes and Jordan (1993) was taken as suitable example. This experiment was completed before any of the thesis experiments had been performed. It was considered a suitable experiment to use since it studied the consistency of attributes present in the interface throughout the learnability curve. It uses Jordan et al's (1991) multi-component definition of usability to define the stages of usability and thus allows direct comparison between the results in the different learnability stages with those found for this thesis.

Most of this chapter will be concerned with explaining the experimental hypothesis, paradigm and results in order to give the reader a clearer idea of why the original conclusions were made. Following on from this is a discussion of how what has been learned from this thesis can be applied, to supply an alternative, more insightful conclusions to the results of the Moyes and Jordan experiment.

9.1 A definition of consistency

For the purpose of understanding why the original experiment was performed it is important to explain how the formalism, created by Jordan (1993), was derived.

An exact definition of consistency is elusive. The most simple definition possible is one suggested by Reisner (1990); "doing similar things in similar ways". In order to gain a firmer grasp on a definition (and consequently to highlight inconsistencies) research has aimed to develop suitable formalisms (e.g. Reisner 1981, Payne and Green 1983, Kieras and Polson 1985, Payne and Green 1986, Payne and Green 1989, Reisner 1990).

The formalisms developed before Jordan generally decompose the methods for performing tasks into a series of action rules. If tasks are done in "similar ways" then a common rule can be applied to each of the tasks in the set, and all tasks can be described as consistent. A task that requires separate rules in order to be described would therefore be identified as inconsistent. The more additional rules required, the more inconsistent the task is deemed to be from the set.

The main difference between formalisms developed has been how they define "doing similar things". Therefore Set Grammar, developed by Payne and Green (1983), and Task Action Grammar (TAG), also developed by Payne and Green (Payne 1984, Payne and Green 1986, 1989) regard sets as being representative of the way that the user would group tasks, rather than as a function of the interface. Payne and Green identified this as being "grammar in the head". Furthermore, the type of user capable of grouping the tasks into sets would be classified as an ideal, experienced user of the interface. Thus Set Grammars and TAG represent user 'competence' rather than performance (e.g. what a user can do as opposed to what they actually do).

Reisner (1990) goes beyond previous notations by suggesting that although there is a need to establish common criteria for putting tasks into sets, there are no universal laws of semantic grouping. She regards it therefore as more constructive to think simply in terms of users giving labels to things, and putting tasks into sets on the basis of them sharing a common label. Based on this idea she developed a framework for identifying consistency called Agent Partitioning Theory (APT).

In APT semantic groupings are simply decided by individual people. Inconsistencies, then, may be viewed as arising from a user grouping things differently from the designer (whose groupings are embodied in the interface). Inconsistency is therefore a property of a human-machine interaction. It is dependent both on the individual user and on the design of the interface rather than being solely a property of either the interface, or grammar in the head. This is the key point made by Reisner (1990) that had been missed by earlier work.

To describe which things belong in which schema set, assignment rules are required, for example, square [SHAPE]; which specifies that the task is 'draw a square' and the actions assigned to this are those that would be performed for any shape. It is also important however to identify who has made this assignment, and therefore an assigning agent has to be specified. The assigning agent could be the designer, whose assignments are embodied in the machine or a competent user, a user who makes mistakes common to many users, and probably ascribable to some design feature. Reisner calls the combination of the assignment rule and the assignment agent a partitioning rule (e.g. square [SHAPE] for agent designer). The central claim of APT is that when a system is inconsistent different assigning agents use different assignment rules. Specifically, in the case of interface inconsistency, the designer and the competent user apply different assignment rules. In a consistent system both user and designer will partition a given task set into the same schema sets. These schema sets can then be used in the generalised rule schemata - the similar ways of handling similar things.

Jordan (1993) makes the point that APT highlights 2 separate issues. It is possible that the degree of difficulty of learning will be influenced by the number of different rules required; the fewer rules the easier it will be to learn. However, he also points out that learning will be affected by the degree of consistency between the rules and the previous experiences and expectations of the user. Therefore the same rule could be used throughout the interface (e.g. for all tasks go to the menu and drag to the appropriate command then release the mouse button), thus ensuring that the tasks will be consistent, but how the tasks have been grouped under menu headings may be inconsistent with the user's notions of what the commands actually represent. For example, the commands 'Open' and 'Save' in a word processing document may be expected to be found under the menu heading 'File' since they are both operations that operate on entire files, whereas 'Spelling' and 'Word Count' may be found in the menu with the heading 'Tools' since they are both utilities that can be used with an appropriate word processing file. It may, however, be equally conceivable that a user interprets 'Word Count' to be an operation that is performed upon an entire file, and therefore consider it to be consistent if placed under the 'File' heading. What this suggests is that it is unlikely that inconsistency can be adequately described without a discovery procedure that describes and compares both designer and user task groupings and looks for mismatches. This suggests that the identification of consistency (or inconsistency) is an empirical question.

Jordan's solution was to develop another classification identifying these differences in consistency. He labelled these two possible distinctions "set compatibility" and "rule compatibility".

9.1.1 Set compatibility

This refers to the first issue raised; the degree of difficulty of learning will be influenced by the number of different rules required. Thus the tasks that the designer groups as similar and the tasks the user groups as similar should be performed in the same way, so requiring only one set of rules. Therefore Jordan recommends that only one icon should be used per task. For example, looking at the Macintosh interface, it is set compatible that the wastebasket is used whenever the subject wishes to delete something. It is often considered set incompatible, however, to insert a disk icon into the wastebasket which results in the disk being ejected from the disk drive, instead of being erased.

9.1.2 Rule compatibility

Rule compatibility is concerned with how 'natural' the task appears. This relates to the second point raised by Jordan; learning will be affected by the degree of consistency between the rules and the previous experiences and expectations of the user. Thus in the example given above, the placing of the command 'Open' under the heading 'Tools' may be rule incompatible to the user, if she expected all commands under the menu heading 'Tools' to refer to additional utilities that can be used on word-processed documents. In such an instance difficulties will arise, not because there is an extra rule schema, but because of the 'unnaturalness' of the correct rule; e.g. it is rule incompatible.

A system, then, will be rule compatible if features of the interface are compatible with the user's expectations based on experience of the "outside world". Included in experience of the outside world are: experiences of doing similar tasks in daily life; experience of using similar machines; experience of using similar interfaces; experience and knowledge of performing other types of tasks on the same machine (e.g. tasks that the user regards as being in a different set).

Investigating the effects of set and rule compatibility on usability, Jordan (1993) studied users performing command activation on a menu-driven interface. Results suggested that rule compatibility had a significant effect during early interactions (guessability and learnability), whilst set compatibility mattered more later on (learnability and EUP). The aim of the experiment performed by Moyes and Jordan (1993) investigated whether the same results could be found studying users performing with an iconic interface.

9.2 Set and rule compatibility and their relation to iconic interface design

Jordan (1993), using his own formalism, equates 'rule compatible' with 'representational'. It would seem possible that the terms 'representational' and 'rule compatible' could be used interchangeably, since they appear to have the same underlying requirement; that the interface contains information that relates to a user's "world knowledge" or previous experience. Therefore when asked to select an appropriate icon for the command 'Print' it may appear 'natural' to select the icon in figure 9.1 from amongst a set of icons. If, however, none of the icons stood out as likely to be for printing, then the user may not know which icon to use; the actions required for completing this task would then be rule incompatible (e.g. having to select an icon that doesn't appear to represent the command).

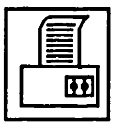


Figure 9.1: Representational icon representing the command 'print' (Blankenberger and Hahn, 1991)

In an iconic interface, set compatibility would concern the universal application of an icon; whether a particular icon could be used a similar way for similar tasks. An interface may contain, for example, a set of object icons (representing files, graphics, or data), and a set of operation icons (representing deleting, printing, or saving). If the same operation icon could be used for, say, printing a table and printing a file, then the action rules for printing tables and files would be set compatible, as they would both contain similar actions, clicking on the same print icon.

As has been supported by experiments 1 to 6, when using computer-based icons, guessability is not the only important component of usability. Since users of iconic interfaces may use the interfaces for long periods of time they get a chance to interact with the icons. If they are unsure of an icon at first, the user may still be able to learn its functionality through interaction. Thus, although a badly-designed icon may not be guessable, it seems possible that it might support reasonable levels of learnability and EUP.

Therefore, based on this idea, and on the results from the experiment performed by Jordan (1993), it can be predicted that if representational type is a rule compatibility issue, effects of representational form on performance should be found for early interactions, but be of less importance later in the learning curve. This is in contrast to set compatibility, which should have little effect early on, but become a more salient issue later.

Using this hypothesis an experiment was designed (Moyes and Jordan, 1993) to investigate the effects of representational type on guessability, learnability, and EUP. These effects were compared with those for set compatibility.

9.3 Consistency Experiment (Moyes and Jordan, 1993)

9.3.1 Experimental overview

The icons used in this experiment (shown in Appendix 8) were taken from the selection of icons studied by Maissel (1990). These were chosen for several reasons. Firstly, the work done by Maissel is experimental rather that theoretical, therefore any conclusions made by Maissel have been experimentally validated. Secondly, Maissel's work is exceptional in that his findings are based on the results gained from more than one experimental method. Finally, apart from being the most recent and extensive piece of experimental research that we had access to, the findings were particularly useful for our purposes since the icons for each operation were ranked in accordance with the results from the experiments. This meant that we were able reliably to select the icon that Maissel's research had rated as the 'best', one which could be described as 'mediocre' and also the 'worst' icon for each operation.

9.3.2 Experimental design

An interface was created using HyperCard (version 2.1) to run on the Apple Macintosh SE/30 (system 7.0). It consisted of sixteen object icons (each representing a different type of file), and eight operation icons (representing different types of file manipulations). Each icon measured 1.5cm x 2cm. Four types of manipulations were simulated by the interface: saving files to disk, copying files, deleting files and printing files. Users performed these tasks by clicking first on the appropriate object (file) icon, and then on the operation (manipulation) icon. They were then given feedback about the effectiveness of their actions via a textual message on the screen. In the case of a successful operation this might read, for example, "Text file successfully saved to disk", or, if the user made inappropriate inputs, "Please try again". The experimental design used for the experiment was a within subjects design.

There were a total of four interfaces, each containing tasks of types A, B, C, and D. These task types are defined as follows:

A: Looking at rule compatibility with set compatibility. Subjects are presented with both the icon rated as the best and also the one rated as worst for that function. However, it is only the best icon that is the correct one for all cases.

B: Looking at the effects of set compatibility without rule compatibility. There is one bad icon and one good one, however, only the 'bad' icon is correct for all cases.

C: Looking at the effects of rule compatibility without set compatibility. Best and worst icon presented, however in three of the cases the worst icon is the correct one to use, and in the test case it is the best one.

D: Neither set nor rule compatible. Here there is one mediocre icon and one bad icon. For three of the files the mediocre one is the correct icon and in the test condition the bad one is correct.

The interfaces were designed to balance task for task type. So, if for one interface, saving a file were a type A task, then on another it might be a type C task. Balancing was done via a Latin square. The position of icons was determined at random for each of the four interfaces.

The experiment was designed so that each type of manipulation would be performed on four types of file. Depending upon whether or not tasks were to be set compatible, it would be necessary to use a total of either one or two operation icons in order to carry out a particular manipulation on all four file types.

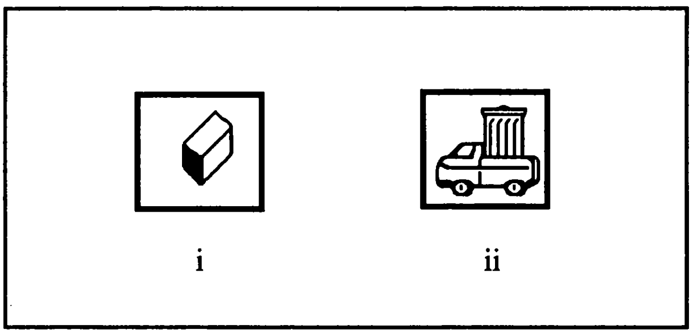


Figure 9.2: Examples of rule incompatible and rule compatible icons

For example, if the icon in Figure 9.2(i)were to be selected to delete spreadsheet files, graphics files, data files, and programming files, then the interface would be set compatible for this group of tasks. However, if this icon could only be used for deleting spreadsheet files, and the icon in Figure 9.2(ii) were used for printing graphics, data, and programming files then there would be a set incompatibility. The icon to be used for a particular task was dictated by whether the task was to be rule compatible or incompatible. For each of the four file manipulations there were two potential operation icons, one of each pair was highly suitable for the operation it represented, whilst the other was not suitable (the basis of these judgements is the work of Maissel, 1990). So, if a suitable icon were used for a manipulation, this would be a rule compatible task, but if the icon were unsuitable the task would be rule incompatible.

9.3.3 Subjects

The subjects were a combination of computing science summer school students and post-graduates from the University of Glasgow. There were a total of 17 subjects; nine female and eight male. The mean age was 20. All had either normal or corrected-to-normal eyesight.

9.3.4 Procedure

Subjects were firstly asked to complete a personal details questionnaire in order to gain administrative information relevant to the experiment. They were then given a handout summarising the aim of the experiment and describing, in general terms, the task which they were to perform. Emphasis was placed on the fact that it was the interface that was being tested and not the subject. The next stage was a practice task. A reduced set of 3 file and 3 function icons was presented, and subjects were asked to perform certain tasks. The files and functions used were, however, different from those used in the experimental interface, and all icons were highly representational. The aim of this was to allow the subject to become familiar with the mechanics of the types of task they would be doing, without becoming familiar with any of the icons used within the experimental session.

Once it was clear that the subject understood the general principles behind interface operation, the experimental session commenced. Depending on what group they were in, subjects were either only set the 4 experimental tasks, or they were given 12 filler tasks, followed by the 4 experimental tasks, followed by a combination of 4 filler tasks and 4 experimental tasks, repeated 12 times. The time taken on each task was recorded from the point at which the experimenter finished reading out the task until the feedback informing the subject that the correct icon had been chosen appeared. Errors made on each task were also recorded.

9.4 Experimental results and discussion

9.4.1 Guessability

Performance on the first experimental trial was taken as representative of guessability.

	Time		Errors	
	Mean	SD	Mean	SD
Condition A	12.688	6.503	0.125	0.354
Condition B	22.747	9.077	1.750	1.282
Condition C	8.689	3.219	0.125	0.354
Condition D	35.548	13.147	2.500	1.512

Figure 9.3: Means and standard deviations for the results in the guessability condition

In the short condition, where subjects were presented only with the experimental tasks, it was expected that conditions where the correct icon was the highly representational one (rule compatible) would have significantly faster reaction times and fewer errors. Those conditions designed to be set compatible would however not show any significant increases in performance as a result of being set compatible. Using the Wilcoxen Signed Rank (WSR) test, conditions A and C, both containing the rule compatible icon as the target icon, were found to be significantly faster when compared with conditions B and D (d.f.=7, $p < 0.1$ and $p < 0.01$ respectively). Similar results were found in terms of the number of errors, with conditions A and C having significantly fewer errors than either of conditions B or D (d.f.=7, $p < 0.02$ and $p < 0.01$). A comparison of the conditions which were set compatible with those which were not was not found to be significant.

These results, therefore, support the majority of experimental work done on icon design. They highlight the fact that the degree of representation between the attribute and the object that it represents is probably one of the most important factors influencing the guessability of an icon.

9.4.2 Learnability

From pilot studies, the criterion considered to indicate learnability was the occurrence of two consecutive error-free trials and performance times below six seconds.

	Time		Errors	
	Mean	SD	Mean	SD
Condition A	2.75	1.04	1.00	0.00
Condition B	4.63	1.92	1.89	1.36
Condition C	6.00	3.11	5.44	3.71
Condition D	5.13	3.44	6.44	5.32

Figure 9.4: Means and standard deviations for the results in the learnability condition

In the longer condition, because subjects had experience with the filler tasks, it was expected that set compatibility would come into effect. Therefore in this case condition A would have the best results throughout followed by condition B (if set compatibility is stronger) or condition C (if rule compatibility is stronger). Condition D, where there is neither set nor rule compatibility was predicted to come last.

Results show that generally this hypothesis is supported. In terms of errors, condition A tasks are most learnable, B is second, C is third and D is last. In terms of time, results are similar, with the exception that condition D tasks appear slightly more learnable than condition C. Comparing condition A which is both rule and set compatible, with condition B, set compatible only, the difference in terms of time was found to be significant (d.f. = 6, $p< 0.1$). This would suggest that rule compatibility still has an effect on the level of learnability of an icon.

However, comparing conditions A with C where the presence or absence of set compatibility was varied, differences were found to be significant in terms of both time and errors (d.f.=7, $p < 0.02$ in terms of time, and d.f.=8, $p < 0.02$ in terms of errors). This therefore suggests that set compatibility is also an important factor in determining the level of learnability.

Since results imply that both set and rule compatibility have an influence on learnability, it may be interesting to consider which type of compatibility has the stronger influence. Looking at the results for condition A, it appears that having a combination of both rule and set compatibility gives significantly better performance results. However, if it is not possible for a particular icon to possess both properties, which property would be the most favourable one to possess ?

Interestingly, when comparing condition B, where there is only set compatibility, with condition C, where there is only rule compatibility, results indicate that condition B tasks are more quickly learned. Subjects took on average about 30% longer on condition C tasks to reach criterion performance levels in terms of time and almost twice as long in terms of errors (although these differences were not statistically significant).

This suggests that set compatibility on its own appears to have a stronger effect on learnability than rule compatibility. The presence or absence of set compatibility appears to have an according effect on the learnability of an icon. Varying rule compatibility when there is no set compatibility does not however, have the same effect. In fact no significant improvement in performance measures was found between condition C, where there was only rule compatibility, and condition D, where there was neither set nor rule compatibility. Conversely, significant differences were found between condition B and condition D tasks in terms of errors (d.f.=7, $p < 0.05$), indicating an effect for set compatibility even in the absence of rule compatibility. These results, therefore, seem to suggest that set compatibility has the stronger influence on learnability than rule compatibility.

9.4.3 Experienced User Performance

Performance on the last three trials was taken as representative of EUP.

	Time		Errors	
	Mean	SD	Mean	SD
Condition A	3.72	0.95	0.00	0.00
Condition B	3.97	1.77	0.00	0.00
Condition C	3.51	0.91	0.29	0.76
Condition D	4.92	4.34	0.86	1.46

Figure 9.5: Means and standard deviations for the results in the EUP condition

Unfortunately, no significant results were found for any condition, either in relation to time taken or the number of errors. This may be due to the nature of the experimental design. By the time users were expected to be reaching EUP no errors were being made and response times were roughly the same. In other words they had learnt what all the icons represented, whether rule compatible or not, and had also learned which test conditions were not set compatible and which ones were. Perhaps in non experimental situations, where user performance can be measured over days or weeks instead of one intense 45 minute period, then more accurate EUP results may have been gained.

9.5 Conclusion

To summarise the results from this experiment, rule compatibility improves usability during the guessability stage, however it becomes less essential for any performance benefit in the learnability stage. Set compatibility on the other hand only starts to influence usability by the learnability stage. In this experiment, by the EUP stage there was no performance advantage to be had by either set or rule compatibility.

The interesting questions to arise from this experiment are:

1. Why does rule compatibility lose performance advantage by the learnability stage?
2. Why is there no performance advantage for either form or consistency by the EUP stage?

For the first question the answer must be that either users are identifying the rule inconsistent icons by some strategy that is of an equivalent cognitive load to learning representational form. Perhaps because there were only 8 icons to learn the cognitive effort involved in learning the 'bad' iconic form was not too demanding. Thus by the end of the learnability stage, users were quite proficient at relating the form to the command. An alternative explanation is that perhaps learning the 'bad' icon form is more cognitively demanding, and therefore users are relying on a different attribute altogether in order to distinguish the appropriate icon. A final suggestion could be that by the learnability stage users no longer rely on form for any condition and have instead changed to rely on another attribute (e.g. position), which may allow faster performance. The results do not suggest a preference for any of these possible hypotheses.

For the second question, it is possible to suggest that by EUP level, rule compatibility was no longer important due to one of the several possible hypotheses just discussed. Set compatibility was however also no longer of importance (in this experiment at least), because there were really only 2 out of the 4 conditions where there was such inconsistency. Therefore with such a restricted interface, and an even more restricted number of set inconsistent occurrences it is likely that it was not too difficult for users to learn these inconsistencies.

A more general explanation for the results could be offered. Although set and rule compatibility have shown to be valid distinctions when referring to consistency, their importance appears to be entirely dependent on the context within which the individual icon is situated.

An icon is only rule compatible if it is consistent with users' expectations. However rule compatibility will be affected by the number of alternative possibilities available within the interface in which it is situated. In other words it can be influenced by the contrast set. For example, in the experiment, if both print icons shown in Figure 9.5 were displayed instead of one representational one and one abstract shape, then the chances of the user selecting the appropriate icon in the guessability stage would have been dramatically reduced. If both these icons had been present then the chances of the user selecting the correct icon may have been 50-50. In other words, the results were entirely dependent on the number of possible alternatives in the contrast set; i.e. context-dependent.

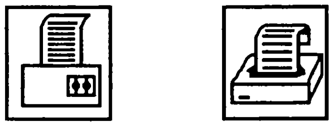


Figure 9.6: Two icons representing the command 'print'

Similarly, the only reason why set compatibility lost importance by the EUP stage was because there were so few inconsistent occurrences that it required little cognitive effort on the users' part to learn these inconstancies. Perhaps in a more complex interface, set compatibility may remain important in the EUP stage. Its importance would, however, still be dependent on the degree of set incompatibility present in the interface as a whole. For example, in an interface using 100 icons, 98 of the icons represent exclusive tasks and two represent the same task, 'print'. The first print icon applies to all types of file (i.e. spreadsheet, word-processing, graphics, etc.), but the second one only applies to painting files. It is likely that users will learn this inconsistency easily enough. If however the first print icon applies to 7 different types of file, and the second icon applies to 7 also, then the cognitive load involved in learning this inconsistency may be too great. In order to get some measure of this cognitive load some secondary task test should in future experiments be employed.

Another scenario could be where all commands in the interface are represented by two possible icons, thus making a total of 50 possible commands. Each command may follow a similar pattern to the first example given; all tasks apply to one icon for each command, except one, which requires the user to select the other icon. It will be even more complicated if the exception to the rule varies between commands. Therefore for the 'print' command the alternative icon should be selected when printing a painting file, whereas for the 'save' command the alternative icon should be selected when the user wishes to save a word-processing file. In such situations it would be predicted that such inconsistencies would have a detrimental effect on performance. Therefore what is being presented is the argument that set incompatibility will only be detrimental to performance if it requires too heavy a cognitive effort to learn.

A large number of different rules could mean a large number of different icons representing the one command; in other words a large contrast set. However, as suggested in the above examples, this is not necessarily the case. Set inconsistency could be just as detrimental to performance with only 2 icons in the contrast set. Conversely, if a rule is consistent over time then the whole interface could contain almost identical icons and performance would not be affected, since users would be sure that only one icon is relevant consistently. Set compatibility, however, does affect the length of time in which the contrast set is reduced. The more inconsistent the rules are, the longer the learning process before the set of possible icons involved in the user's decision can be confidently reduced.

Therefore by re-analysing this experiment using the new findings established from the previous experiments, it is possible to reiterate several important points:

1. A truer picture of an icon's usability can only be achieved by analysing performance over an extended period of time. It is apparent from both experiments that the attributes used by people in the guessability stage to recognise particular icons are not necessarily the same ones as used in the learnability stage. The emphasis may shift to attributes that offer a greater performance advantage.

2. Interface context is extremely important. The size of contrast set the user is required to chose from when selecting an icon can greatly influence performance, irrespective of how compatible the actual icon attributes may be with previous knowledge and expectation.

3. If the experimental design had looked at the icons in isolation any effects of set compatibility / incompatibility would have been lost. This therefore supports the argument at there are many attributes and issues affecting usability that are lost when icons are studied in isolation.

Criticisms of this experiment, however, are that in real life it is possible to expect that icons could be rule incompatible. Therefore are many occasions where it is not possible to use attributes that draw upon a user's previous knowledge or experience simply because the concept being represented is far too abstract to represent in such a concrete manner. Accordingly, many designers have attempted to use metaphor (e.g. the form of a piggy bank to represent 'save'), which may appear evident to them, but which appear 'unnatural' to the user. Other times, as in this experiment, designers may prefer to use abstract icons which may appear less like the correct icon to the user than other possible alternatives within the interface.

However, in real-life interfaces designers would be unlikely to 'randomly' use one rule for some tasks and another rule for other similar tasks. Generally there should be some reason behind this. The problem occurs because the reason behind the design is not evident to the user. If the icon attributes were designed so that the reason was also communicated, then the icons used for the two different rules would be more distinguishable, the extra information available from inspection as opposed to 'trial and error', and perhaps the disruption caused by set incompatibility reduced. Therefore by testing an interface for set incompatibilities the designer can compensate by either removing them or supplying additional information to communicate why they occur.

Whether this is indeed the case is perhaps an issue suitable for further study. It may be possible that as Gibson (1979) suggests there is always a vast amount of information present in an environment - whether it is used or not depends on our ability to see it. Perhaps if users assume all tasks to use the same rule for that particular command then they will ignore the information communicated by the interface.

10. Conclusions

The research performed for this thesis differs from previous research in that it specifically considers the affect on user performance when icons are placed into an interface context, and when users are able to interact with the icons over extended periods of time. The main achievement of this thesis is that the experimental results question the assumptions about icon design that have been established through previous research (e.g. Rogers, 1986; Gittins, 1986; Jordan et al., 1991). What results in this thesis suggest is that previous findings should not be used as generalisations or fixed guidelines for the design of icons but should instead be looked upon as rules-of-thumb. The central conclusions to this thesis, each undermining previous assumptions, are described in the next four sections.

10.1 Current categorisations of icon attributes do not necessarily identify the attribute the user selects to recognise the icon in the interface.

The categorisations for icon design currently in existence (e.g. Gaver, 1986; Gittins, 1986; Blattner at al., 1989; Rogers, 1989a; Webb, Sorenson and Lyons, 1989; Maissel ,1990) present the designer with guidelines of how to create attributes which will convey the icon's underlying meaning to the user. What this thesis has however stressed is that once the icon is placed into the interface, there is no means of predicting that this will be the attribute users will rely on in order to recognise the icon for the approapriate command. Results from the pre-emptive move data appear to suggest that even in conditions where users were presented with representational form they would use position to identify some of the icons. Similarly, users presented with abstract icon forms, and therefore predicted to use position to identify the appropriate icon for the command, appeared to identify some icons by their abstract form. In fact, the attribute users relied upon for each icon was not generalisable between subjects, thus suggesting that the attributes people use to identify each icon are almost entirely unpredictable and completely idiosyncratic.

The conclusions reached by this thesis are that the likelihood of an attribute being relied upon can, to some extent, be determined by the interface context. It suggests that the more distinguishable the attribute is from the rest of the attributes present on the interface (i.e. the contrast set), the more likely subjects will be to rely on this attribute. Therefore if there are two icons displaying subtle variations of the representational form of a printer, then it can be predicted that the user will be unlikely to rely on form when searching for the appropraire icon for either command.

According to Kaptelinin (1993), those features considered to be particularly discriminable are global features (e.g. colour size, position) as opposed to local features (e.g. the icon form). If the thesis findings comply with Kaptelinin's theory then once an initial strive at understanding was made (by focusing on form), users would then be predicted to rely on global features. However, the pre-emptive move data, even at EUP level, indicated that users in all conditions were still relying upon form for some icons, suggesting that the move from local to global feature processing is not necessarily uniform. Therefore, results imply that Kaptelinin's theory is not necessarily generalisable to all icons.

10.2 Learning and changes in attribute use occur throughout the learning curve as defined by Jordan et al (1991) - and beyond.

This is an extremely important finding. Results strongly suggest that the attribute considered to be important in the guessability stage of performance is not necessarily the attribute used throughout the rest of the usability curve. There is a distinct contrast between the "effort after meaning" made by the user in the guessability stage, and demand for a highly discriminable attribute with a small cognitive load in the learnability and EUP stages. Therefore as the experiments have suggested, icons which contain attributes which appear to offer no meaning to the user in the guessability stage perform as well as icons possessing these meaningful attributes at EUP. Although icon-oriented interfaces tend to be aimed at the computer novice, by restricting study of icon design to how to attain meaningful icons, researchers are ignoring the potential design issues apparent to the majority of the usability curve.

However results have shown that learning and fine tuning of procedures to recognise particular icons with minimum effort and cognitive load continue to occur after experienced user performance level has been achieved using the classical performance measures of time and error. This may suggest that Jordan et al.'s (1991) model of usability may be too simplistic. With the ability to capture continued learning through measures such as the number of pre-emptive moves a more detailed picture may be achieved.

The results may therefore question whether EUP (as defined by Jordan et al.) does actually exist, or whether it is simply a term used to describe the leveling off of traditional measures of performance.

10.3 Users learn more than just the minimum number of attributes required to identify an icon.

This is related to the previous point. In order to switch from one attribute to another, users must be aware of their presence on the interface. Results suggest that when faced with multiple properties users do not simply learn one, but perhaps through incidental learning learn at least two, and perhaps several. Experimental evidence suggests that at changeover there appeared to be a disruption followed by a quick recovery. Therefore, if learning was required at the changeover point then it occurred far faster than initial learning of the original attribute. If the original attribute was considered to be the attribute that was easiest to learn then why was learning of an inferior attribute actually faster? Obviously practice effects are going to have some influence, but comparison between initial performance times and changeover times are, in some conditions, so drastic that this is not a valid explanation. Therefore, as was suggested in Chapter 8, it is likely that users were aware of the other attributes that were present and therefore at changeover were shifting to an attribute that they had previously learned to some extent. The degree to which they had learned this attribute is, however, unknown. It can only be suggested that some degree of 'polishing' was required before the subject was able to perform at an optimal performance rate again.

The important point is that attributes a subject relies upon may often not be those initially categorised as attributes that are beneficial to learn. For example, evidence from the experimental results suggest that some people still learned abstract icon form, even though it appeared not to reduce cognitive load or have any apparent performance benefits. This finding contradicts the view held by Draper (1991) and others who suggested that users economise on the amount of information that they will learn, and therefore, will only learn those attributes that offer smaller cognitive load or increased performance.

10.4 Evaluation paradigms do not evaluate icons within a suitable context.

In order to establish good icon design, potential icons need to be evaluated. Several paradigms can be identified in the previous research with the aim of isolating the most appropriate icon for the command. Several of these are discussed with the emphasis being on why, in the light of the experimental results, they should be considered invalid.

a. Appropriateness ranking

For each individual command subjects are presented with a set of alternative symbols and asked to sort them into order of appropriateness for that command. Frequency tables are then compiled. Rogers and Oborne (1987) used this paradigm in the second part of an experiment. The initial experiment required a group of subjects to draw pictures which they thought represented the meaning of a set of verbs. These drawings were then used as the set of symbols in the appropriateness ranking experiment.

b. Comprehension test

This is sometimes also called the "recognition test" and usually involves presenting users with a symbol and asking them to write down what they think it means. The responses are then categorised as correct, plausibly incorrect, other and don't know. Researchers who have used this paradigm include Brainard et al (1961), Howell and Fuchs (1968), and Easterby and Zwaga (1976).

c. Matching test

Here the respondent is asked to select for a given referent one symbol from a number of symbols. In the matching test the score of the symbol is based on the number of times it is selected correctly from the group of symbols presented to the respondents. A score of incorrect choice of symbol is also recorded. Zwaga and Boersema (1983) used this paradigm when evaluating a set of Netherlands railway symbols.

A less stringent version of this test has also been used (Milner, 1987; Rogers 1986). In this case functions were read or handed out to a group of subjects who then marked whichever icon they thought best matched the description (out of a full set of icons). Each icon could only be picked once. This paradigm has its own design problems in that subjects were unable to reselect any of the icons if they had selected incorrectly for an earlier command they may be forced to choose from an inappropriate selection of icons for a later command.

d. Timed search and select paradigm

This paradigm is very similar to the matching test. A set of icons is placed within a matrix. The subject is presented with a command and asked to select the appropriate icon from the matrix. Reaction times and error rates are recorded. Researchers using this paradigm include Arend, Muthig and Wandmacher (1987), Browne and Stammers (1987), George (1988), Bewely, Roberts, Schroit and Verplank (1983). Criticism of this paradigm has been that it places importance on the discriminability of the icon attributes rather than the importance of the communicating meaning. As suggested in the previous section, however, discriminability and interpretation are perhaps less unrelated than might have been presumed by previous research.

The main flaw in methodologies a and b is that both attempt to evaluate the appropriate icon for the command by considering each of the set in isolation to the rest of the interface. Such methodologies could perhaps be used as a way of establishing a basis for further research, but should in no way be used as sole justification for selecting one particular icon over another. The ISO (1981)[2] recommend that procedures a, b and c are used when performing icon evaluation; performed in the order that they appear here. Therefore the appropriateness ranking and the comprehension test narrow down the number of alternative icons for each command, while the matching test analyses these design decisions in context. Unfortunately, very few researchers have performed an evaluation involving all three stages, the exception being Maissel (1990).

---

2 Note that these recommendations are for building information signs, not computer icons.

However, even when the paradigm includes surrounding the target icon with the icons with which it will be situated in the interface, the paradigm is often still invalid since many researchers failed to study the icon within the interface itself. Often the paradigm included the use of enlarged pictures of icons (Bewely, Roberts, Schroit and Verplank (early stages only), 1983; Browne and Stammers, 1987; George, 1988) or hand drawn images (Rogers, 1986; Milner, 1987). Others simply transferred the icons from the interface onto paper in order to perform paper and pencil tests (Maissel, 1990). By doing this many attributes apparent within the interface are lost (e.g. relative position), thus reducing the possibility that some meaningful information could be presented by situating the icon within the interface as a whole.

The suggestion is, therefore, that experimental methodologies used as evaluation techniques in the past should be used in the early stages of evaluation, but before any final design decisions are made the icon should be assessed within the computer generated interface. The main point to be made is that even if these design methodologies conclude that one icon design is 'perfect', every time the interface is modified in some way it is likely that the icon's effectiveness would have to be re-assessed.

A further point to consider is that all experimental paradigms described use error rates and/or reaction times as their dependant variables. However, the experiments have shown that learning continues after error rates and reaction times have reached optimal levels, therefore these paradigms are not suitable as measures for experienced levels of perfomance.

Finally, just as there is evidence that the effects of interface context have been ignored in the design of icons, there is also evidence that some researchers, at least, have been aware of its influence. For example, Easterby (1970, pp. 150) suggests:

> The important inference here is that it is of little use deciding on the visual qualities of the symbol until the context in which they are interpreted is agreed, together with adequate definition of the required symbolic referents and the interrelations between these individual concepts.

Similarly, Cahill (1975, pp. 378) states:

> No symbol should be conceived in isolation, for it will not, and cannot be expected to stand alone. The signification a symbol conveys is embedded in the entire context in which it will be used...Context situations must therefore be carefully analysed and tested for their contributions to the information carried by particular symbols intended for display within these contexts.

Unfortunately so far, the number of researchers who have been aware of the importance of contextual information are far outnumbered by those who simply ignore its influence. The ultimate aim of this thesis is to suggest that icon research should no longer be concerned with measuring the degree to which a particular representational form of a solitary icon can increase performance within the guessability stage of usability. Instead research should be widened to consider how attributes in the interface interact, and how this interaction may vary over time as user knowledge of the interface increases. According to Gibson (1966) "the attempt to analyse how a picture conveys information is a necessary but highly ambitious undertaking". This sentiment applies equally to the iconic interface. The results from this thesis strongly suggest that only by looking at the interface as a whole, as opposed to just a small fraction of it, will any significant results be achieved.

10.5 Future Work

The findings from this thesis have stimulated a plethora of new research questions. The main result from this work has been to emphasise the importance of attribute discriminability in relation to the rest of the interface. The main question however is how important is discriminability as a design issue? Is it more important than consistency, or the ability to represent some essence of the underlying object to the user? Does its importance increase or diminish as more attributes are presented and the number of icons in the interface increases (as the interface in general becomes more complex). These questions have not been addressed by the experiments performed so far.

There are also many questions raised by considering performance throughout the learning curve. As suggested, we could ask does EUP exist? Do users ever reach a plateau, or do they continue fine tuning their performance? Do users fine tune their performance to such an extent and then switch off to any changes in the environment, or do they continue to monitor change? How can we measure EUP beyond the stage where reaction time and error rates reach optimal performance levels? Pre-emptive move data may be an important step towards this, however, a new experimental paradigm may also be useful. A suggested paradigm could be one where subjects were presented with tasks which competed with icon selection for the user's cognitive attention. Measuring performance levels (e.g. reaction time, errors, pre-emptive move data) once the additional load was introduced may give some idea just how much of the user's cognitive attention was actually being used to identify the appropriate icon.

It may also be interesting to look at the issue of re-usability (Jordan et al, 1991) and other ideas related to extended user performance. Re-usability is the measure of how usable an icon is when the user has interacted with the interface for a period of time, but has then taken an extended break from use. Will users rely on the same attributes that they had relied on when they had reached experienced user level, or will they return to relying on attributes that offer users meaning? Obviously the length of time away from the interface is going to have an important influence on the results. Also, the experiences of the user during this period may have an additional effect on results. The issue of re-usability is an important one, since in real life many software packages are used only intermittently. A series of experiments performed over a much longer time period may begin to answer these questions.

Another issue, which may be important to consider, is which attribute are subjects more likely to rely upon when there are competing, equally discriminable attributes available? For example, if given the choice between a highly discriminable form and a highly discriminable position which would subjects most likely rely upon? From the experimental results it would be predicted that many subjects would rely on position if using a mouse due to the increased performance advantage, but how about discriminable colour versus discriminable form? Future work should therefore look at alternative attributes, since it could be argued that perhaps the findings here are unique to form and position.

10.6 Conclusions for designers

The ultimate aims of any HCI research should be to enlighten designers and be of practical benefit in design. Firm conclusions are perhaps impossible to draw from the results of the experiments in this thesis. However, it is possible to argue that several conditions may be suggested as rules-of-thumb for designers. Furthermore, the literature review performed in Chapter 2 makes several points apparent. The practical design points that can be taken from this thesis are, therefore, as follows:

1. Suggestions when designing guessable attributes.

To design an easily guessable icon the only clear assumption that remains unchallenged by empirical investigation is that the more direct the mapping between the surface form of the icon and its underlying referent, the more efficient the icon; in other words, the easier it will be to understand, learn and use (Hemenway, 1982; Gaver, 1986; Rogers, 1988). The implication therefore is that designers should attempt, wherever possible, to design icons that are as close as possible to the real thing thus ensuring that the sign-referent mapping is nearest the representational endpoint on the continuous scale of abstraction.

Designers should however remember that form should not be the only attribute to be considered. There are a number of other attributes such as position, shape and colour that could carry representational meaning.

Tools are generally the easiest system feature to design guessable iconic forms for. Tools that are particularly well suited to being displayed iconically are those that support the numerous drawing and painting techniques necessary for graphical manipulative drawing processes. Their success lies in the fact that the tools people would use in an equivalent paper-based drawing task are tangible and concrete and therefore easily represented as icons. However, they are only tangible because, as Kohlers (1969) suggests, they are representational to people who know the iconic code (e.g., the mapping between how something is represented and what it represents)

This is not to say, however, that intangible objects or functions can never have guessable attributes. Through convention, many iconic forms can become established, and therefore guessable. The example presented previously was that of inverse video to represent when an icon has been selected. Although not a natural phenomena in any culture, it has been established through convention, and therefore would appear to be obvious and natural to many computer users.

Perhaps the most important perspective for designers to consider when designing icons is the point made by Hutchins, Hollan and Norman (1985), that:

> If we restrict ourselves to only build an interface that allows us to do things that we can already do and in ways we already think, we will miss the most exciting potential of new technology: to provide new ways to think and to interact with a domain.

Unfortunately, currently one of the main drawbacks is that what may look good on paper may not work so well when transferred onto the screen. As technology improves and becomes cheaper, current limitations (e.g. resolution) should be removed, allowing the designer greater scope in the design used.

2. Be aware of the importance of interface context in design.

Designers should be aware that interface context, as defined in this thesis, exerts a strong influence on any icon's ability to communicate its meaning to the user. As has been consistently stressed in this thesis, this point has often been ignored by researchers in the field of icon design. Designers therefore should be aware of this when reading icon research, and should question whether results would have been so valid if the icons had been tested in an environment that contained more than one icon.

3. Using interface context to help communicate an icon's meaning.

The surrounding context may reduce the guessable icon's ability to communicate an icon's meaning. Conversely, it may also increase the likelihood that a non-guessable attribute will be understood. Most previous research has seen the inability to design icons in a form representational to the real world as a reason for not using iconic interfaces. Few researchers have considered how communication could be aided by presenting icons within a suitable context, allowing contextual relationships to become apparent.

It is conceivable that once an icon is placed within a context new attributes may become apparent which may communicate the intended meaning. Many of the Gestalt graphic design principles are perhaps important to consider. Meaning can be established by applying the rules of proximity, similarity, good continuation, closure, and symmetry to the design of icon attributes. The main design principle being suggested is that increased information should allow the user to make a more informed judgment as to what any particular icon may mean, and ultimately result in a more usable interface.

4. Designers should include discriminability as a design feature.

The main design guidelines to be extracted from this thesis are that discriminability is highly important in icon design, and that it is entirely context dependent. Therefore, designers should be aware that by changing one icon attribute, that icon's relationship with all other icons in the interface may also be altered.

Findings suggest that people will rely on the characteristic that is most easy to discriminate from a set of characteristics. In other words when faced with a forced choice, users rely on the attribute contained in the icon which has the greatest contrast to other icons within that set. Therefore, designers should be aware that it is impossible to say that any particular property is going to be **the** property that people rely on until you know the degree to which the property value varies from the value of that property for other commands. Some contrasts may be particularly obscure and idiosyncratic.

In certain cases it may simply be the attribute that appears most discriminable solely to that user that will be the one he or she will rely upon. This may not necessarily be the attribute judged by the designer to be the most discriminable. This complies with the experimental finding that what attributes people rely upon differs both between people AND between icons.

The main advantage of increased discriminability appears to be smaller cognitive load. In turn a smaller cognitive load is likely to be equated to mean a more usable interface. By ensuring that an icon contains at least one discriminable attribute, the contrast set (the number of possible options) will be reduced for any particular command and therefore the faster the user will recognise the appropriate icon. In other words, less time will be spent comparing icons before the selection is made, and consequently there will be less of a cognitive load on the user. The theory would therefore support the idea that designers should aim to create attributes that are as distinct from the surrounding attributes as possible. If this design principle is complied with the theory would predict that the user is increasingly likely to rely upon that attribute to identify the icon, and will be able to perform the task of recognition significantly faster and with greater ease.

5. Designers should implement a design that is suitable for all stages of usability.

Many abstract attributes (e.g. form, location, colour, shape) will not be guessable, but it is possible that their meaning may be extremely learnable. So much emphasis has been placed on icons being usable at the guessability stage, however, guessability only appears to apply to a small percentage of the total learnability curve By ignoring any non-guessable attributes, designers should be aware that they are potentially ignoring design issues that may have greater lasting consequences over the majority of the learning curve.

This thesis has suggested that by considering the entire learning curve when designing the interface it becomes apparent that the commonly held assumption, that a 'good' icon is a guessable icon, is far too restrictive to be regarded as a valid design principle. The second assertion is that a design applying to usability throughout the learning curve may result in a more usable interface overall.

Results have suggested that people's reliance on particular attributes evolves over time. Therefore they may initially rely on form, but as they learn position, they shift to rely on this attribute instead.

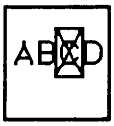
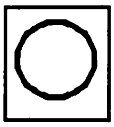
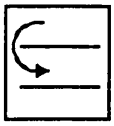
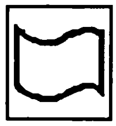
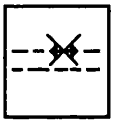
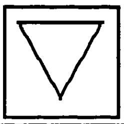
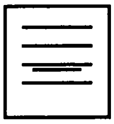
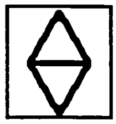
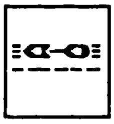
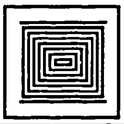
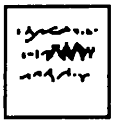
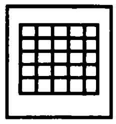
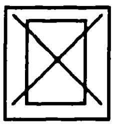
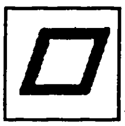
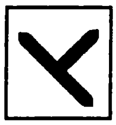
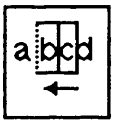
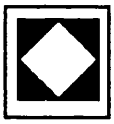
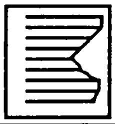
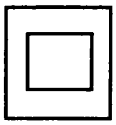
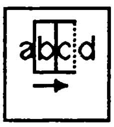
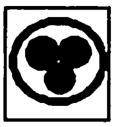
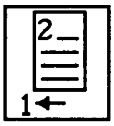
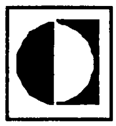
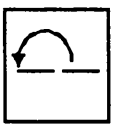
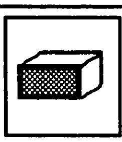
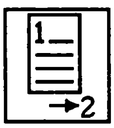
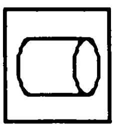
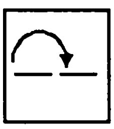
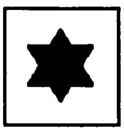
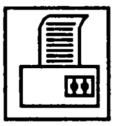
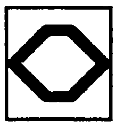
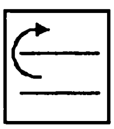
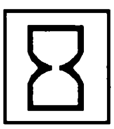
To summarise, there are perhaps three phases to how attributes are used over the learning curve:

1. Use the attribute that allows you to get the job done (performance is significantly faster if a guessable attribute is present)
2. Learn other attributes perhaps through incidental learning or because they offer some level of performance advantage
3. Forget the previously used attribute as the user becomes more experienced in using the attribute offering the performance advantage

Finally, designers should be aware that people continue to learn and rely on other attributes even when performance measures appear to have reached asymptote - suggesting that perhaps learning never ends. This also suggests that reaction time and error rate are not good measures to capture learning in the experienced user level of performance and therefore more elaborate or subtle means of performance capture should be devised.

Appendix 1: Materials used for experiments 1-6

1. Icon sets used

Command	Near Icon Set	Far Icon Set	Command	Near Icon Set	Far Icon Set
Delete One Character			Cursor One Line Down		
Delete One Word			Underline		
Delete One Line			Bold		
Delete the Whole Text			Right Justified		
Cursor One Character Left			Not Right Justified		
Cursor One Character Right			Page Up		
Cursor One Word Left			Page Down		
Cursor One Word Right			Print		
Cursor One Line Up					

Near and Far are the descriptions originally given by Blankenberger and Hahn (1991) to describe representational and abstract form respectively.

Icon frame size: 27X29 mm.

2. Experimental Protocol

1. Give subject the "Overview of the experiment" to read.

Answer any general questions, but not any pertaining to the meanings of particular icons.

2. Give basic instructions session.

a) Say: "In order to perform the tasks that you will be set during the experiment you will need to known the general principles of how such tasks are performed. I am going to demonstrate this to you, and then let you run through a short practice session, until you feel competent at it."

b) Perform a sample task on the demonstration screen, then let the subject try the short training display. While demonstrating the task, remember to **say:** " You will notice that there is a short tone half a second before the next trial begins...There is also a tone emitted whenever the wrong icon is selected".

c) At the end of this session **say:** "Do you feel confident performing the task ?" If they are, then **say:** "That concludes the basic instruction session, we shall soon be ready to start the experimental session." Otherwise, allow a few more practice trials and then repeat this stage.

3. Start the experimental session.

a) Load the experimental interface. **Say:** "We are now going to start the experimental session. The procedure will be the same as the training session. If you have any difficulties or get anything wrong keep trying. Remember, however, that it is the word processor under test not you, so don't worry if you have difficulties, or get some things wrong."

b) Note down times and errors on the experimenter's sheet.

c) If subjects are having problems with the physical mechanics of performing tasks help them with this, but don't give any assistance if asked about icons' functions.

4. Questionnaire

Once users have completed the experimental session, they are given a questionnaire to complete. **Say:** "Could you now please complete this questionnaire. All the answers will aid my analysis of the experiment and will be kept in the strictest of confidence".

5. Conclude experiment.

When all tasks are complete **say:** "That concludes the experiment, thank you very much indeed for your co-operation".

3. Experimental Overview

The purpose of this experiment is to investigate the ease of use, or usability, of a simulated iconic interface. The interface simulates a word processing package, designed to enable a range of manipulations to be performed on a word-processed document.

During the experimental session you will be set a series of tasks to perform, using the interface under test. The tasks will be presented randomly by the computer. You will be required to select the appropriate icon for the task.

Remember, it is the interface that is being tested, not you, so don't worry if you have any difficulties, or cannot complete a task.

Do you have any questions ?

4. Questionnaire

PART A : PERSONAL DETAILS

Subject No:

Gender: M / F Date of birth:

Do you have either good or corrected-to-good vision?

Have you used a mouse before?

Have you used any word processing packages before?

Which ones have you used?

How long ago was the first time that you used one?

Since then how regularly, on average, did you use one?

How regularly do you use them now?

Have you used a graphical or iconic user interface (i.e. interfaces that use icons)before?

Which ones have you used?

How long ago was the first time that you used one?

Since then how regularly, on average, did you use them?

How regularly do you use them now?

PART B: INTERFACE DETAILS **Subject No:**

(please use the back of this sheet of paper for any additional comments you feel you want to make)

What strategies (if any) did you use for selecting the appropriate icons for the task?

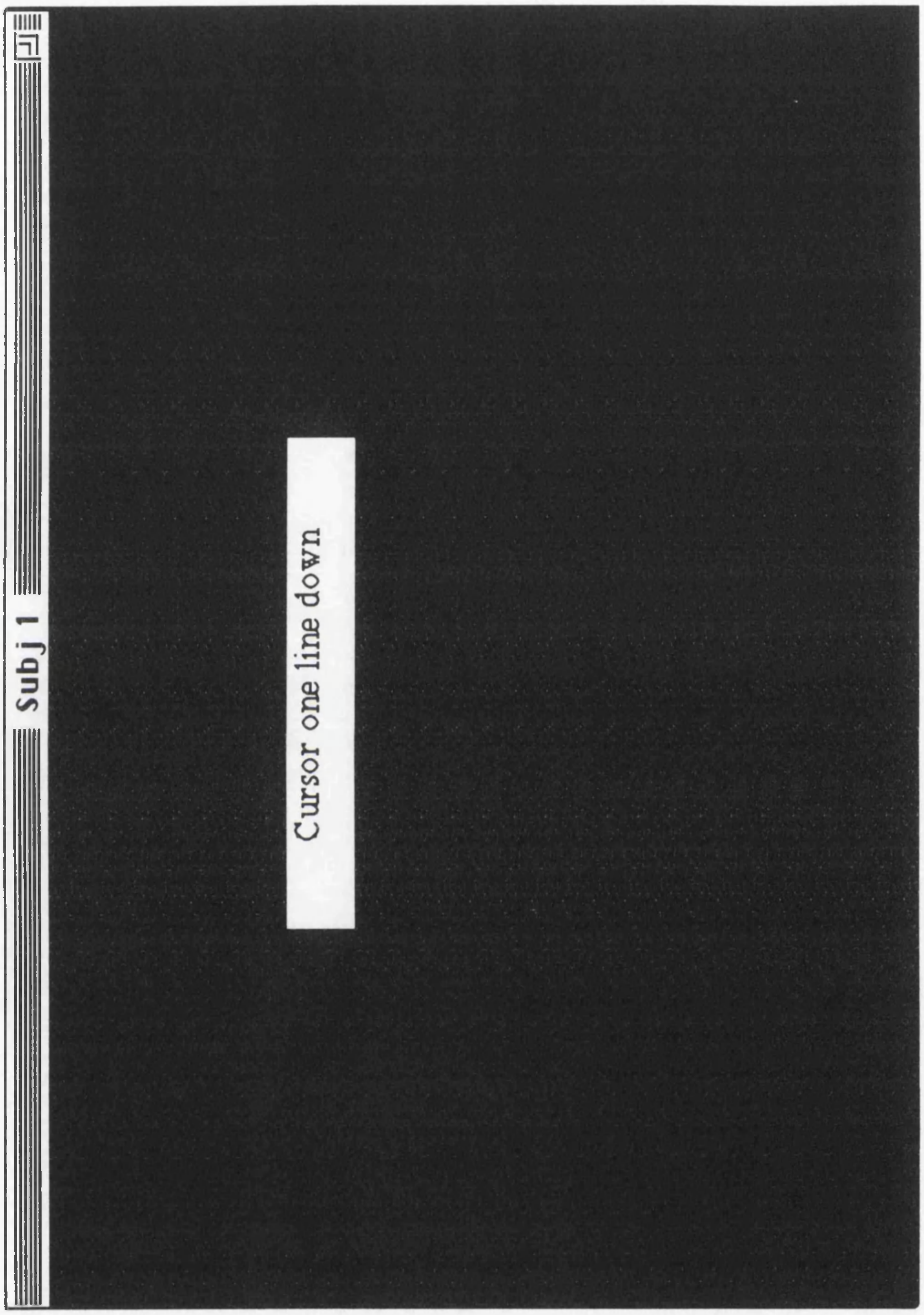
Did your notice a change in the interface half way through the experiment? If so, did you notice it immediately?

Were you surprised by the change?

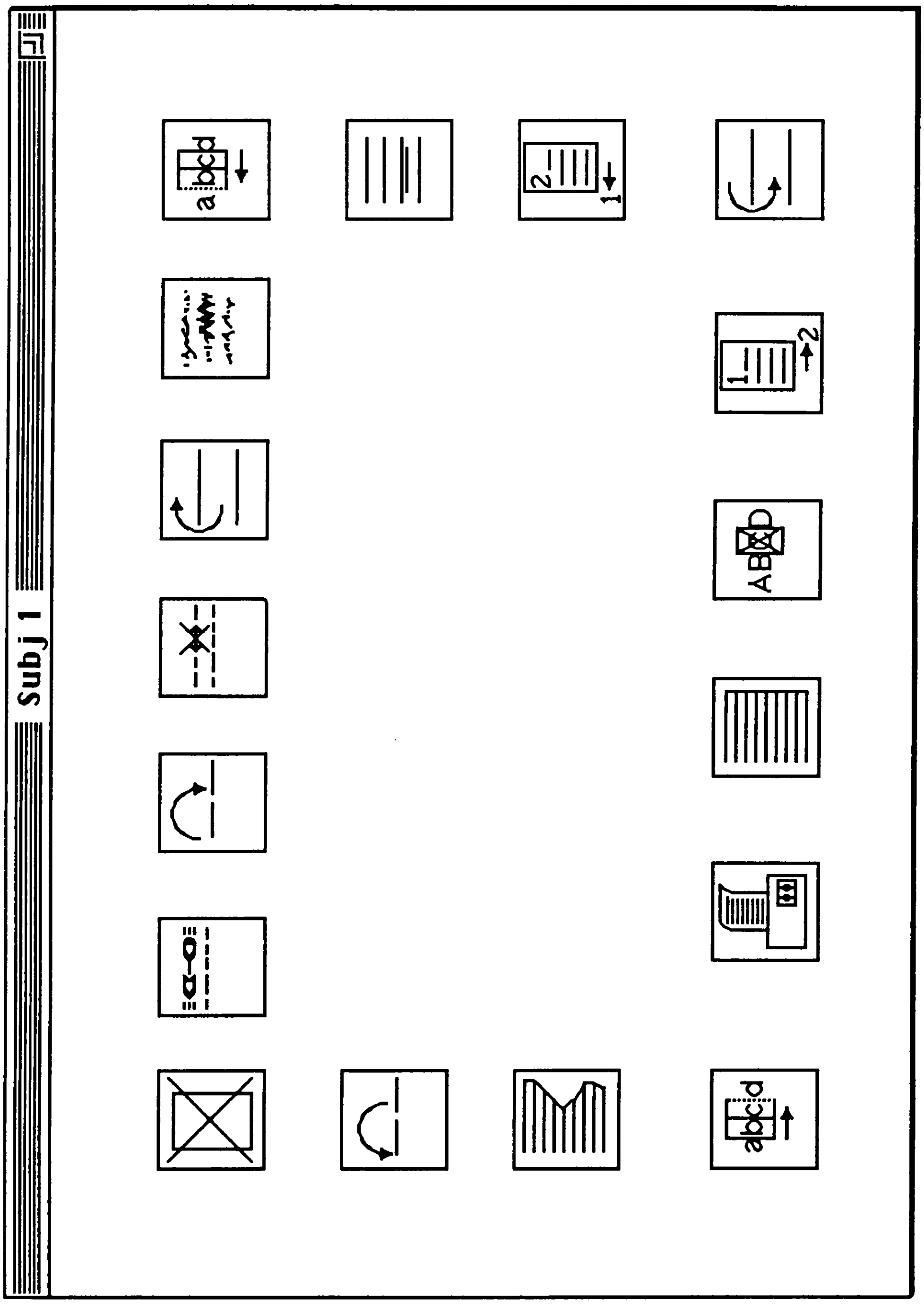
How greatly do you feel the change disrupted your performance (e.g. large disruption on one trial only, or consistent disruption over several trials, etc.)?

Do you feel that the change in the interface caused a change in your strategy for selecting the appropriate icon?

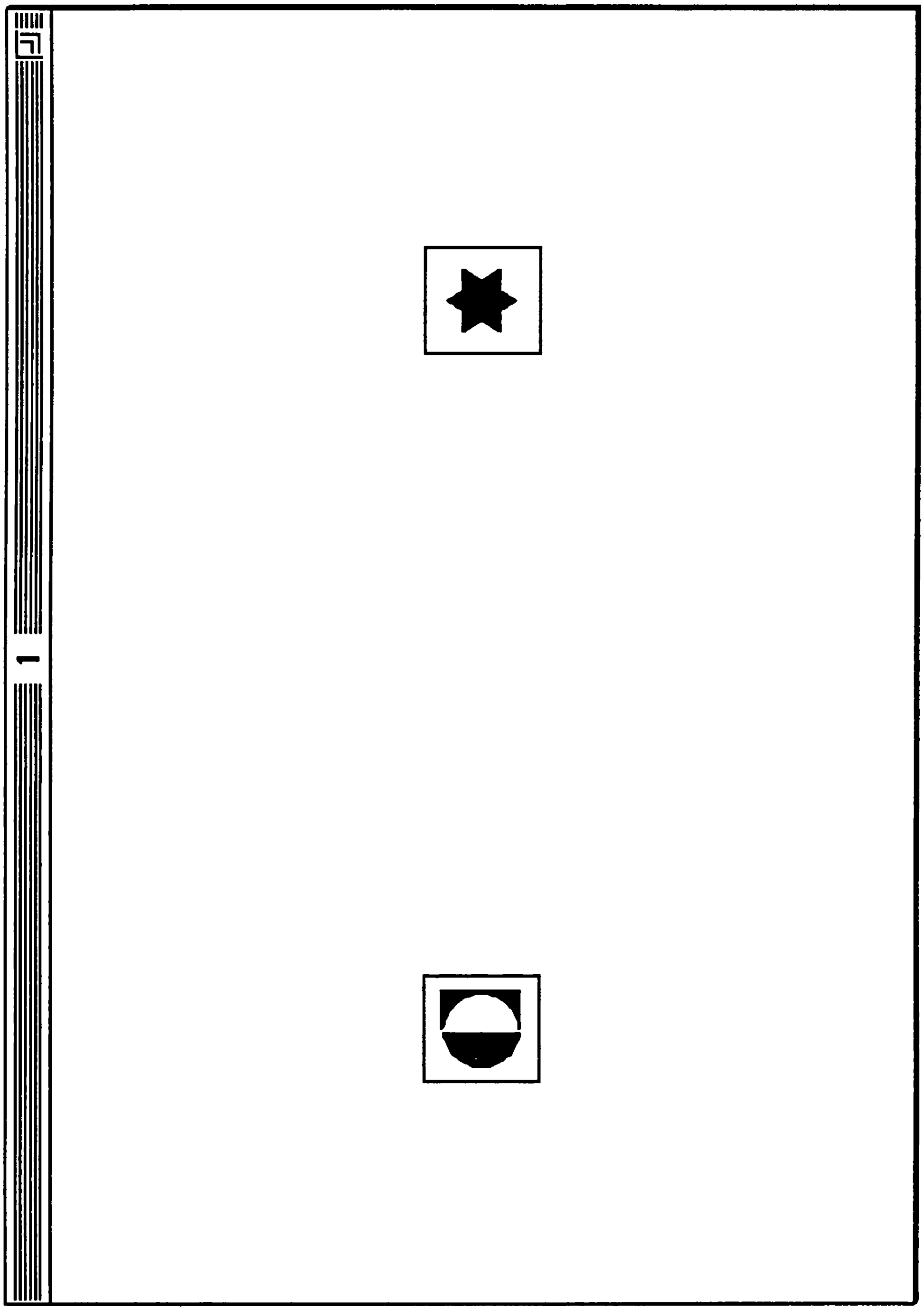
5. Interfaces



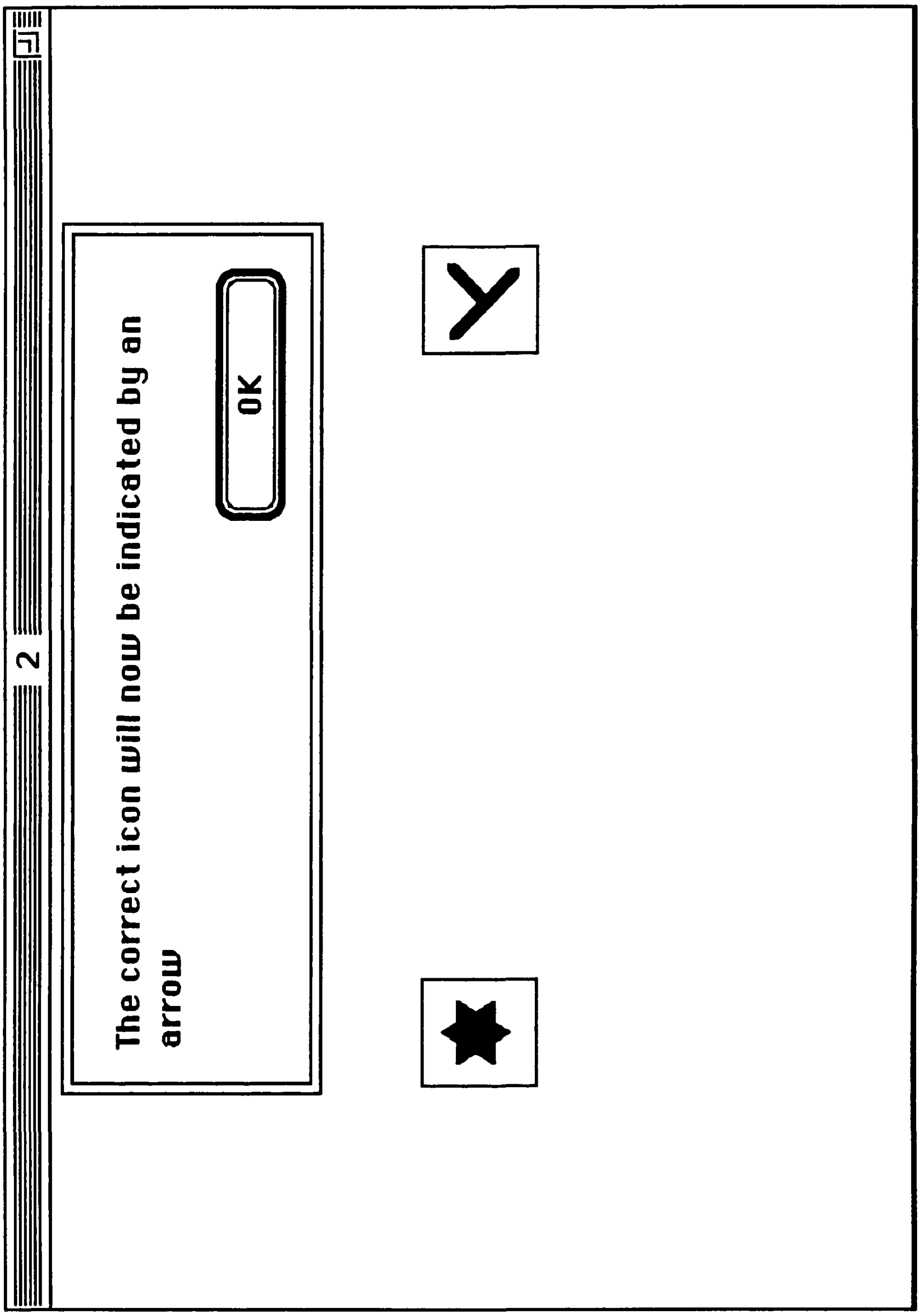
Command card used in experiment 1. Users were presented with this card for 1.5 seconds before the interface card appeared, therefore not requiring the centre button (as shown in Figure 4.3).



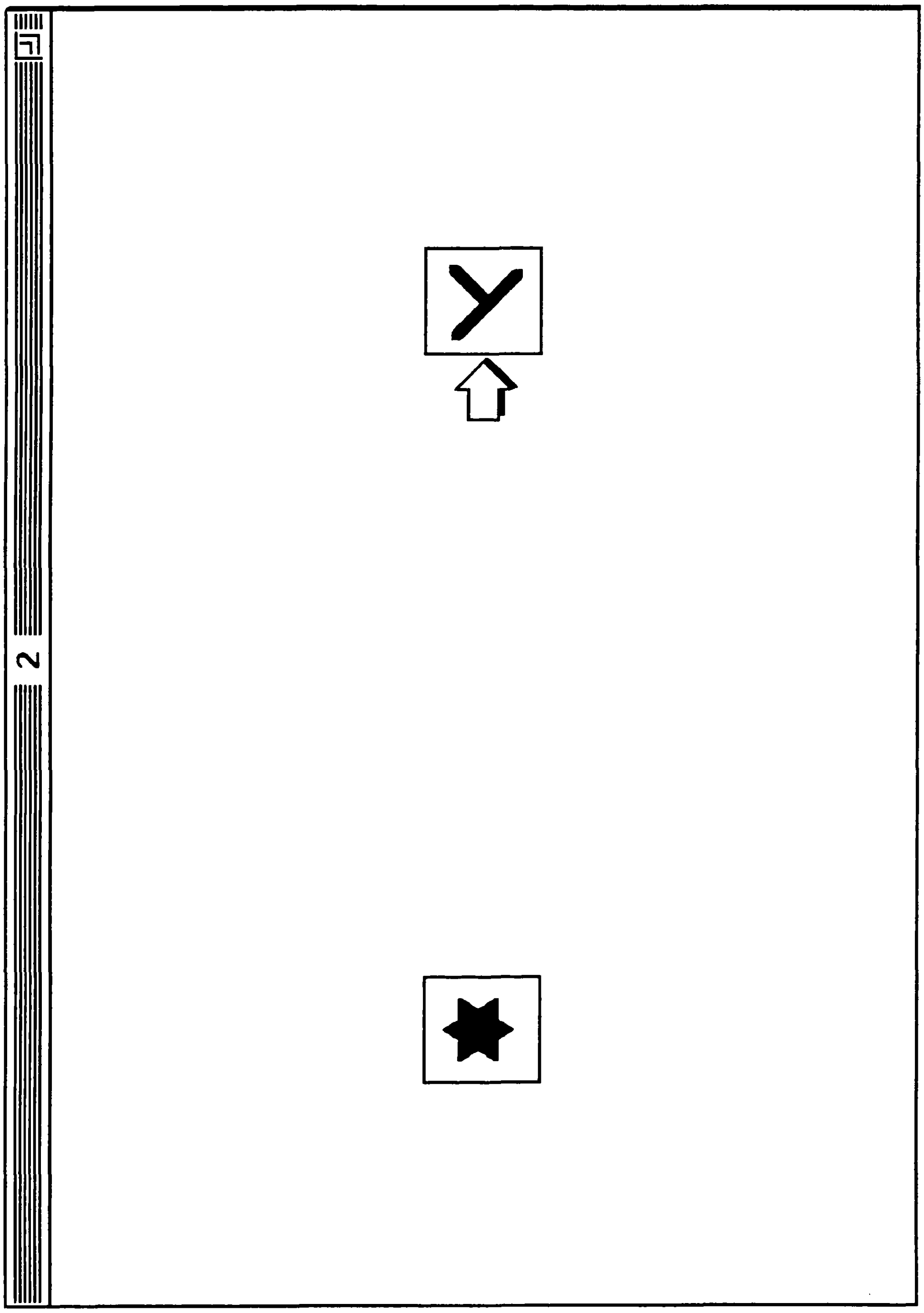
An example of the representational form interface for experiment 1. The rectangular layout was not re-used since the icons were not equi-distant from the centre button on the command card and therefore differences in reaction times could be attributed to differences in the distance the mouse was required to travel.



An example of the abstract form interface used in experiments 2 to 5. The icons are equi-distant to the centre button on the command card.

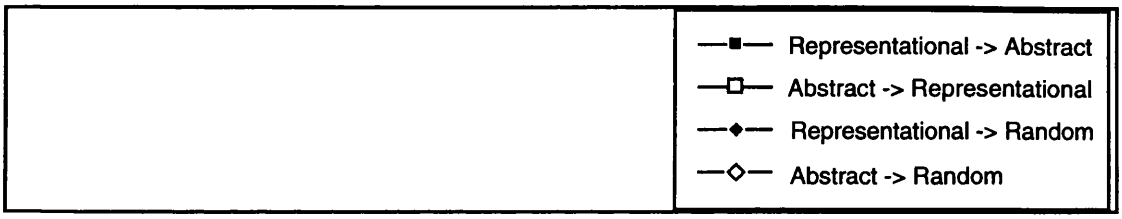


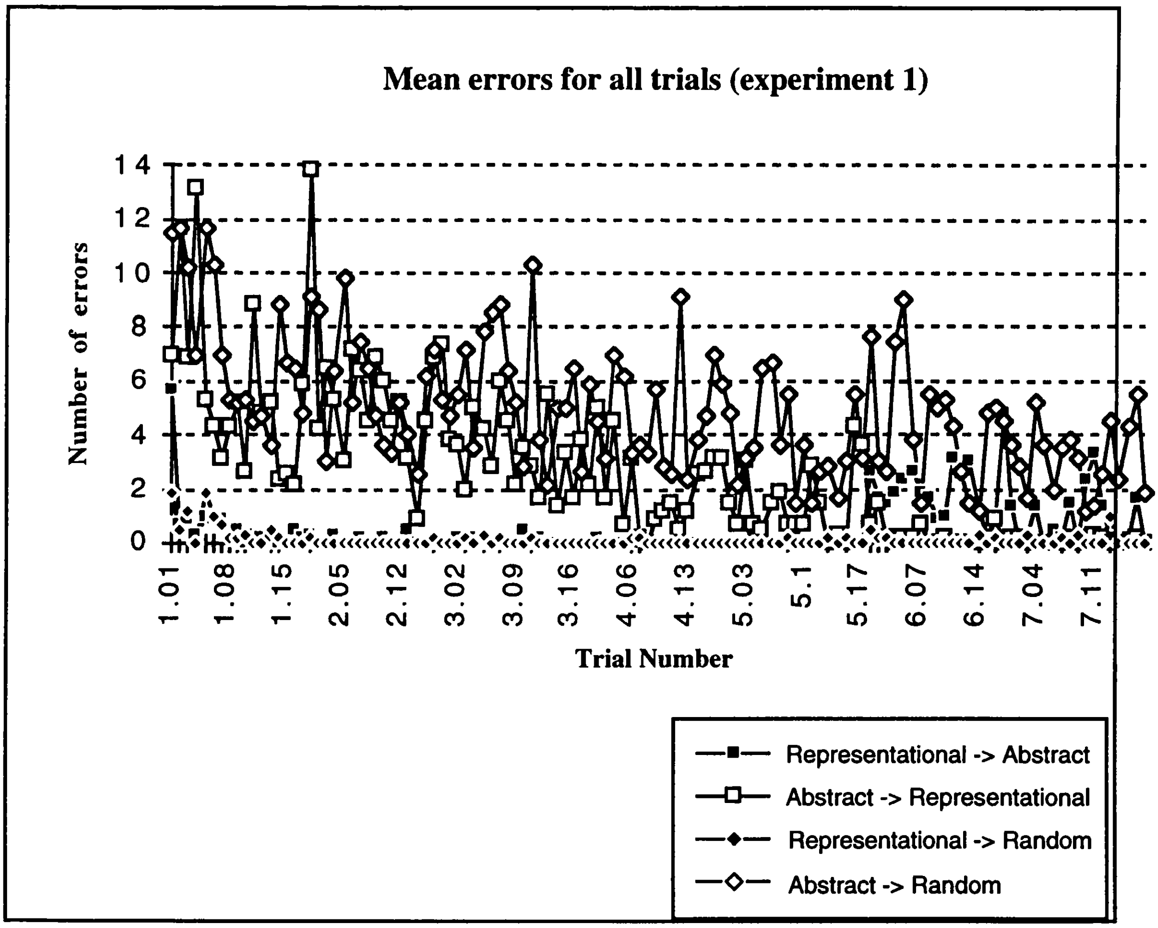
In experiments 2 to 6 the occurrence of an error resulted in the penalty being enforced. A dialogue box appeared and the user was required to select the OK box in order to proceed.



Once the dialogue box had disappeared subjects were shown which icon they should have selected. The next trial commenced immediately after this.

Appendix 2 - Reaction time and error graphs for experiment 1





Appendix 3 - Statistical results from experiment 2

B = Block; G = Condition (Group)

Condition 1 = Representational -> Abstract

Condition 2 = Abstract -> Abstract (different)

Condition 3 = Representational -> Random

Condition 4 = Abstract -> Random

1. ANOVA

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
G	3	119.532	39.844	3.263	.0361	
Error	28	341.889	12.210			
B	2	157.980	78.990	6.073	.0041	
GB	6	137.667	22.945	1.764	.1233	
Error	56	728.379	13.007			.58

G=Group (1,2,3,4)
B=Block (20,21,22)

2. Simple Effects MSW.Cell

Effect	MSn	DFn	DFe	MSe	F	p
G at B 1	1.069	3	84	12.741	.084	.969
G at B 2	84.047	3	84	12.741	6.596	.000
G at B 3	.618	3	84	12.741	.048	.986
B at G 1	119.264	2	56	13.007	9.169	.000
B at G 2	19.376	2	56	13.007	1.490	.234
B at G 3	2.356	2	56	13.007	.181	.835
B at G 4	6.827	2	56	13.007	.525	.594

3. Standard deviations for blocks 20, 21 and 22

Block Number	Condition 1	Condition 2	Condition 3	Condition 4
20	0.87	0.11	0.2	0.25
21	2.29	0.78	0.36	2.2
22	0.53	1.03	1.69	1.21

Appendix 4 - Statistical results from experiment 3

1. ANOVA of reaction times

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
G	3	13.194	4.398	.961	.4248	
Error	28	128.135	4.576			
B	2	72.517	36.259	8.140	.0008	
GB	6	19.966	3.328	.747	.6143	
Error	56	249.455	4.455			.69

G=Group (1,2,3,4)
B=Block (5,6,7)

2. Simple Effects MSW.Cell

Effect	MSn	DFn	DFe	MSe	F	p
G at Block 1	2.126	3	84	4.495	.473	.702
G at Block 2	4.508	3	84	4.495	1.003	.396
G at Block 3	4.420	3	84	4.495	.983	.405
B at Condition 1	15.194	2	56	4.455	3.411	.040
B at Condition 2	8.574	2	56	4.455	1.925	.155
B at Condition 3	20.390	2	56	4.455	4.577	.014
B at Condition 4	2.085	2	56	4.455	.468	.629

3. Standard deviations for each block of trials

Block Number	Condition 1	Condition 2	Condition 3	Condition 4
1	1.23	3.52	2.1	4.8
2	1.85	6.06	3.86	4.15
3	2.07	7.38	1.08	3.2
4	1.83	1.75	0.72	1.93
5	0.77	0.42	0.29	1.72
6	1.86	1.07	1.97	1.82
7	2.03	0.27	0.27	0.28
8	0.47	0.46	0.24	0.4
9	0.51	0.04	0.26	0.24
10	0.19	0.31	0.17	0.47

Appendix 5 - Statistical results from experiment 4

B = Block; C = Condition

Condition 1 = Representational -> Abstract

Condition 2 = Abstract -> Representational

Condition 3 = Representational -> Random

Condition 4 = Abstract -> Random

1. ANOVA (1 between and 1 within variable)

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
C	3	47.238	15.746	4.708	.0101	
Error	24	80.263	3.344			
B	2	144.805	72.403	19.858	.0000	
CB	6	48.977	8.163	2.239	.0552	
Error	48	175.009	3.646			.56

2. MSW.Cell

Effect	MSn	DFn	DFe	MSe	F	p
C at Block 20	.167	3	72	3.545	.047	.986
C at Block 21	29.490	3	72	3.545	8.318	.000
C at Block 22	2.415	3	72	3.545	.681	.566
B at Cond. 1	70.481	2	48	3.646	19.331	.000
B at Cond. 2	4.749	2	48	3.646	1.303	.281
B at Cond. 3	7.407	2	48	3.646	2.032	.142
B at Cond. 4	14.254	2	48	3.646	3.909	.027

3. Standard Deviations for Blocks 20, 21 and 22.

Block Number	Condition 1	Condition 2	Condition 3	Condition 4
20	0.30	0.25	0.24	0.71
21	5.51	0.94	1.67	1.95
22	1.59	0.89	0.51	0.68

4. Tukey (hsd) for Block

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C
A. B 1	X	-	s
B. B 3	-	X	s
C. B 2	s	s	X

5. Tukey (hsd) for Condition

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C	D
A. C 2	X	-	-	s
B. C 3	-	X	-	s
C. C 4	-	-	X	-
D. C 1	-	-	-	X

Appendix 6 - Statistical results from experiment 5

B = Block; C = Condition

Condition 1 = Representational -> Abstract

Condition 2 = Abstract -> Abstract (different)

Condition 3 = Representational -> Random

Condition 4 = Abstract -> Random

1. ANOVA (One between and one within variable)

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
C	3	73.549	24.516	5.216	.0065	
Error	24	112.809	4.700			
B	2	160.301	80.150	18.766	.0000	
CB	6	42.780	7.130	1.669	.1491	
Error	48	205.014	4.271			.57

2. Simple effects MSW.Cell

Effect	MSn	DFn	DFe	MSe	F	p
C at B 1	1.922	3	72	4.414	.435	.728
C at B 2	32.496	3	72	4.414	7.362	.000
C at B 3	4.359	3	72	4.414	.987	.404
B at C 1	70.481	2	48	4.271	16.502	.000
B at C 2	9.398	2	48	4.271	2.200	.122
B at C 3	7.407	2	48	4.271	1.734	.187
B at C 4	14.254	2	48	4.271	3.337	.044

3. Tukey (hsd) for Block

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C	D
A. C 2	X	-	-	s
B. C 3	-	X	-	-
C. C 4	-	-	X	-
D. C 1	s	-	-	X

4. Tukey (hsd) for Condition

Upper Triangle: .05 level ; Lower Triangle: .01 level			
	A	B	C
A. B 1	X	-	s
B. B 3	-	X	s
C. B 2	s	s	X

Appendix 7 - Statistical results from experiment 6

Total of 16 subjects.

16 Icons and 40 blocks of trials (20/20 split)

Condition 1: Representational -> Blank

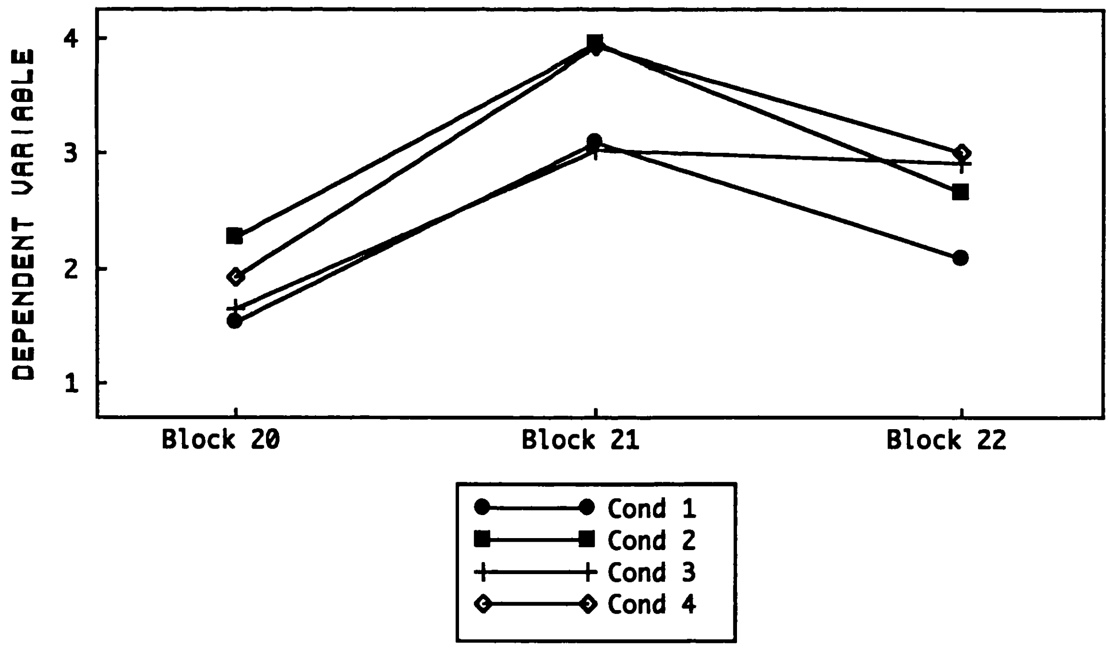
Condition 2: Abstract -> Blank

Condition 3: Representational -> Random

Condition 4: Abstract -> Random

1. Reaction time data

1.1 Graph of mean reaction times over blocks 19, 21 and 21



1.2 ANOVA summary table (1 between and 1 within variable)

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
G	3	70.586	23.529	4.637	.0036	
Error	252	1278.569	5.074			
B	2	350.474	175.237	39.342	.0000	
GB	6	32.295	5.383	1.208	.3003	
Error	504	2244.913	4.454			.86

1.3 Tukey (hsd) for Block

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C
A. Block 20	X	s	s
B. Block 22	s	X	s
C. Block 21	s	s	X

1.4 Tukey (hsd) for Condition

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C	D
A. Cond 1	X	-	s	s
B. Cond 3	-	X	-	-
C. Cond 4	-	-	X	-
D. Cond 2	-	-	-	X

1.5 MSW.Cell

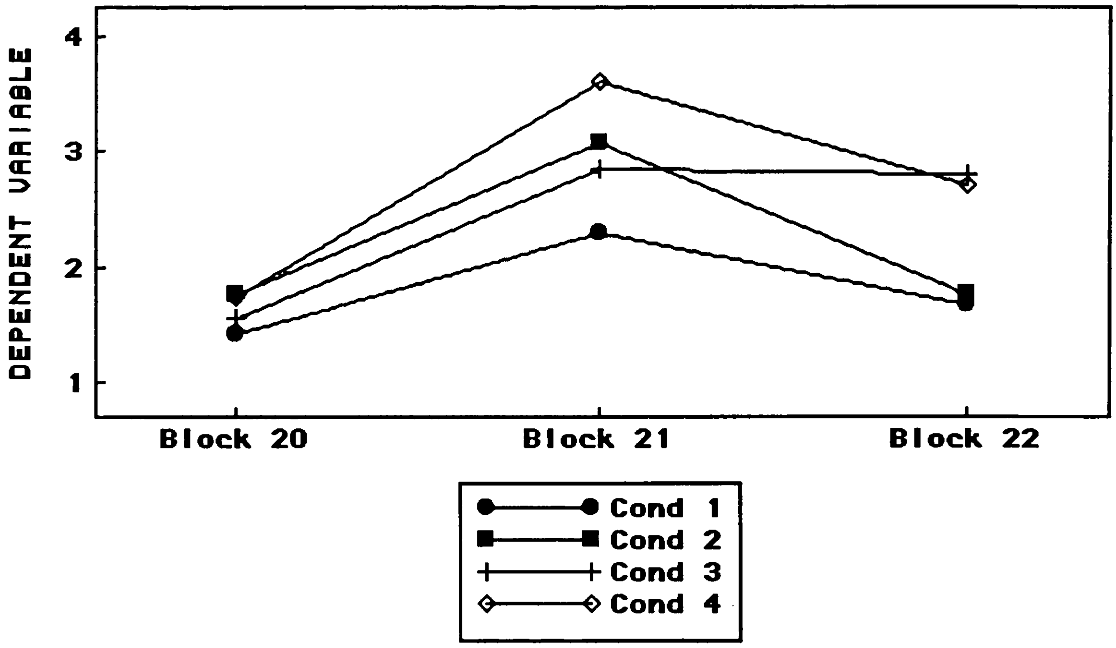
Effect	MSn	DFn	DFe	MSe	F	p
G at Block 20	6.962	3	753	4.661	1.494	.215
G at Block 21	16.135	3	753	4.661	3.462	.016
G at Block 22	11.196	3	753	4.661	2.402	.066
B at Cond 1	40.058	2	504	4.454	8.993	.000
B at Cond 2	49.328	2	504	4.454	11.075	.000
B at Cond 3	38.631	2	504	4.454	8.673	.000
B at Cond 4	63.367	2	504	4.454	14.226	.000

1.6 Standard Deviations (blocks 20, 21, and 22)

Block No.	Condition 1	Condition 2	Condition 3	Condition 4
20	1.13021105	2.31591661	0.53151345	1.95847979
21	2.82489512	3.7611772	1.56966918	3.00988084
22	1.39622576	2.06130629	1.23812701	1.94632336

2. Reaction time statistics (for pure R/T)

2.1 Graph of means across blocks 20, 21 and 22.



2.2 ANOVA summary table

ANOVA Summary Table for Exp. 7 Stats:Pure Stats

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
G	3	73.409	24.470	7.706	.0001	
Error	231	733.556	3.176			
B	2	204.300	102.150	34.849	.0000	
GB	6	44.986	7.498	2.558	.0190	
Error	462	1354.213	2.931			.73

2.3 Tukey (hsd) for Block

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C
A. Block 20	X	s	s
B. Block 22	s	X	s
C. Block 21	s	s	X

2.4 Tukey (hsd) for Condition

Upper Triangle: .05 level ; Lower Triangle: .01 level

	A	B	C	D
A. Cond 1	X	-	s	s
B. Cond 2	-	X	-	-
C. Cond 3	-	-	X	-
D. Cond 4	s	-	-	X

2.5 Simple effects MSW.Cell

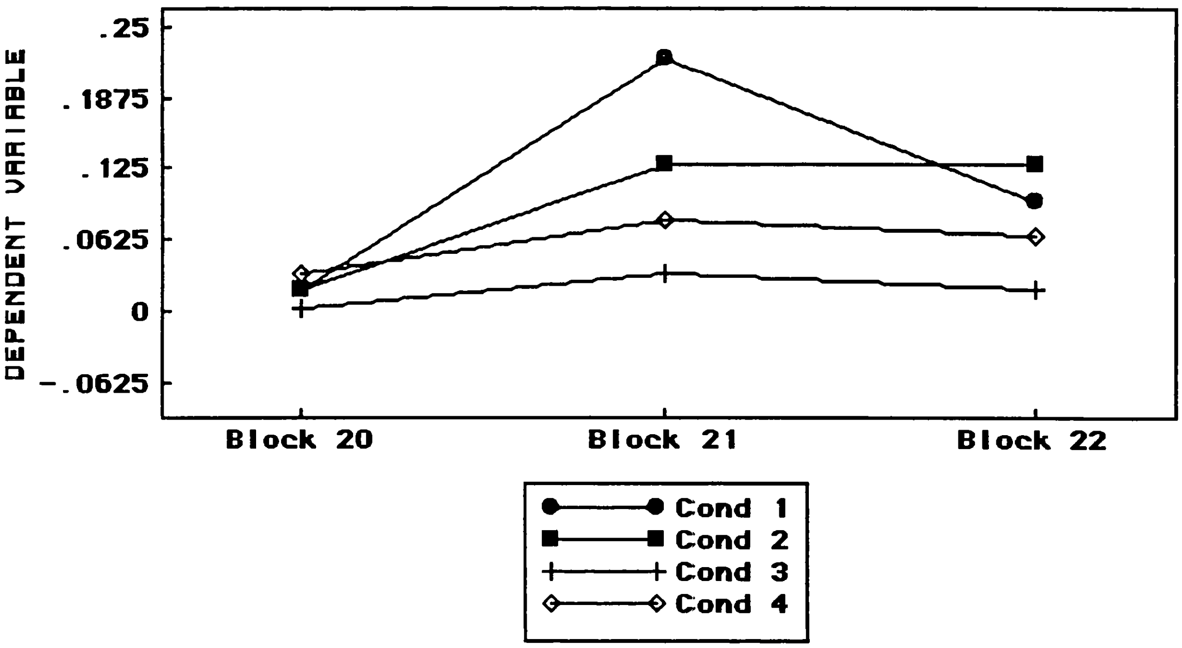
Effect	MSn	DFn	DFe	MSe	F	p
G at Block 20	1.557	3	692	3.013	.517	.671
G at Block 21	17.361	3	692	3.013	5.763	.001
G at Block 22	20.546	3	692	3.013	6.820	.000
B at Cond 1	12.779	2	462	2.931	4.360	.013
B at Cond 2	24.550	2	462	2.931	8.375	.000
B at Cond 3	34.342	2	462	2.931	11.716	.000
B at Cond 4	56.201	2	462	2.931	19.173	.000

2.6 Standard Deviations for blocks 20, 21 and 22

Block No.	Condition 1	Condition 2	Condition 3	Condition 4
20	0.88530088	2.29281343	0.5347181	1.50353199
21	2.20339431	3.22333844	1.33460626	2.82583634
22	0.87672822	1.39828762	1.16498745	1.78810955

3. Error rate statistics

3.1 Graph of mean errors across blocks 20. 21 and 22



3.2 ANOVA summary table

Source of Variation	df	Sum of Squares	Mean Square	F	p	Epsilon Correction
G	3	.958	.319	4.973	.0023	
Error	252	16.188	.064			
B	2	1.237	.618	10.612	.0000	
GB	6	.721	.120	2.063	.0561	
Error	504	29.375	.058			.92

3.3 Tukey (hsd) for Block

	A	B	C
A. Block 20	X	s	s
B. Block 22	-	X	-
C. Block 21	s	-	X

3.4 Tukey (hsd) for Condition

	A	B	C	D
A. Cond 3	X	-	s	s
B. Cond 4	-	X	-	-
C. Cond 2	-	-	X	-
D. Cond 1	s	-	-	X

3.5 Simple effects MSW.Cell

Effect	MSn	DFn	DFe	MSe	F	p
G at Block 20	.010	3	754	.060	.173	.915
G at Block 21	.410	3	754	.060	6.806	.000
G at Block 22	.139	3	754	.060	2.312	.075
B at Cond 1	.672	2	504	.058	11.528	.000
B at Cond 2	.255	2	504	.058	4.379	.013
B at Cond 3	.016	2	504	.058	.268	.765
B at Cond 4	.036	2	504	.058	.626	.535

Appendix 8 - Materials for consistency experiment

1. Overview of the experiment

(to be read by the subject at the start of experiment)

The purpose of this experiment is to investigate the ease of use, or usability of a simulated iconic interface. The interface simulates a file handling package to enable a range of manipulations to be performed on different types of files.

During the experimental session you will be set a series of tasks to perform, using the interface under test. Each task will involve performing a simulated file manipulation. This will be done using the icons on the interface.

Remember, it is the interface that is being tested, not you, so don't worry if you have any difficulties, or cannot complete a task.

Do you have any questions?

2. Experimental Protocol

(to be read by the experimenter before running a subject)

1. Give subject the "Overview of the experiment" to read.
Answer any general questions, but not any pertaining to the meanings of particular icons.

2. Administer "Personal details interview". Say: "I am now going to ask you for some personal details. The details that I shall ask for are only things that are directly relevant to the experiment, and shall be treated in the strictest confidence."

3. Give basic instructions session. Say: "In order to perform the tasks that you will be set during the experiment you will need to known the general principles of how such tasks are performed. I am going to demonstrate this to you, and then let you practice at it a few times, until you feel competent at it."

Perform a sample task on the demonstration screen, then let the subject have a try.

At the end of this session say: "That concludes the basic instruction session, we shall soon be ready to start the experimental session".

4. Start the experimental session. Load the experimental interface. Say: "We are now going to start the experimental session. I shall give you instructions by telling you the task that I wish you to perform. If you have any difficulties or get anything wrong keep trying. Remember, however, that it is the word processor under test not you, so don't worry if you have difficulties, or get some things wrong."

Note down times and errors on the experimenter's sheet.

As soon as a task is complete set the next one without delay, until the experiment is over.

If subjects are having problems with the physical mechanics of performing tasks (e.g. if they click on things in the wrong order) help them with this, but don't give any assistance if asked about an icon's function.

5. Conclude experiment. When all tasks are complete say: "That concludes the experiment, thank you very much indeed for your co-operation."

3. Icons used

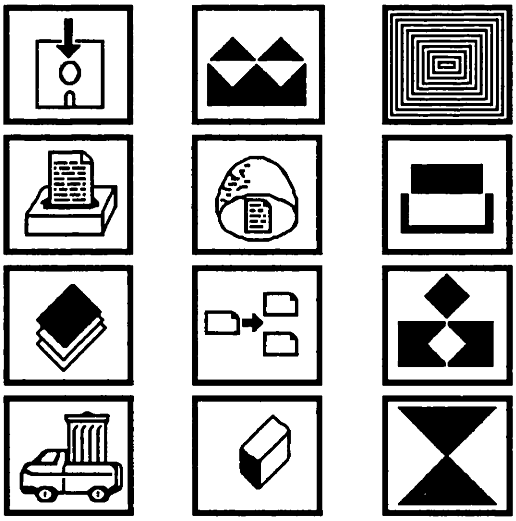
Best Mediocre Worst

Save to disk

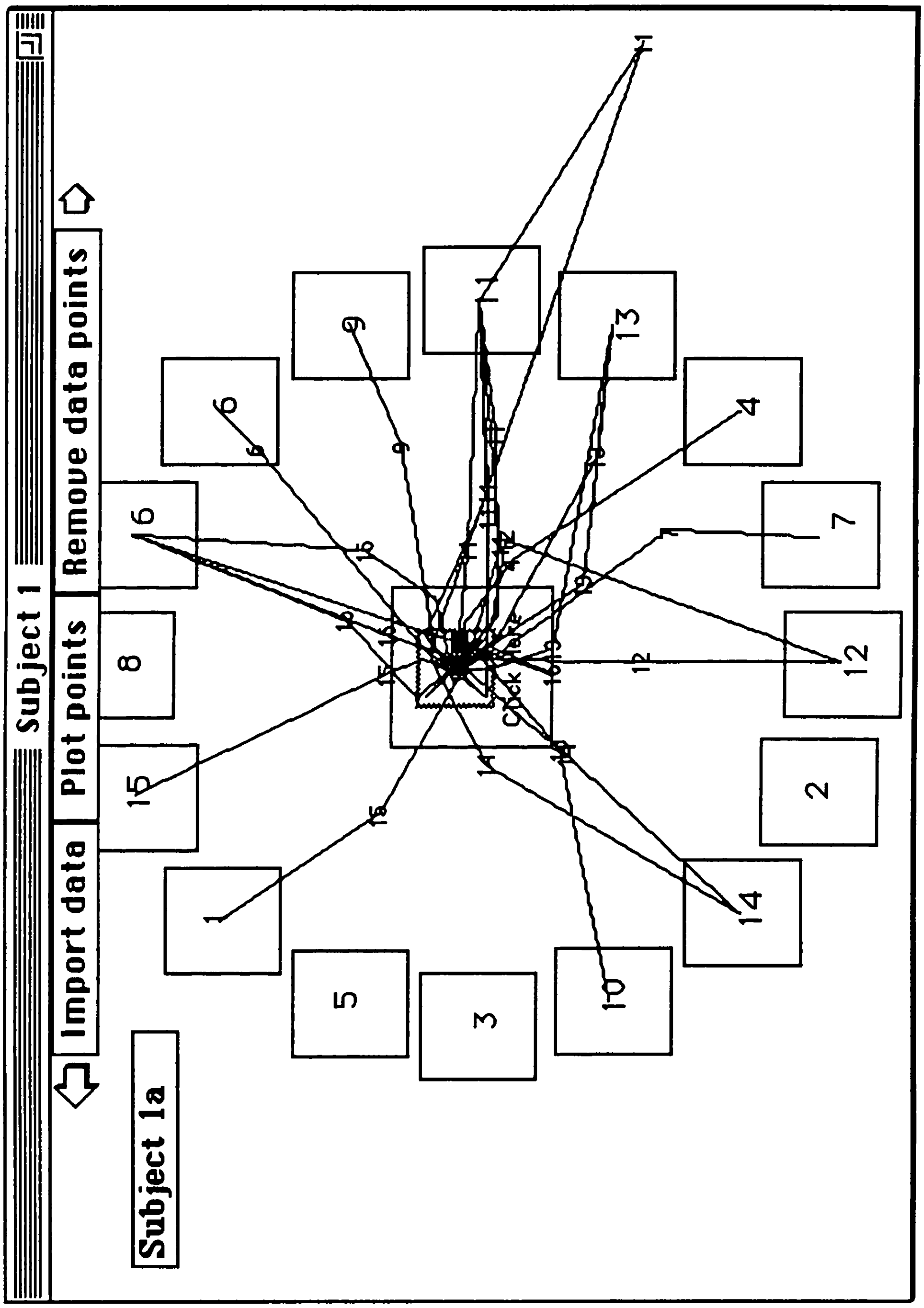
Print

Copy

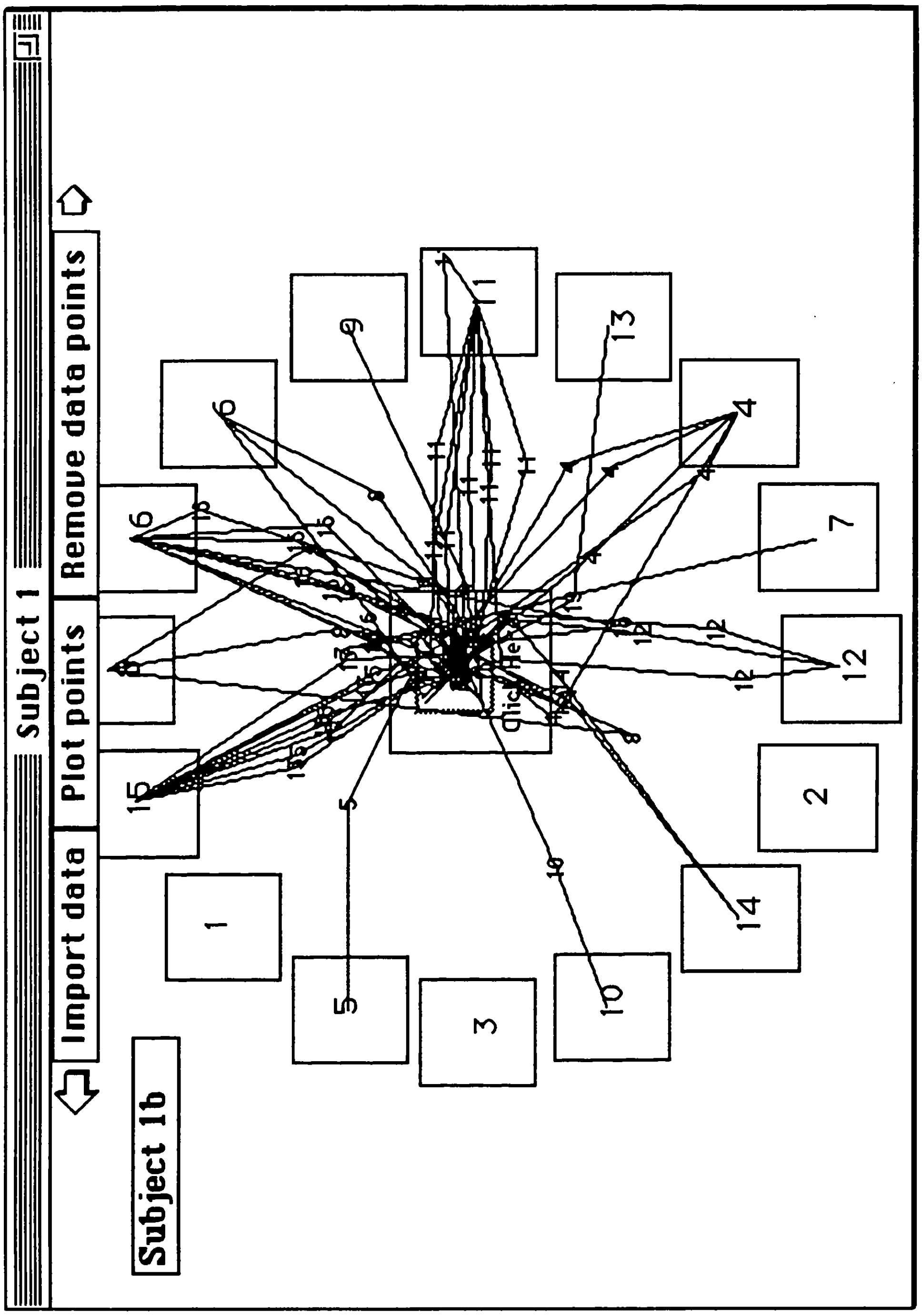
Delete



Appendix 9- Plotter stack interfaces



Pre-emptive move data was recorded using a HyperCard Stack created to plot the position of the mouse before the interface card appeared. The rectangles refer to the icon positions. The numbers in the rectangles relate to the icons (e.g. Number 1 = "Delete one character", etc.). The lines show the plot of the mouse movement, ending at the icon the user finally selected. The numbers attached to the lines represent the number of the command that had been presented (e.g. Number 1 = "Delete one character", etc.)



For each subject the plot data was separated into two cards; one for before the changeover and one for all trails after the changeover. This allowed the researcher to compare the degree of pre-emptive move activity before and after the changeover.

Bibliography

AIGA (1981). *Symbol Signs*, New York : Hastings House.

Arend, U., Muthig, K.P. & Wandamacher, J. (1987). Evidence for global feature superiority in menu selection by icons. *Behaviour and Information Technology*, 6 (4), pp. 411-426.

Arnheim, R. (1969). *Visual Thinking*, Berkeley: University of California Press.

Bannon, L.J. & Bødker, S. (1991). Beyond the interface:encountering artifacts in use. In Carroll, J.M. (ed.) *Designing Interaction: psychology at the human-computer interface,* CUP, Ch 12, pp. 227-253.

Barnard, P.J. & Marcel, A.J. (1977). A preliminary investigation of factors influencing the interpretation of pictorial instructions for the use of apparatus. In *Proceedings of the 8th International Symposium on Human Factors in Telecommunications*, Cambridge.

Bartlett F.C. (1932). *Remembering: A study in experimental and social psychology*, London: Cambridge University Press.

Bates, E. (1979). *The emergence of symbols: cognition and communication in infancy*, New York: Academic.

Baxter, I. & Oatley, K. (1991). *Comparing the learnability of spreadsheets*, Unpublished Report, University of Glasgow: Department of Psychology.

Bennett, J.L. (1984). Managing to meet usability requirements. Establishing and meeting software goals. In*Visual Display Terminals*, Bennett, J.L. & Care, D. et al. (eds.), pp. 161-183.

Bennett, K.B. & Flach, J.M. (1992). Graphical displays: Implications for divided attention, focused attention and problem solving. *Human Factors*, 34 (5), pp. 513 - 533.

Bewley, W.L., Roberts, T.L., Schroit, D. & Verplank, W.L. (1983). Human Factors Testing in the Design of Xerox 8010 "Star" Office Workstation. *CHI'83 Proceedings*, pp. 72-77.

Bertin, J. (1983). *Semiology of Graphics*, The University of Wisconsin Press.

Biederman, I. & Ju, G. (1988). Surface versus Edge-Based Determinants of Visual Recognition. *Cognitive Psychology*, 20, pp. 38-64.

Blankenberger, S. & Hahn, K. (1991). Effects of icon design on human-computer interaction. *International Journal of Man-Machine Studies*, 35, pp. 363-377.

Blattner, M. M., Sumikawa, D. A. & Greenberg, R. M. (1989). Earcons and Icons: their Structure and Common Design Principles. *Human-Computer Interaction*, 4, pp.11-44.

Bødker, S. (1989). A Human Activity Approach to User Interfaces. *Human-Computer Interaction*, 6, pp. 171-195.

Brainard, R.W., Campbell, R.J. & Elkin, E.H. (1961). Design and Interpretability of Road Signs. *Journal of Applied Psychology*, 45(2), pp. 130-136.

Brooke, J., Bevan, N., Brigham, F., Harker, S. & Youmans, D. (1990). Usability statements and standardisation - work in progress in ISO. In *Human-Computer Interaction - INTERACT '90*, Diaper, D. et al. (eds.), Elsevier Science Publishers B.V (North-Holland).

Browne, R.M. & Stammers, R.S. (1986). The evaluation and generation of icons for a computer drawing package. *Contemporary Ergonomics* (Megaw, E.D., ed.), Taylor & Francis: London.

Bruce, V. & Green, P. (1987). *Visual perception. Physiology, Psychology and Ecology*, Lawrence Erlbaum Associates, Ltd.

Byrne, M.D. (1993). Using Icons to Find Documents: Simplicity is Critical. In *Bridges Between Worlds INTERCHI '93 Conference Proceedings*, Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press, pp. 446-453.

Cahill, M.-C. (1975). Interpretability of Graphic Symbols as a Function of Context and Experience Factors. *Journal of Applied Psychology*, 60 (3), pp. 376-380.

Card, S.K., Moran, T.P. & Newell, A. (1983). *The psychology of human computer interaction*, Hillsdale, NJ: Lawrence Erlbaum.

Carmichael, L., Hogan, H.P. & Walter, A.A. (1932). An experimental study ofthe effect of language on the reproduction of a visually perceived form. *Journal of Experimental Psychology*, 15, pp. 73-86.

Carroll, J.M. (1993). Creating a design science of human-computer interaction. *Interacting with Computers*, 5 (1), pp. 3-12.

Cohen, L. & Holliday, M. (1984). *Statistics for Social Scientists. An Introductory Text with Computer Programs in Basic*, London: Harper & Row, pp. 8-9.

Consise Oxford English Dictionary (1991). Eighth edition, Allen, R.E. (ed.), Oxford University Press.

Daniel, T.C. (1972). Nature of the effect of verbal labels on recognition memory for form. *Journal of Experimental Psychology*, 96, pp. 152-157.

de Groot, A.D. (1965). *Thought and Choice in Chess*, Mouton, The Hague.

de Groot, A.D. (1966). Perception and memory versus thought. In *Problem-Solving: Research, Method and Theory*, Kleinmuntz (ed.), Wiley.

Deregowski, J.B. (1989). Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences*, 12, pp. 51-119.

Dewar, R.E. & Ells, J.G. (1977). The semantic differential as an index of traffic sign perception and comprehension. *Human Factors*, 19 (2), pp. 183-189.

Dipboye, R.L. & Flanagan, M.F. (1970). Research settings in industrial and organisational psychology. *American Psychologist*, 34, pp. 141-150.

Draper, S.W. (1991). *Icons, Representation, Information and Affordance* , Technical Report, December 1991.

Draper, S.W. (1993a). The Notion of Task in HCI. In *Bridges Between Worlds, INTERCHI '93 Adjunct Proceedings*. Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press, pp. 207-208.

Draper, S.W. (1993b). *Søgaard's example*. Technical memo, June 23.

Draper, S.W & Barton, S. (1993). Learning by exploration, and affordance bugs. In *Bridges Between Worlds, INTERCHI '93 Adjunct Proceedings*, Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press, pp. 75-76.

Dretske, F. (1988). *Explaining Behaviour*, MIT.

Eason, K.D. (1984). *Behaviour and Information Technology*, Volume 3 (2).

Easterby, R.S. (1970). The perception of symbols for machine displays. *Ergonomics*, 13, pp. 149-158.

Easterby, R.S. & Hakiel, S.R. (1977). Safety Labelling of consumer products: Field studies of sign recognition. *Report AP 76*, Applied Psychology Department, University of Aston, Birmingham.

Easterby, R.S. & Zwaga H. (1976).. Evaluation of public information symbols, ISO tests: 1975 Series. *Report AP 60,* Applied Psychology Department, University of Aston, Birmingham.

Ells, J.G. & Dewar, R.E. (1979). Rapid comprehension of verbal and symbolic traffic sign messages. *Human Factors*, 21, pp. 161-168.

Elmes, D.G., Kantowitz, B.H. & Roediger, H.L. (1989). *Research Methods in Psychology* (3rd Edition.), St. Paul, MN: West Publishing.

Erikson, C.W. (1953). Object Location in a Complex Perceptual Field. *Journal of Experimental Psychology*, 45 (2), pp. 126-132.

Erikson, C.W. (1952). Location of objects in a visual display as a function of the number of dimensions on which the objects differ. *Journal of Experimental Psychology*, 44, pp. 56-61.

Familant M., E. & Detweiler, M.C. (1991, October). *Iconic reference: Evolving Perspectives and an Organising framework.*, TM-INS-020268.

Fisch, M.H. (1986). *Peirce, semiotic, and pragmatism: essays by Max H. Fisch.* Ketner, K.L. & Kloesel, C.J.W. (eds.), Indiana University Press.

Fitts, P.M. (1951). Engineering Psychology and Equipment Design. In *Handbook of Experimental Psychology*, Stevens, S. (ed.), pp. 1267-1286.

Franks, J.J. (1974). Toward understanding understanding. In *Cognition and symbolic processes*, Weimer, W.B & Palermo, D.S. (eds.), Hillsdale, NJ: Erlbaum.

Garcia, M., Badre A.N. & Statsko, J.T. (1994). Development and validation of icons varying in their abstractness. *Interacting with Computers*, 6 (2), pp.191-212.

Garnham, A. (1985). *Psycholinguistics - Central topics*, London: Methuen.

Gaver, W.W. (1986). Auditory Icons: Using Sound in Computer Interfaces. *Human Computer Interaction* , 2, pp. 167-177.

Gaver, W.W. (1989). The Sonicfinder: An Interface that Uses Auditory Icons. *Human Computer Interaction* , 4 , pp. 67-94.

George, D.A. (1988). *User Performance with Abstract and Concrete Icons.* Bsc. Project, Applied Psychology Department, University of Aston, Birmingham.

Gibson, J.J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Gilb, T. (1981). *Design by Objectives.* Unpublished Manuscript.

Gittins, D. (1986). Icon-Based Human-Computer Interaction. *International Journal of Man Machine Studies*, 24, London: Academic Press Inc., pp. 519-543.

Greenlee, D. (1973). *Peirce's Concept of Sign.*, Mouton & Co. N.V. publishers, The Hague.

Hakiel, S.R. & Easterby, R.S. (1987). Methods for the Design and Evaluation of Icons for Human - Computer Interfaces. *IEE 2nd International Conference on Command, Control, Communications and Management'.*

Hall, J.F. (1989). *Learning and Memory,* Allyn & Bacon.

Heil, J. (1983). *Perception and Cognition.*, University of California Press: Berkeley.

Hemenway, K (1982). Psychological Issues in the Use of Icons in Command Menus. *Proceedings on Human Factors in Computer Systems*, ACM, Gaithersberg, MD, USA.

Howell, W.C. & Fuchs, A.H.(1968). Population Stereotypy in Code Design. *Organisational Behaviour and Human Performance*, 3, pp. 310-339.

Hutchins, E.L., Hollan, J.D. & Norman, D.A. (1986). Direct Manipulation Interfaces. *User Centered Interface Design*, Norman, D.A. & Draper, S.W. (eds.), Ch. 5, pp. 86-124.

ISO (1981). *Procedure for the development of symbols,* Ref. ISO TC145/SC1.

ISO (1992). *Graphic Symbols Used on Screens: Interactive Icons: Part 1 - General,* Project ISO/IEC 1.18.21.02, Thursday, July 23, 1992.

Jervell, H.R. & Olsen, K.A. (1985). Icons in Man-Machine Communications. *Behaviour and Information Technology.*, 4 (3), pp. 249-254.

Johnson, P. & Johnson, H. (1989). Knowledge Analysis of Tasks: Task analysis and specifications for human-computer systems. In Downton, A. (ed.) *Engineering the human computer interface*, New York: McGraw Hill.

Jones, S. (1983). Stereotypy in pictograms of abstract concepts. *Ergoniomics*, 26, pp. 605-611.

Jordan, P.W. (1993). *Consistency and Usability.* Thesis (Ph.D.), University of Glasgow.

Jordan, P W. & Draper S W (In Review). Set and Rule Compatibility and their Effects on Guessability, Learnability and EUP. Submitted to *Behaviour and Information Technology.*

Jordan, P W, Draper, S W, MacFarlane, K K & McNulty, S-A. (1991). Guessability, Learnability, and Experienced User Performance. In *People and Computers VI,* Diaper, D. & Hamond, N. (eds.), Cambridge University Press.

Jordan, P.W. & O'Donnell, P.J. (1992). Quantifying guessability, learnability and experienced user performance. *Contemporary Ergonomics 1992*, Lovesey, E.J. (ed.), London: Taylor & Francis, pp. 404-410.

Kantowitz, B.H. (1992). Selecting Measures for Human Factors Research.*Human Factors*, 34 (4), pp. 387-398.

Kaptelinin, V. (1993). Icon Recognition in Menu Selection: The Effect Of Practice. In *INTERCHI'93 Adjunct Proceedings*, Ashlund, S. et al. (eds.), ACM Press.

Kieras, D.E. & Polson, P.G. (1985). An approach to the formal analysis of human complexity. *International Journal of Man-Machine Studies*, 22, pp. 365-394.

Koffka, K. (1935). *Principles of Gestalt Psychology*, New York: Harcourt & Brace.

Kohlers, P.A. (1969). Some formal characteristics of pictograms. *American Scientist*, 57, pp. 348-363.

Lansdale, M.W. (1988). On the Memorability of Icons in an Information Retrieval Task. *Behaviour and Information Technology*, 7 (2), pp. 131-151.

Lansdale, M.W., Simpson, M. & Stroud, T.R.M. (1990). A comparison of words and icons as external memory aids in an information retrieval task. *Behaviour and Information Technology*, 9 (2), pp. 111-131.

Leech, G.N. (1983) *Principles of Pragmatics*. New York: Longman.

Lodding, K.N. (1983). Iconic Interfacing, *Computer Graphics and Applications*, 24, pp.11-20.

Mackinley, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5 (2).

Maissel, J. (1990). Development of a Methodology for Icon Evaluation. *NPL Report*, DITC 159/90.

Maissel, J., Macleod, M., Thomas, C., Rengger, R., Coorooran, R., Dillon, A., Maguire, M., Sweeney, M. (1993). *Usability Context Analysis: A practical guide.* Esprit project 5429.

Marcus, A. (1984). Corporate Identity for Iconic Interface design: The Graphic Design Perspective. *IEEE Computer Graphics and Applications*, December, pp. 24-32.

Marshall, E. & Walker, P. (1987). Visual memory for pictorial stimuli in a serial choice reaction-time task. *British Journal of Psychology*, 78, pp. 213-231.

Mayes, J.T., Draper, S.W., McGregor, M.A., Oatley, K. (1988). Information flow in a user interface: the effect of experience and context on the recall of MacWrite screens. In *People and Computers IV*, Jones, D.M. & Winder, R. (eds.), Cambridge University Press, pp. 275-289.

Milner S.J. (1987). The development of an iconic interface for a commercially available CAD package. *MSc. Project*, Loughborough University of Technology.

Moyes, J. & Jordan, P.W. (1993). Icon Design and its Effect on Guessability, Learnability and Experienced User Performance. In *People and Computers, Proceedings of HCI'93*, Alty, J.L., Diaper, D. & Guest, S. (eds.), Cambridge University Press.

Nickerson, R.S. & Adams, M.J. (1979). Long-Term Memory for a Common Object. *Cognitive Psychology*, 11, pp. 287-307.

Norman, D.A. (1986). Cognitive Engineering. In *User centred system design: New perspectives in human-computer interaction*, Norman, D.A. & Draper, S.D. (eds.), Hillsdale, NJ:Lawrence Erlbaum Associates.

Norman, D.A. (1991). Cognitive Artifacts. In *Designing interaction: Psychology at the human-computer interface* , Carroll, J.M (ed.), Cambridge University Press, pp. 17-38.

Norman, D.A. (1993). *Things That Make Us Smart*, Addison-Wesley Publishing Company.

Norman, D.A., Rumelhart, D.E. & LNR (1975). *Explorations in Cognition*, Freeman.

Olson, R.K. & Attneave, F. (1970). What variables produce similar grouping? *American Journal of Psychology*, 83, pp. 1-21.

O'Malley, C. (1988). *Iconic Interfaces: Some Issues for the Design of Learning Environments*, Pre-Delta Project TF III, Task a.

Osgood, C., Suci, G. & Tannenbaum, P. (1957). *The Measurement of Meaning,* Urbana: University of Illinois Press.

Paivio, A., Rogers, T.B. & Smythe, P.C. (1968). Why are pictures easier to recall than words? *Psychonomic Science,* 11 (4), pp. 137-138.

Palmer, F.R. (1988). *Semantics - A New Outline,* Cambridge University Press.

Palmer, S. (1978). Fundamental Aspects of Cognitive Representation. In *Cognition and Categorization,* Rosh, E. & Lloyd,B.B.(eds.), Lawrence Erlbaum Associates.

Payne, S.J. (1984). Task Action Grammars. In Human-Computer Interaction - INTERACT '84, Shackle, B. (ed.), North Holland: Elsevier, pp. 527-532.

Payne, S.J. & Green, T.R.G. (1983). The user's perception of the interaction language: a two level model. In *CHI'93 conference proceedings,* New York: Academic Press, pp. 202-206.

Payne, S.J. & Green, T.R.G. (1986). Task-action grammars:a model of the mental representation of task languages, *Human-Computer Interaction* , 2, pp. 93-133.

Payne, S.J. & Green, T.R.G. (1989). Task-Action Grammars: the model and its developments. In *Task Analysis for Human-Computer Interaction,* Ellis Horwood, pp. 75-107.

Peirce, C S (1931-1958) .In *Collected papers of Charles Sanders Peirce,* Hartshorne, C., Weiss, P. & Burks, A. (eds.), 8 Vols, Cambridge M A.: Harvard University Press.

Pollock, C. (1988). Training for Optimising Transfer between Word Processors. In *Proceedings of HCI'88 Conference on People and Computers,* Cambridge University Press, pp. 309-328.

Pomerantz, J.R. & Schwaltzberg, S.D. (1975). Grouping by proximity: Selective attention measures. *Perception and Psychophysics,* 18, pp. 355-361.

Pomerantz, J.R. (1983). Global and local Precedence: Selective attention in form and motion perception. *Journal of Experimental Psychology: General.,* 112, pp. 689-697.

Ravden, S.J. & Johnson, G.I. (1989). *Evaluating Usability of Human-Computer Interfaces: A practical Method,* Chichester: Ellis Horwood.

Reisner, P (1990). What is Inconsistency? In *Human-Computer Interaction. INTERACT '90,* Diaper, D et al. (eds.), North Holland: Elsevier, pp. 175 - 181.

Reisner, P. (1981). Formal grammar and human factors design of an interactive graphics system. *IEEE Transactions on Software Engineering,* SE-7 (2), pp. 229-240.

Restle, F. (1961). *Psychology of Judgement and Choice,* New York:Wiley.

Rohr, G. & Keppel, E. (1984). Iconic Interfaces: Where to Use and How to Construct. *Human Factors in Organisational Design and Management,* Amsterdam: North Holland.

Rogers, Y. (1986). Evaluating the Meaningfulness of Icon Sets to Represent Command Operations. In *People and Computers, Proceedings of HCI'86,* Harrison, M. & Monk, A (eds.), Cambridge University Press, pp. 586-603.

Rogers, Y. (1989a). Icons at the Interface: their Usefulness. *Interacting With Computers,* 1, pp. 105-117.

Rogers, Y. (1989b). Icon Design for the User Interface. *International Reviews of Ergonomics,* 2, pp. 129-154.

Rogers, Y. & Oborne, D.J. (1985). Some Psychological Attributes of Potential Computer Command Names. *Behaviour and Information Technology,* 4 (4), pp.349-365.

Rogers, Y. & Oborne, D.J. (1987). Pictorial Communication of Abstract Verbs in Relation to Human-Computer Interaction. *British Journal of Psychology*, 78, pp. 99-112.

Rosch, E. & Lloyd, B.B. (1978). *Cognition and Categorisation,* Lawrence Erlbaum Associates.

Russo, P. & Boor, S. (1993). How Fluent is Your Interface? Designing for International Users. *Bridges Between Worlds, INTERCHI '93 Proceedings,* Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press, pp.342-347.

Ryan, T.A. & Schwartz, C.B. (1956). Speed of perception as a function of mode of presentation. *American Journal of Psychology*, 69, pp. 60-69.

Sanford, A.J. (1986). *Cognition and Cognitive Psychology*, Harper Collins.

Scott, D. & Findlay, J.M. (1991). Future Displays: a Visual Search Comparison of Computer Icons and Words. In *Proceedings of the Ergonomics Society's 1991 Annual Conference*, Lovesey, E.J. (ed.), Taylor & Francis.

Schneiderman, B. (1987). *Designing the User Interface*, Reading, Mass.: Addison-Wesley.

Selker, T. & Koved, L. (1988). Elements of Visual Language. *IEEE*, pp. 38-44.

Shackel, B. (1986) Ergonomics in design for usability. In *People and Computers: designing for Usability. Proceedings of the second conference of the BCS HCI specialist group*, Harrison, M.D. & Monk, A.F. (eds.), Cambridge: Cambridge University Press.

Shackel, B. (1991). Usability - context, framework, definition, design and evaluation. In *Human Factors for Informatics Usability*, Shackel, B. and Richardson, S. (eds.), Cambridge: Cambridge University Press, pp. 21-37.

Shinar, D. & Drory, A. (1983). Sign registration in day-time and night-time driving. *Human Factors*, 25, pp. 117-122.

Smith, D.C., Irby, C., Kimball, R., Verplank, B. & Harslem, E. (April 1982). Designing the Star User Interface. *Byte*, 7, pp. 242-282.

Stanat, D.F. & McAllister, D.F. (1977). *Discrete Mathematics in Computing Science,* Prentice/Hall International, Inc.

Standing, L., Conezio, J. & Haber, R.N. (1970). Perception and Memory for Pictures: single trial learning of 2500 visual stimuli. *Psychonomic Science*, 19 (2), pp. 73-74.

Staples, L. (1993). Representation in Virtual Space: Visual Convention in the Graphical User Interface. In *Bridges Between Worlds, INTERCHI '93 Proceedings,* Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press, pp. 349-354.

Stothard M.L., Stammers R.B. & Carey M.S. (1988). Comprehension of icons for a computer drawing package: a follow-up study. *Contemporary Ergonomics*, Megaw, E.D. (ed.), Taylor & Francis: London, pp. 252-257.

Strijland, P. (1993). Icon Design. Tutorial Notes for *Bridges Between Worlds, INTERCHI '93,* Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press.

Tipper, S.P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *The Quarterly Journal of Experimental Psychology*, 37A, pp. 571-590.

Tipper, S.P. & Driver, J. (1988). Negative priming between pictures and words in a selective attention task: evidence for semantic processing of ignored stimuli. *Memory and Cognition*, 16 (1), pp. 64-70.

Velolia, D. (19930. Facile 3D Direct Manipulation. *Bridges Between Worlds, INTERCHI '93,* Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E. & White, T. (eds.), ACM Press.

Webb, J.M., Sorenson, P.F. & Lyons, N.P. (1989). An Empirical Approach to the Evaluation of Icons. *SIGCHI Bulletin*, July, pp. 87-90.

Werner H. & Kaplan, B. (1967). *Symbol Formation*, New York:Wiley.

Whiteside, J., Jones, S., Levey, P.S. & Wixon, D. (1985). User performance with command, menu and iconic interfaces. In *Proceedings of the CHI'85 Conference on Human Factors in Computing Systems*, New York: ACM.

Zwaga H. & Boersema T. (1983). Evaluation of a set of graphic symbols. *Applied Ergonomics* , 14, pp. 43-54.