INTERNATIONAL JOURNAL OF COGNITIVE ERGONOMICS, 2001, *5*(2), 153–178

The Effects of Visual Information on Users' Mental Models: An Evaluation of Pathfinder Analysis as a Measure of Icon Usability

Siné J. P. McDougall
Psychology Department
University of Wales Swansea

Martin B. Curry
Sowerby Research Center
BAE SYSTEMS

Oscar de Bruijn
Department of Electrical Engineering
Imperial College

ABSTRACT

Research has shown that individuals' knowledge structures change as a result of learning and experience. This article investigates the possibility that the content of graphical user interfaces can play a role in determining the nature of the knowledge structures users develop.

Users employed either concrete, abstract, or arbitrary icon sets in a computer-based problem-solving task. The effects of these icons were assessed using standard measures of performance. On the basis of the assumption that users' mental models should be better if appropriate icons were presented on the interface, Pathfinder analysis was used to elicit users' knowledge structures as they gained experience with the interface. The efficacy of this measure was then compared with performance measures.

Our findings show that users' knowledge structures do depend on the nature of the graphical information presented at the interface but do not rely as much on the use of the visual metaphor as previously thought. Although most measures were sensitive to initial differences between icon sets, only some measures were sensitive to the long-term differences that remained after users had gained experience with the icon set. The implications of these findings for interface design are discussed.

1. INTRODUCTION

Considerable efforts have been made by interface designers to ensure that icons used in displays are as easy to use as possible. An assumption implicit in their design is that

Requests for reprints should be sent to Siné McDougall, Department of Psychology, University of Wales Swansea, Swansea SA2 8PP, UK. E-mail: s.mcdougall@swansea.ac.uk

better visual information on the interface helps the user to develop a better mental model, or knowledge structure, of the system being used. This assumption was tested directly in this study by investigating whether users' knowledge structures differed as a result of the type of visual information presented to them on the interface. This was done by providing groups of users with different sets of icons that could be used to solve a set of computer-based problems. Differences in the knowledge structures created as a result of experience with each type of icon were then compared using standard measures of performance and Pathfinder analysis.

The effects of icon characteristics on users' understanding and performance were also explored. Of particular interest was the degree of icon concreteness (i.e., the extent to which icons depict objects or people from the real world) and the closeness of the relation between an icon and its function (referred to as *semantic distance*). To examine the effects of these icon characteristics, the icon sets that participants were asked to use were systematically varied.

1.1. Knowledge Representation and Learning

It is well known that individuals' knowledge structures change with learning and experience. As people become more expert in a particular domain, their knowledge representation changes in systematic ways to provide better support for their activities. Such changes have been demonstrated in areas as diverse as chess (Chase & Simon, 1973; de Groot, 1965; Saarilouma, 1994); computer programming (Adelson, 1981; McKeithen, Reitman, Rueter, & Hirtle, 1981); and the understanding of physics and mathematics (Chi, Feltovich, & Glaser, 1981; Schonfield & Herrmann, 1982). Research has also shown that knowledge structures converge and approximate better to those of an expert as experience increases (Barsalou, 1989; Goldsmith & Johnson, 1990; Rowe, Cooke, Hall & Halgren, 1996; Ye, 1997).

The extent to which knowledge structures converge on a given standard may depend on the perspective that individuals bring to bear on learning. Gomez, Schvaneveldt, and Staudenmayer (in press) compared doctors' and patients' understanding of multiple allergy sensitivity (a condition that is often difficult to attribute to a single physical cause). Patients believed their condition was the result of sensitivity to a wide range of chemicals in their environment, whereas doctors perceived the symptoms to be strongly linked to emotional causes. Systematic differences in knowledge structures were also observed in a study where participants were given different roles in a team involved in a simulated rescue-and-relief helicopter mission ~~(Rowe,~~ Stout, Rivera, & Salas, 1998). It seems likely Cooke therefore that, if individuals experience the same problem-solving task via different visual interfaces (i.e., from different visual perspectives), the knowledge structures that they develop may differ systematically as a result because the perspective they are given of the interface is different. This possibility was investigated in this experiment.

1.2. Eliciting Knowledge Structures

Before any claims can be made about changes in users' knowledge structures, appropriate measures of these changes need to be obtained. The difficulty with standard measures of problem-solving performance such as accuracy or response times is that they are only indirect measures of users' knowledge structures. Although they provide an index of how well individuals' knowledge structures support the task domain, they give little information about how that knowledge is organized.

Another way of eliciting users' mental models is by using interview or "think aloud" protocols (e.g., Ericsson & Simon, 1984; Gray, Waite, & Draper, 1990). However, as Staggers and Norcio (1993) pointed out, the limitations inherent in these techniques are that novices have difficulty articulating their knowledge whereas expert users may no longer be consciously aware of their conception of the system. Furthermore, the data sets that result are typically large and unwieldy and difficult to interpret.

Cognitive task analysis methods have also been used to model cognitive aspects of system use (e.g., Card, Moran, & Newell, 1983; Goh & Coury, 1994; Kieras, 1978; Morphew, Thordsen, & Klein, 1998). These methods lead to the production of lists of goals and details of their associated activities. Although the process of cognitive task analysis has proved useful in detailing the information users require and the behavioral sequences they need to follow, there may be a considerable disparity between these idealized models and those that the individual user employs in practice.

Another method that has been widely used to elicit information about users' knowledge structures is the Pathfinder technique developed by Schvaneveldt (1990). This involves gathering data about users' perceptions of the interrelations between key concepts in a given domain. Subjective ratings are used to assess the strength of the relations between pairs of concepts. These pairwise proximity ratings are then submitted to a multivariate analysis from which a mathematically derived "cognitive map" of users' domain knowledge can be generated. This map takes the form of a network in which concepts are represented as nodes that are connected by links to other concepts that the user considers to be related.

Although Pathfinder networks are related to semantic networks (Collins & Loftus, 1975; Meyer & Schvaneveldt, 1976; Quillian, 1969), they are distinguished by being derived directly from empirical data rather than being primarily theoretical in nature. They have, for example, been used to compare the knowledge structures of novices and experts (Gomez, Hadfield, & Housner, in press; Schvaneveldt, Durso, Goldsmith, Breen, & Cooke, 1985; Wilson, 1994) and differences between human factors experts and software development specialists (Gillan, Breedin, & Cooke, 1992), as well as to improve navigation of Internet and database systems (Chen, 1998).

Pairwise proximity judgment data has also been analyzed using multidimensional scaling. This enables each concept to be located in an *N*-dimensional space where the distance between points reflects the psychological proximity of the corresponding concepts. The dimensions defining the space represent the main properties along which the concepts within the domain are organized (i.e., categorical information). For example, a series of concepts such as *ruby, iron, sparrow, trout,* and *daisy* might yield dimensions such as living or nonliving (sparrow, trout, daisy vs. ruby, iron) and plants or animals (daisy vs. sparrow, trout; see Schvaneveldt et al., 1985). Research has shown that multidimensional scaling can capture changes that occur in knowledge structures as a function of learning (Adelson, 1981; Gonzalvo, Canas, & Bajo, 1994) and can be used to measure differences in the cognitive representations of novices and experts (Goldsmith, Johnson, & Acton, 1991; Schvaneveldt et al., 1985).

Although multidimensional scaling and Pathfinder analyses both reduce large amounts of pairwise proximity data to an interpretable form, they achieve this using mechanisms that highlight different aspects of the underlying cognitive structures. The main strength of multidimensional scaling appears to lie in its ability to capture the global categorical dimensions of a knowledge domain, whereas Pathfinder is better at capturing the local relations between concepts (Gonzalvo et al., 1994; Schvaneveldt et al., 1985). Because we were interested in the extent to which users were able to use icons to solve specific problems, Pathfinder analyses were considered more appropriate. Despite its widespread use, we are not aware of any study carried out to date that has used Pathfinder to explore how users' knowledge structures might be affected by the way visual information is presented or how these knowledge structures change as users learn how to use the interface.

1.3. Icon Usability

The effects on user performance of a wide range of icon characteristics have been investigated. These have included color (e.g., Christ, 1975; Christ & Corso, 1982; Tullis, 1981); shape (Arend, Muthig, & Wandmacher, 1987); visual complexity (Byrne, 1993; Tullis, 1983); meaningfulness (McDougall, Curry, & de Bruijn, 1999; Rogers, 1989); and the extent to which icons are concrete or abstract (McDougall, de Bruijn, & Curry, 2000; Rogers, 1989; Stammers & Hoffman, 1991).

Difficulties in interpreting these findings often arise because the icon sets chosen by researchers are not always representative examples of the characteristics they purport to investigate. This is particularly true of research examining the effects of icon concreteness. For example, researchers often chose concrete icons that were more visually complex, as well as more concrete, than their abstract counterparts (Arend et al., 1987; Green & Barnard, 1990; Rogers, 1986; Stammers, George, & Carey, 1989). As a result, the de facto assumption arose that to depict items in the real world a concrete icon had to contain more detail and would therefore be more visually complex. However, more recent research, where the concreteness and complexity of icons have been properly controlled, has shown that icon concreteness and complexity act in very different ways to determine usability (McDougall et al., 2000).

The assumption with regard to concrete icons that we wished to investigate in this study is associated with the notion of the visual metaphor that lies at the heart of modern graphical user interfaces. Concrete icons are thought to be effective in graphical user interfaces because they allow users to draw on correspondences between the real-world objects they depict and the representations of those objects embedded in the computer interface (Smith, Irby, Kimball, & Verplank, 1982). In contrast, the meanings of abstract icons are thought to be less apparent because they typically do not depict objects and, as a result, associations with the real world cannot be used to facilitate understanding of the interface. Assumptions about the benefits of employing concrete icons therefore combine the depiction of items from the real world (i.e., icon concreteness) with the strength of the relation between the icon and function by assuming that the two will always go together. However, interface designers often use abstract icons because it is simply not possible to "pictorialize" a function on an interface in a way that is likely to be meaningful to the user, that is, stronger icon-referent relations can be produced by using *less* pictorial representations. Thus, in practice, there often seems to be a dissociation between concreteness and the strength of the icon–referent relation.

In this study, the effects of concreteness and the strength of the icon–referent relation (henceforth referred to as *semantic distance*) on usability were considered (see Blankenberger & Hahn, 1991; Blattner, Sumikawa, & Greenberg, 1989; Familant &

Detweiler, 1993; McDougall et al., 1999, for further discussion of semantic distance). One of three sets of icons was presented to participants: concrete, abstract, or arbitrary. Concrete icons used depictions of real-world items to represent icon functions. Abstract and arbitrary icons were created so that they were equally nonconcrete but differed in the semantic distance between icon and referent. Abstract icons had systematic icon–referent relations (i.e., the semantic distance was less), whereas the visual representations for arbitrary icons were randomly chosen (i.e., the semantic distance was greater; see Appendix A for examples of the icons used). The hope was that, in dissociating the effects of concreteness from semantic distance, we would be able to see which icon characteristic was the primary determinant of usability. If users' understanding of the interface was better when using concrete icons, and abstract and arbitrary icons are equally ineffective, then this would suggest that it is concreteness that is the primary determinant of icon usability. If, however, abstract icons were more effective than arbitrary icons, then this would suggest that it is differences in semantic distance that are more important in determining icon usability.

To summarize, this study investigated whether the type of visual information presented at the interface affected users' mental models. This was assessed by asking users to carry out a problem-solving task using either concrete, abstract, or arbitrary icons. Pathfinder analysis was used to examine how users' knowledge structures changed as a result of their experience with the interface. Several measures of performance were also obtained to observe the relation between users' knowledge structures and their performance.

2. METHOD

2.1. Participants

Participants were undergraduates from the University of Wales, Swansea, who were given course credits or paid £10 for their participation in this study. A total of 113 individuals took part in this study. Fifty-three participants provided subjective ratings of the characteristics of icons in each of the three icon sets employed in the study so that icon characteristics could be properly controlled. Sixty participants took part in the experiment. Twenty participants each used concrete, abstract, or arbitrary icon sets in a series of problem-solving tasks.

2.2. Materials and Apparatus

The experiment was conducted using personal computers equipped with an Intel Pentium 166MHz processor and color monitor. Participants responded to the problem-solving task using a Microsoft mouse with its track speed set midway between slow and fast. Reaction time was measured using the system's multimedia timer, which allows measurement to within 1 msec. Using a mouse as an input device adds an estimated uncertainty of up to ±30 msec to timings obtained (Microsoft Corporation, 1997).

Three sets of 16 icons were used to represent possible solutions to the problems presented. The majority of the icons in the concrete and abstract sets were newly created for this study. Abstract icons were derived from the concrete sets but did not include concrete, or pictorial, components. The arbitrary icons were selected from the set of 239 icons rated by McDougall et al. (1999), but the original function labels assigned to these labels bore no relation to the functions employed in the study (hence the arbitrary nature of the icon–function relations). The icons in each set and their functions are shown in Appendix A. Subjective ratings were employed to control icon characteristics between icon sets. Ratings of the visual complexity, concreteness, and semantic distance of icons were obtained using the same procedures as those reported by McDougall et al. Ratings for each of these characteristics were obtained from separate groups of participants: 18 provided ratings of the visual complexity of the icons, 19 provided concreteness ratings, and 17 provided ratings of semantic distance. Ratings for each icon characteristic were obtained as follows:

1. Concreteness. Participants were told that icons should be regarded as concrete if they represented "real objects, materials, or people" (see Gilhooly & Logie, 1980; Paivio, Yuille, & Madigan, 1968). Those that did not should be regarded as abstract. A 5-point rating scale was employed; a rating of 5 indicated that the icon was definitely concrete, whereas a rating of 1 indicated that the icon was definitely abstract.

2. Semantic distance. This measured the closeness of the relation between the icon and the function it represented. The strength of icon–function relations can vary along a continuum from very close to very distant (see McDougall et al., 1999). Where the icon–function relation is very weak, users must rely on learning what the icon means to understand and use it (e.g., a circle with a cross to indicate female).

Participants were shown the icon function next to one of the three possible icons (a full list of icon functions and matching icons is shown in Appendix A) and asked to rate the semantic distance between icon and function. Participants were asked to use a rating of 1 to indicate that icon and function were not closely related and 5 to indicate that they were strongly related.

3. Complexity. Participants were told that the visual complexity of an icon was determined by the amount of "detail or intricacy" it contained (see Snodgrass & Vanderwart, 1980). A rating of 5 indicated that the icon was very complex, whereas a rating of 1 indicated that icon was very simple. Because this characteristic was not systematically varied across sets, care was taken to ensure that icon sets were uniformly complex.

Subjective ratings of icon meaningfulness and familiarity were not obtained because these characteristics are known to be closely related to icon concreteness (see McDougall et al., 1999).

The mean ratings and standard deviations for each characteristic of the three icon sets are shown in Table 1. Three one-way analyses of variance were carried out to examine differences in ratings of each characteristic among icon sets. Differences were apparent among icon sets in ratings of concreteness and semantic distance. Newman–Keuls comparisons confirmed that concreteness ratings were higher for the concrete icon set ($p < .05$) but were similar for the abstract and arbitrary sets ($p > .05$). Comparisons of semantic distance ratings showed that icon–function relations were significantly closer for concrete than for abstract icons and that abstract icon–function relations were significantly closer than for arbitrary icons. As can be seen from Appendix A, abstract icons related systematically to their functions whereas arbitrary icons did not. Thus whereas abstract and arbitrary icons were equally concrete, the abstract icons were better related to the functions they represented. No differences in visual complexity ratings among icon sets were evident.

TABLE 1
Mean Ratings and Standard Deviations of Icon Characteristics for Each Icon Set

	Icon Sets						
	Concrete		*Abstract*		*Arbitrary*		*ANOVA & Newman–Keuls Analysis*
Ratings	*M*	*SD*	*M*	*SD*	*M*	*SD*	
Concreteness	4.08	0.50	2.09	0.38	1.93	0.22	$F(2, 45) = 109.32, p < .001$; conc > abs = arb
Semantic distance	3.65	0.58	2.53	0.57	1.17	0.13	$F(2, 45) = 153.79, p < .001$; conc > abs > arb
Visual complexity	2.64	0.69	2.48	0.62	2.66	0.80	$F < 1$; conc = abs = arb

Note. ANOVA = analysis of variance; conc = concrete icons; abs = abstract icons; arb = arbitrary icons.

2.3. Procedure

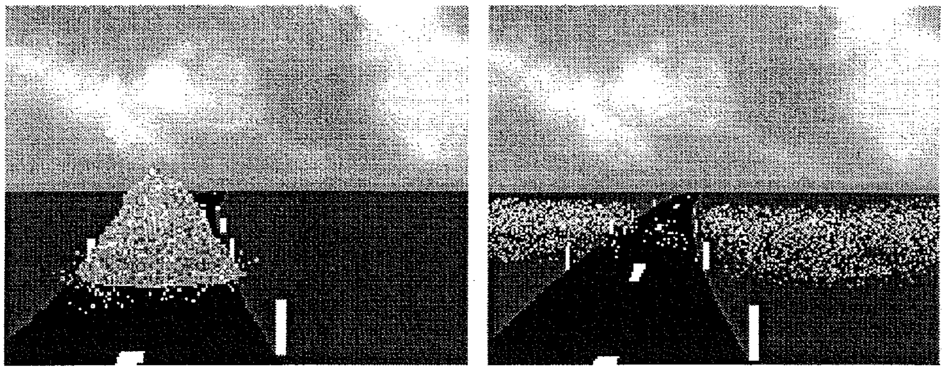
2.3.1. Problem-Solving Task

Participants were presented with a set of problems that resembled real-world situations. The task differed from the simpler matching or recognition tasks that are more commonly used to assess icon usability but was more likely to be analogous to the problem-solving interactions users have when using a new interface.

Participants were told that their job was to get from A to B as quickly as possible using a multipurpose vehicle. To do this, however, they would have to deal with a number of obstacles they would encounter along the way and would need to use the icon functions to move on. For example, participants might encounter a pile of sand in the middle of the road (see Figure 1a). To continue, participants could use one of three solutions. The sand could be either pushed, shoveled, or blown off the road using appropriate icons (see Appendix B for a full list of obstacles and their icon "function" solutions). To emulate real-world situations, the number of possible solutions to each problem varied (there were one, two, or three possible solutions). When participants employed the correct icon to deal with a problem, the obstacle would disappear from the display (see Figure 1b) and a statement would appear on the display telling them that the obstacles had been cleared (e.g., "You have cleared the sand from the road."). A new obstacle would then appear on the display.

Participants were encouraged to try to find solutions as quickly as possible to the problems that they encountered but also to explore alternatives (e.g., chopping or burning a fallen tree lying across the road). Points were awarded depending on how quickly participants solved the problems they encountered and the number of different solutions they were able to discover for each problem.

Problem-solving trials were presented in blocks of 16 because there were 16 types of obstacles that could be encountered (see Appendix B). A block of trials therefore provided participants with one chance to solve each problem. There were a total of 22 blocks of 16 trials. Participants were presented with problem-solving trials over three 1-hr sessions. In the first session, participants were introduced to the task and presented with the



(a) Sand obstacle in road

(b) Sand dispersed after appropriate icon use

FIGURE 1 Example of problems presented on the Multipurpose Vehicle Interface used in the experiment.

first block of trials. After a short break, participants then completed trial blocks 2 through 8. In the second session they completed blocks 9 through 15 and in the third session completed blocks 16 through 22.

The first block of trials differed from subsequent trials because it was the first time obstacles were encountered and participants had to guess which icons might be appropriate. In addition, because there were only 16 trials, it was only possible to provide one solution for each problem in this first block. After the first block of trials was completed, participants were told about the points scoring system because they were now able to explore alternative solutions to the problems presented. One score was given for the speed with which participants were able to solve the problem. For each problem solved, participants gained a total of 25 points, with points being deducted for each second taken. A bonus score was awarded on the basis of the number of extra solutions discovered for each problem. Bonus points were scored across a whole experimental session (i.e., trial blocks 2 through 8, 9 through 15, and 16 through 22) because only one solution could be given in each block of trials.

2.3.2. Performance Measures

Measures of performance obtained from the problem-solving task were as follows:

1. The number of correct solutions to problems at the first attempt. During the first block of learning trials this was a measure of how easy it was to guess the function of an icon when it was first encountered. In later blocks of trials, it provided an index of the extent to which participants had learned icon-function relations in previous blocks.
2. The number of attempts required for successful problem solving. The number of attempts required reflected the amount of effort required to solve problems.

3. The time taken to solve problems successfully. This measure also reflects the extent to which icon–function relations had been learned and how quickly users were able to deal with the problems they encountered as a result.

The scores participants obtained as a result of the speed at which they solved problems were not used as a measure of performance, because this was simply another measure of the time taken to solve problems successfully (i.e., measure 3). Similarly, the bonus points awarded when users correctly used alternative icon solutions to problems are not discussed further within the context of this article.

2.3.3. *Knowledge Structure Measures*

2.3.3.1. Relatedness ratings. As noted previously, participants were able to gain experience with their icon set over the course of the problem-solving trials. As they did so, data was obtained about the way in which they perceived the relations between the problems they encountered and the icons they were given to use to solve those problems. This was done by asking participants to provide ratings of the relatedness of all possible pairwise combinations of problems and icon solutions (e.g., sand–shovel, sand–blow, sand–accelerate, sheep–go round, sheep–illuminate, and so on; see Appendix B). A 7-point rating scale was employed to measure the perceived closeness of icon–function relations. Ratings of 7 indicated a very close icon–function relation whereas ratings of 1 indicated a very distant relation.

These ratings were obtained at four time points: at Time 1 after users had solved each problem once, and at Time 2 (after eight problem-solving trials), Time 3 (after 15 trials), and Time 4 (after 22 trials). If participants had an ideal understanding of the icons to use when they encountered a problem, they would produce high relatedness ratings for correct icon–function solutions and low relatedness ratings for incorrect icon–function pairings.

2.3.3.2. Evaluation of knowledge structures using pathfinder. Pathfinder analysis was used to examine changes in users' conceptual structures as they gained experience with the icon set. The analysis was carried out to derive the simplest conceptual structure from the relatedness ratings ($q = 31$, $r =$ infinity; see Schvaneveldt, 1990). These conceptual structures were compared with an ideal standard based on the correct possible solutions given in Appendix B.

Once Pathfinder networks had been obtained, they were compared to the ideal network using the C statistic, a measure of shared links for matching nodes across two different networks (Goldsmith & Davenport, 1990). This provides a measure (ranging from 0 to 1) of the degree of association between the two networks. The expectation was that as participants gained experience with the interface, the relation between the ideal network and participants' networks would increase and that this would be mediated by the nature of the interface (i.e., whether participants were using concrete, abstract, or arbitrary icons for problem solving).

3. RESULTS

3.1. Performance Measures

3.1.1. Number of Correct Solutions at the First Attempt

This provided a measure of initial icon "guessability" (in the first block of trials) and of icon–function learning (in later blocks of trials). As Figure 2 shows, the number of correct solutions increased steadily as participants became familiar with the interface and gained an understanding of the functions that icons represented. It should be noted that the pattern of participants' performance was influenced by the encouragement they were given to explore alternative solutions to problems. This produced a cyclical pattern of performance. Correct solutions were produced most frequently in the first block of trials in a session (i.e., blocks 2, 8, and 15) when participants are providing the first set of solutions to problems. This is followed by a dip in performance while participants explored other solutions (which were often incorrect) and a gradual rise toward the end of the session as participants successfully applied the new solutions they had discovered.

A mixed two-way analysis of variance (ANOVA) was carried out to examine this data further where the effects of icon set (concrete, abstract, or arbitrary) and user experience (trial blocks 1 through 22) were considered. The effect of experience was significant reflecting participants' increasing accuracy over blocks of trials, $F(21, 1113) = 21.79$, $p < .001$. There was also a small, but statistically reliable, difference in accuracy depending on the type of icon set participants used, $F(2, 53) = 3.52$, $p < .05$. Accuracy was higher when participants used concrete icons ($M = 9.79$, $SD = 1.26$), less when abstract icons were used ($M = 8.56$, $SD = 1.57$), and least when arbitrary icons were used ($M = 8.49$, $SD = 2.24$).

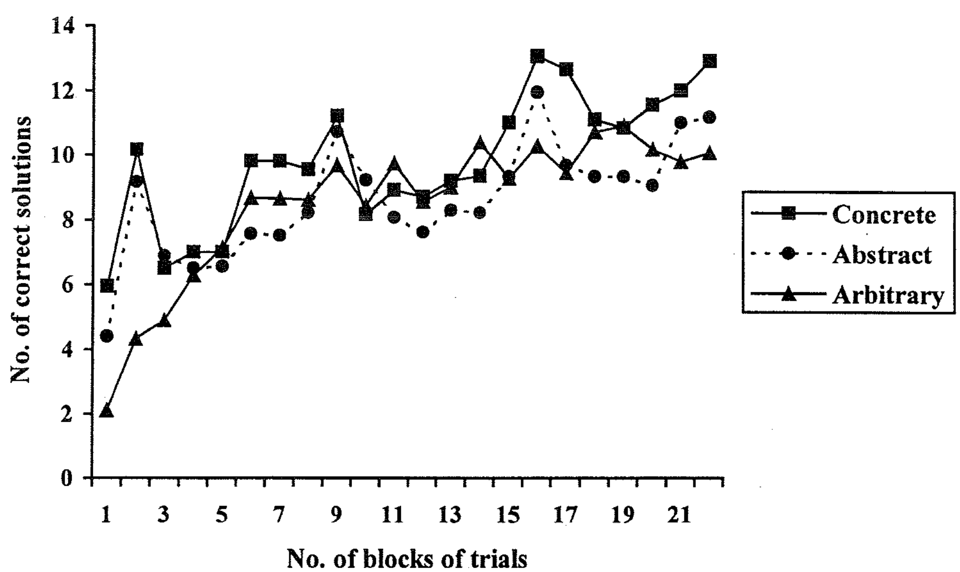


FIGURE 2 The number of correct solutions to problems at the first attempt as users gain experience with icon sets.

The interaction between experience and type of icon used was also significant, $F(42, 1113) = 2.23$, $p < .001$. This was the result of variation in the type of icon that provided the most accurate performance across blocks of trials. This interaction was examined further using analysis of simple main effects followed by Newman–Keuls comparisons. The results are summarized in Table 2. Only those blocks on which significant differences between icon conditions were found are listed. Although these show the same trend as the main effect of icon type, some exceptions to this trend were apparent, producing the interaction between experience and icon type.

3.1.2. Number of Attempts Required to Successfully Solve Problems

Figure 3 shows that the amount of effort required to find a solution reduced as participants gained experience with the interface. The number of attempts required, particularly in early trials, appeared to be determined by the type of icon set participants were asked to use. Those required to use arbitrary icons needed approximately twice as many attempts to solve problems successfully as those using concrete icons on the first block of trials. Those using abstract icons also required more attempts on the first block of trials in comparison to those using concrete icons.

A two-way ANOVA was also carried out on this data to examine the effects of experience and the type of icon set used on the number of attempts required to solve problems. Accuracy improved with experience, $F(21, 1197) = 22.18$, $p < .001$, and there were differences between icon sets, $F(2, 57) = 7.32$, $p < .001$. Those using concrete icons took least attempts to arrive at a correct solution ($M = 2.35$, $SD = 0.07$); those using abstract icons took slightly more ($M = 2.65$, $SD = 0.50$); and those using arbitrary icons took most attempts ($M = 3.25$, $SD = 0.97$). A significant interaction was also observed between experience and the type of icon set used, $F(42, 1197) = 4.65$, $p < .001$. The interaction was examined more closely using analysis of simple main effects followed by Newman–Keuls comparisons (see Table 3). The interaction appears to result from the fact that differences between icons sets are greatest during initial learning trials (see blocks 1, 2, and 3) and are attenuated in later trials.

TABLE 2
Correct Solutions to Problems: Summary of Simple Main Effects Analyses

Number of Block	*Simple Main Effect*	*Newman–Keuls Comparisons*
1	$F(2, 53) = 13.73$, $p < .001$	conc > abs > arb
2	$F(2, 53) = 22.48$, $p < .001$	conc = abs > arb
16	$F(2, 53) = 3.59$, $p < .05$	conc > arb
17	$F(2, 53) = 5.50$, $p < .01$	conc > abs = arb
20	$F(2, 53) = 3.84$, $p < .05$	conc > abs
22	$F(2, 53) = 3.64$, $p < .05$	conc > arb

Note. conc = concrete icons; abs = abstract icons; arb = arbitrary icons.

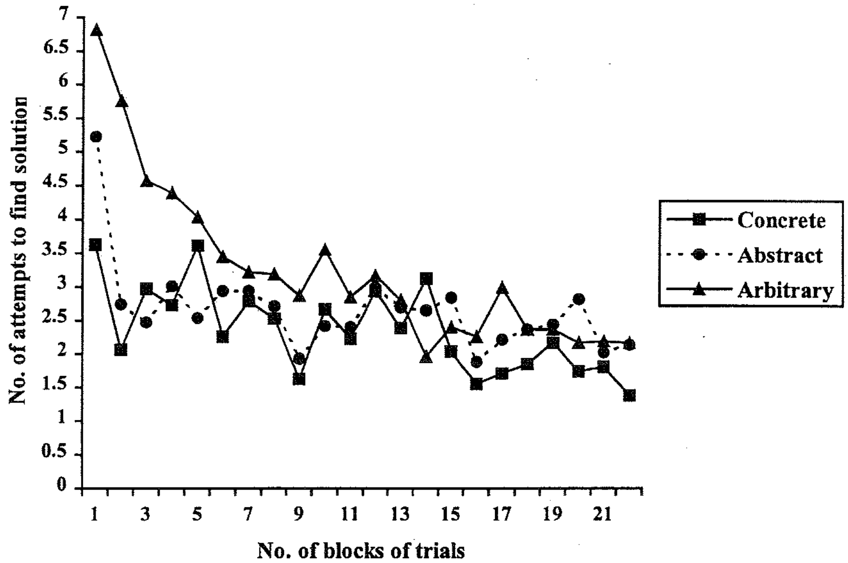


FIGURE 3 Number of attempts required to solve problems as users gain experience with icon sets.

3.1.3. Time Taken to Solve Problems

Figure 4 shows that the time taken to find solutions to problems decreased dramatically as users gained experience with icon sets. It also shows that performance again fluctuated over each experimental session. Faster times are recorded at the beginning of each session (see blocks 9 and 16) followed by a slight rise in response times as participants attempted to find other solutions to problems. Differences between icon sets were less apparent than when accuracy was considered. This may be due to the fact that, irrespective of the type of icon used, participants adopted a strategy where they quickly selected a variety of alternative solutions to discover one that was correct.

An ANOVA examining the effects of icon set and experience on response times found no significant main effect of icon set, $F(2, 57) = 0.94$. The effects of experience, $F(21, 1197) = 81.92$, $p < .001$, and the interaction between icon type and experience, $F(21, 1197) = 2.04$, $p < .001$, were, however, significant. Simple main effects analyses followed by Newman–Keuls comparisons showed that differences between icon types were apparent at block 2, $F(2, 57) = 21.00$, $p < .001$, arbitrary > concrete = abstract; and block 14, $F(2, 57) = 4.36$, $p < .05$, arbitrary > concrete = abstract.

3.1.4. Analyses of Covariance

To compare the effects of concreteness and semantic distance, identical analyses of variance were carried out on each of the performance measures in which either icon concreteness or semantic distance was entered as a covariate. When semantic distance was entered as a covariate, the main effect of condition (i.e., differences between icon sets) was no longer sig-

TABLE 3
Number of Attempts of Solve Problems: Summary of Simple Main Effects Analyses

Number of Block	*Simple Main Effect*	*Newman–Keuls Comparisons*
1	$F(2, 57) = 30.23, p < .001$	arb > abs > conc
2	$F(2, 57) = 35.32, p < .001$	arb > abs = conc
3	$F(2, 57) = 10.44, p < .001$	arb > abs = conc
4	$F(2, 57) = 6.74, p < .01$	arb > abs = conc
6	$F(2, 57) = 3.47, p < .05$	arb > conc
9	$F(2, 57) = 7.22, p < .05$	arb > abs = conc
14	$F(2, 57) = 3.17, p = .05$	conc > arb
17	$F(2, 57) = 6.30, p < .01$	arb > abs = conc
20	$F(2, 57) = 4.17, p < .05$	abs > conc
22	$F(2, 57) = 4.37, p < .05$	arb = abs > conc

Note. conc = concrete icon , abs = abstract icons, arb = arbitrary icons.

nificant for either measure of accuracy—number correct at first attempt: $F(2, 53) = 1.56, p > .05$; number of attempts required: $F(2, 57) = 0.28, p > .05$. No effect of icon set had initially been observed for the time taken to solve problems, but an interaction between experience and icon set had been observed. This was no longer apparent when semantic distance was entered as a covariate, $F(42, 1197) = 1.00, p > .05$.

When icon concreteness was entered as a covariate, the main effect of icon set remained significant for both accuracy measures—number correct at first attempt: $F(2, 53) = 2.76, p < .001$; number of attempts required: $F(2, 57) = 4.92, p < .001$—as did the interaction between experience and icon set for the time taken to solve problems, $F(42, 1197) = 2.17, p > .01$.

Taken together, these findings suggest that semantic distance exerts a greater influence on performance than icon concreteness.

3.2. Knowledge Structure Measures

3.2.1. Relatedness Ratings

Icon usability can be measured by standard measures of performance but can also be evaluated by assessing users' knowledge of the interface. This was achieved by asking participants to rate the relatedness of icons and problems (e.g., sand in road–shovel, sand in road–accelerate). The assumption was that if participants were asked to employ a usable icon set, they would be more likely to achieve a good understanding of which icons were appropriate for a given problem, and ratings between correct icon–function pairings would therefore be higher (i.e., approaching 7). Those employing a better icon set should also be more likely to perceive incorrect icon–function pairings as more distantly related (i.e., approaching 1) than those employing less usable sets. These relatedness ratings formed the basis for the Pathfinder analysis (see Section 3.2.2.). Ratings to correct and incorrect icon–function pairings were compared across each of the three experimental icon conditions.

3.2.1.1. Correct icon–function pairings. Relatedness ratings to icon–function pairings were obtained at four time points: after the first block of experimental trials and,

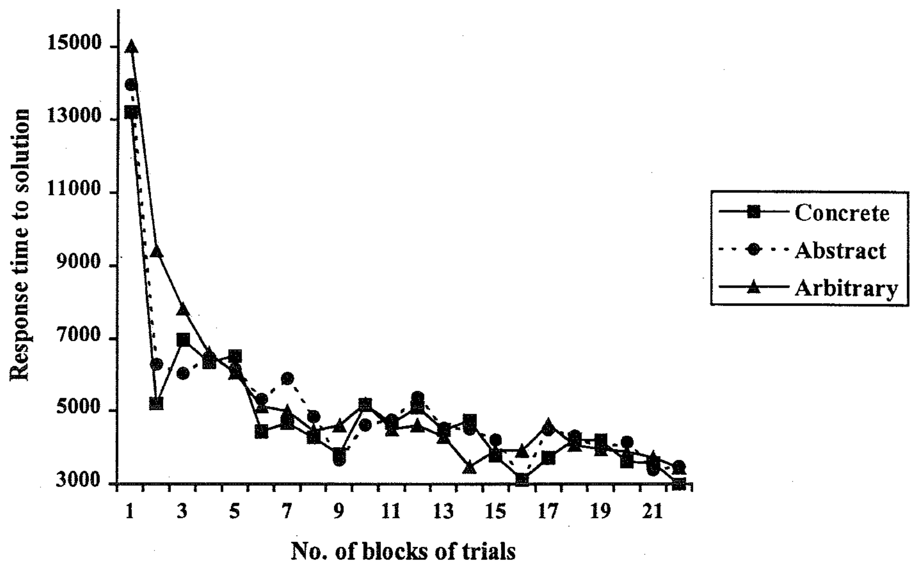


FIGURE 4 Time taken to solve problems.

subsequently, at the end of each 1-hr experimental session (i.e., after 8, 15, and 22 blocks of trials). Figure 5 shows the relatedness ratings obtained for each icon set. The relatedness ratings for correct icon–function pairs appear to be consistently higher for those using concrete and abstract icons in comparison to those using arbitrary icons.

An ANOVA was carried out to examine the effect of icon set (concrete, abstract, or arbitrary) and experience (after blocks 1, 8, 15, and 22) on relatedness ratings. This revealed significant effects of icon set, $F(2, 56) = 27.52$, $p < .001$, experience, $F(3, 84) = 92.29$, $p < .001$, and a significant interaction, $F(6, 168) = 3.37$, $p < .01$, between the two effects. Simple main effects followed by Newman–Keuls comparisons revealed that relatedness ratings for those using the concrete and abstract icons were higher at all time points than for those using the arbitrary icons (all *p*s < .05). No differences were found in ratings between those using concrete and abstract icons (all *p*s > .05).

3.2.1.2. Incorrect icon–function pairings. Data for the relatedness ratings of incorrect icon–function pairings were also obtained as a matter of course at the same four time points. It was assumed that if users had a good understanding of the interface, ratings to incorrect icon–function pairs would be lower.

An identical ANOVA was used to examine incorrect pairings as was used for correct icon pairings. Significant effects of icon type, $F(2, 452) = 25.75$, $p < .001$, and experience, $F(3, 678) = 231.02$, $p < .001$, were observed. There was also a significant interaction between these two effects, $F(6, 1356) = 21.78$, $p < .001$. Simple main effects followed by Newman–Keuls comparisons revealed a slightly more complicated pattern of findings than for correct icon–function pairings (see Figure 5b). After block 1, those using the concrete and abstract icon sets provided significantly lower ratings than those using the arbitrary icon set (*p*s < .05). After blocks 8 and 15, those using concrete icons provided lower ratings than

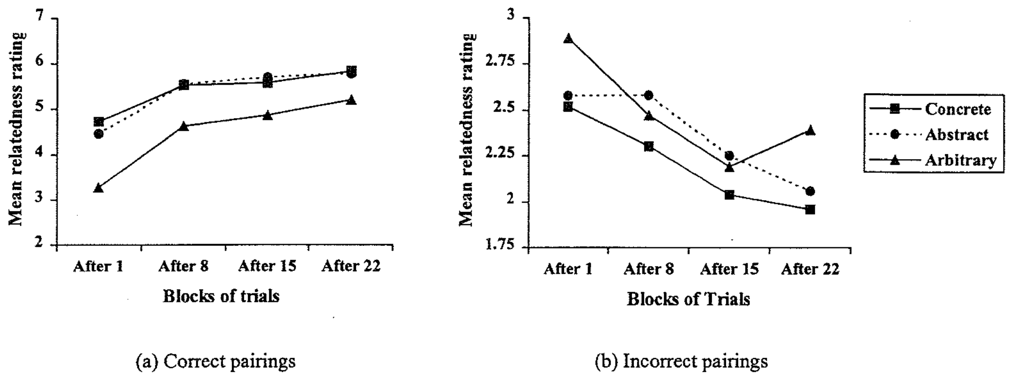


FIGURE 5 Relatedness ratings for icon–function pairings.

those using either abstract or arbitrary sets. After block 22, those using the concrete icons provided the lowest ratings, those using abstract icons provide significantly higher ratings, and those using the arbitrary set provided ratings that were significantly higher than both other groups. The general trend that can be observed in these findings is that those using concrete icons provide the lowest ratings.

When the pattern of findings for correct and incorrect icon–function pairings are considered together, they suggest that, although those using concrete and abstract icons are just as likely to perceive correct icon–function solutions, those using abstract icons appear to have a less clear conception of icon problem interrelationships because they are able to differentiate less clearly between correct and incorrect icon solutions (cf. Figures 5a and 5b).

3.2.2. Evaluation of Knowledge Structures Using Pathfinder

Pathfinder analysis was used to explore differences in users' conceptual structure as a result of the icon set they used. The Pathfinder algorithm is a structural and procedural modeling technique that extracts underlying patterns in proximity data and represents them spatially in a Pathfinder network. The relatedness ratings, which provided an estimate of the perceived proximity of an icon and a function, formed the basis of this analysis. The analysis was carried out to derive the simplest proximity structure from the relatedness ratings ($q = 31$, $r =$ infinity; see Schvaneveldt, 1990). Pathfinder solutions for users' initial conceptualizations of the relations between problems and solutions are shown in Figure 6 (after the first experimental trial) and their final understanding is shown in Figure 7 (after Trial 22). These solutions represent the understanding on average for users in each icon group.

Obstacles presented in the road (i.e., problems) are shown in the gray-filled conceptual nodes whereas possible solutions to problems (represented as icons on the interface) are shown in italics. Full details of problems and their correct solutions are shown in Appendix B. Links between problems and solutions are shown as filled lines if the link is correct (i.e., a correct problem–solution pairing) and as dashed lines if the link is incorrect.

Differences in the initial understanding between those using arbitrary icons in comparison to concrete or abstract icons are most apparent. At this stage, arbitrary icon users appear

→ Pathfinder solutions showing user groups' initial conceptualizations of the interface.

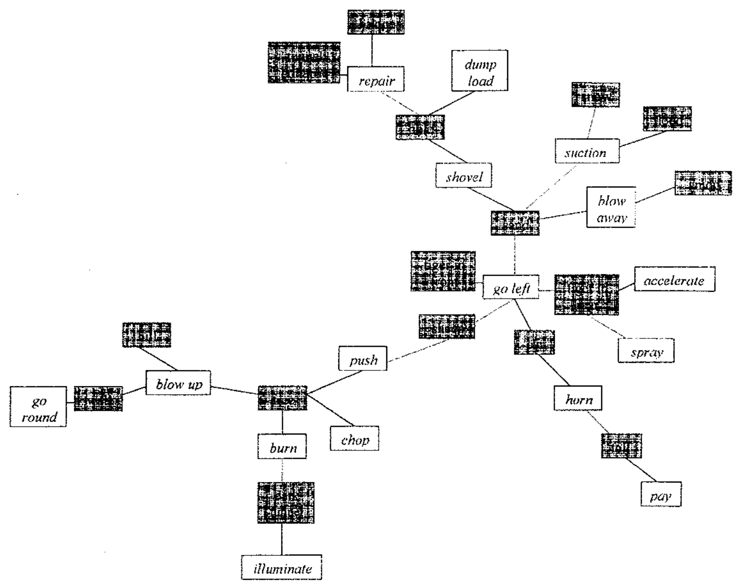


FIGURE 6a ~~Pathfinder solutions showing user groups' initial conceptualizations of the interface.~~ → Concrete ~~Items~~ Icons Time 1

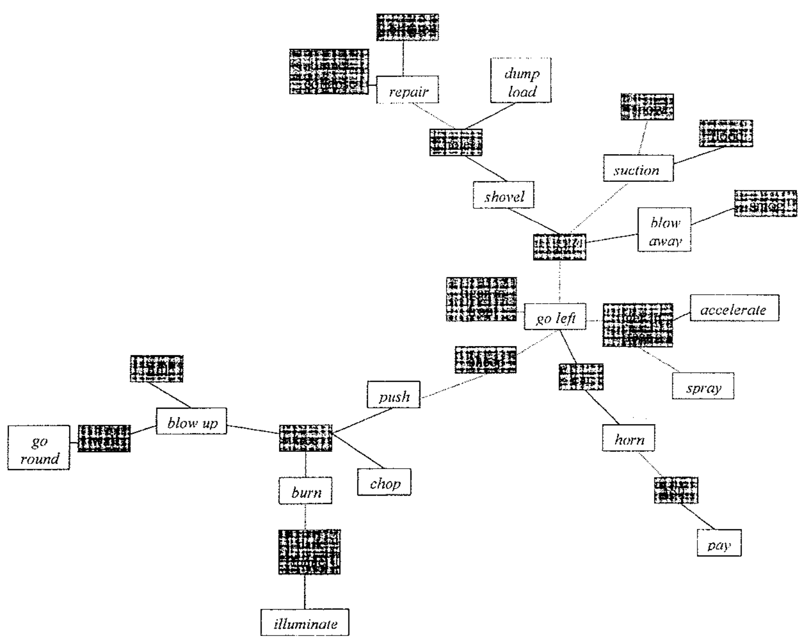


FIGURE 6b Abstract icons Time 1.

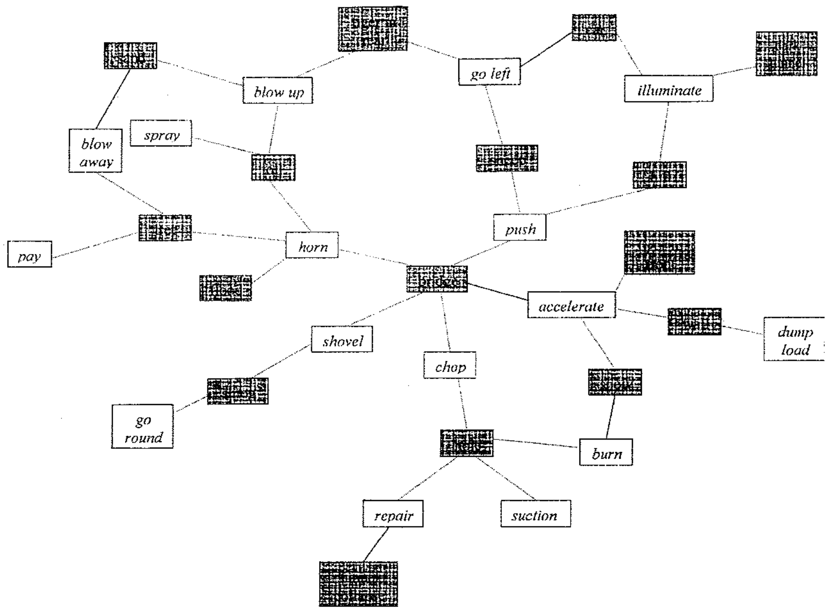


FIGURE 6c Arbitrary icons Time 1.

to have a minimal understanding of icon functions and, as a result, problem-solution pairings are mostly incorrect.

3.2.2.1. C statistic. The degree of association between these networks and the ideal standard of all possible correct icons solutions to problems (see Appendix B) was calculated using the C statistic in accordance with the following equation:

$$C(A,B) = \tfrac{1}{n}\Sigma \frac{|Av \cap Bv|}{|Av \cup Bv|}$$

This defines the similarity, C, between two Pathfinder networks, A and B, where there is a finite set V of n nodes. The similarity of a neighborhood of a given node in A is compared with the neighborhood of the same node in B. A neighborhood is defined as the set of items linked directly to a given node (e.g., the neighborhood of "repair" for concrete icons at Time 1 is 3—"bridge," "tunnel collapse," and "hole"). The index of similarity for a common node in two graphs is the intersection of the node neighborhoods divided by the union of the neighborhoods. Overall graph similarity is the mean of these n values (for further details see Goldsmith & Davenport, 1990).

Even at Time 1, there was a considerable degree of commonality between the ideal standard and networks for concrete and abstract icons users ($C = 0.60$ and 0.59, respectively). However, there was little association between arbitrary icons and the ideal network ($C = 0.15$). At Time 4 there were high levels of association between all icon networks and the ideal, C(concrete) = 0.79; C(abstract) = 0.82, and C(arbitrary) = 0.77.

Pathfinder solutions showing user groups' final conceptualizations of the interface.

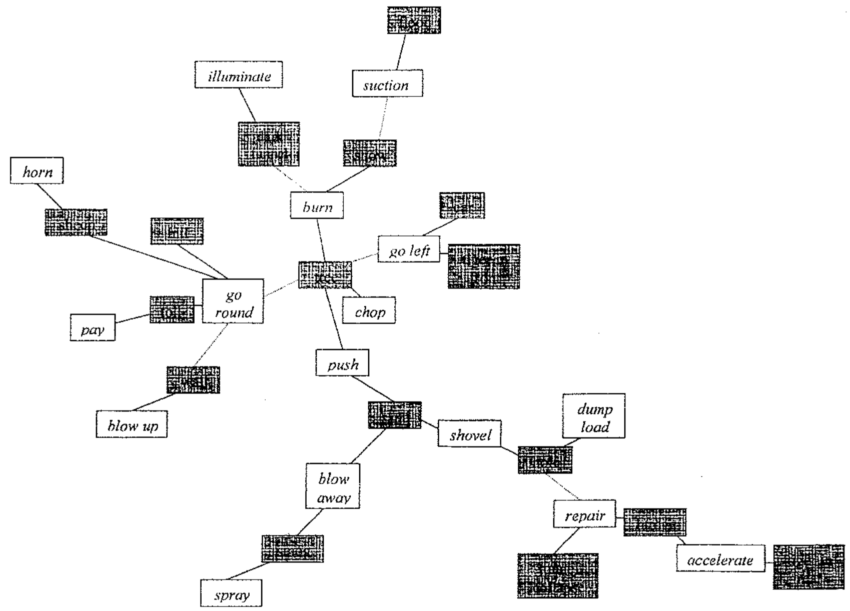


FIGURE 7a ~~Pathfinder solutions showing~~ user groups' final conceptualizations ~~of the interface~~.
Concrete icons Time 4

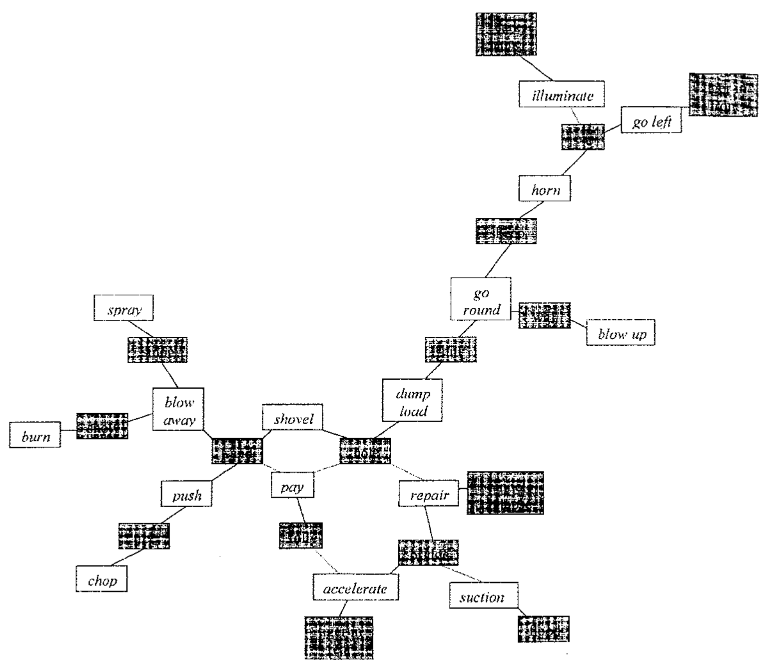


FIGURE 7b Abstract icons Time 4.

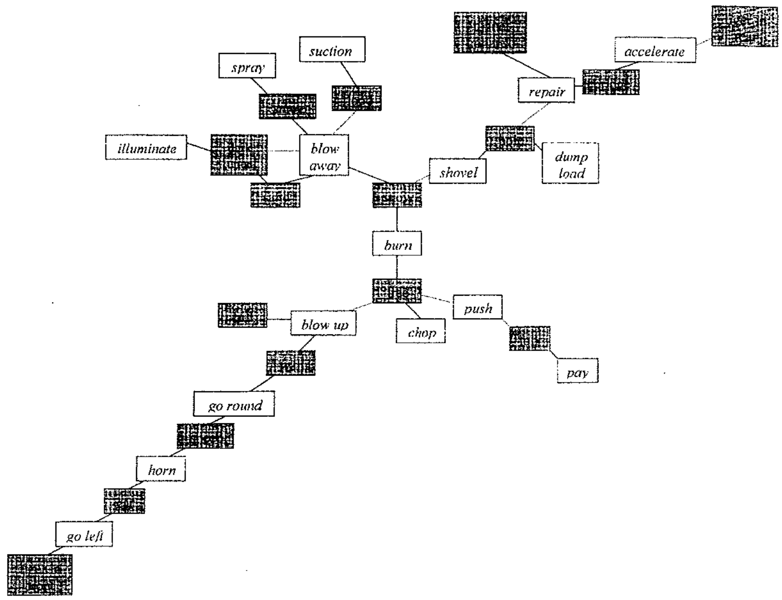


FIGURE 7c Arbitrary icons Time 4.

A two-way ANOVA was carried out to examine these differences further. The main effect of Time (1 vs. 4) was significant, $F(1, 32) = 56.49$, $p < .011$, as was the effect of icon type, $F(1, 31) = 29.24$, $p < .001$. Most important, the Time X icon type interaction was also significant, $F(1, 32) = 18.36$, $p < .001$. This interaction is the result of arbitrary icon users' poor understanding of the interface at Time 1.

3.2.2.2. Qualitative analysis. Figures 6 and 7 show that, despite having little experience of the interface, users of concrete and abstract icons gained a very rapid understanding of icon–function relations. Where erroneous links are made, they are plausible in some cases but not in others. In both concrete and abstract Pathfinder solutions, a plausible link was made between a hole in the road ("hole") and "repair" as a solution. When abstract icons were used, "accelerate" was given as a solution to seven problems, which included accelerating through a wall, a sheep, and a toll booth. This suggests that users will make what they feel are plausible links between icons and solutions wherever possible and will sometimes overgeneralize the use of icons they have found successful to many problems even when these are not plausible.

By Time 4, the arbitrary icon user group had arrived at a much better understanding of the interface. However, some misunderstandings remained for all three user groups. All groups maintained the erroneous assumption that "repair" is appropriate solution to a hole in the road. This suggests that considerable corrective feedback would be required for users to change their assumptions about this plausible icon–function relation. In addition, some solutions still appear to be overgeneralized and used to solve more problems than they were designed to (e.g., "go round" in the concrete icon set and "blow away" in the arbitrary icon set). This is most apparent for problems with multiple solutions (see Appendix B), which sometimes appear to elicit far too many solutions (e.g., "tree in the road" in the concrete and arbitrary icon sets and "hole in the road" in the abstract icon set).

These figures illustrate the way in which Pathfinder solutions can help designers to identify users' misconceptions about the interface. Because these solutions represent the averaged view across a user group, individual viewpoints are much less likely to influence the design process.

4. DISCUSSION

4.1. Knowledge Structure Measures

Our findings lend support to the view that the underlying mental structures users develop as they interact with a system are significantly affected by the way in which information is presented to them. If different visual information is used to represent the same functions, then different mental structures are produced.

Pathfinder networks provide a rich source of information for the designer who wishes to identify precisely where users are failing to learn correct icon–function links. In both the initial and later stages of learning, the particular difficulties with each icon set could be easily seen in the networks. Pathfinder networks are particularly useful for such qualitative analyses because they are based on the average responses of a group of users and are therefore less likely to be biased by individual idiosyncratic responses. They also highlight users' search for any plausible links between icons and functions. For example, users consistently made a link between a hole in the road and "repair" as a plausible solution. Although this was not regarded as "correct" in the interface we presented to participants, designers would be able to capitalize on their awareness of any unexpected links users were making between icon and function when creating the next iteration of an interface. By the same token, when users fail to link an icon with an appropriate function, this could also act as a prompt for further design work.

Before using Pathfinder analysis for interface evaluation, however, it is important to assess the costs involved. The amount of effort in obtaining relatedness ratings, or other proximity data, for Pathfinder analysis is considerable. Obtaining ratings even for the simple interface used in this study involved participants rating hundreds of possible icon–function pairs at each time point where data was gathered. If one wishes, as we did, to examine users' knowledge structures after they had gained experience with the interface, then there is also the time commitment involved in allowing users to gain experience. With existing interfaces, however, groups of novices and experts may be compared (see Schvaneveldt et al., 1985; Wilson, 1994).

Other, more "economical" alternatives might be considered. These include asking users to name icons (i.e., produce function labels) or to match icons to functions (Nolan, 1989). For a more fine-grained analysis, users can be asked rate the comprehensibility of an icon (Rogers, 1986). Although these methods provide an measure of users' understanding of individual icon–function relations, they do not provide an overview of their understanding of the interface as a whole in the way that Pathfinder networks do so effectively. Other methods such as card sorting or categorization tasks, where users are asked to sort cards with icons or function labels on them into similarity clusters, can go some way to providing designers with some of the information available from Pathfinder analysis (Neilsen, 1994). The disadvantage of these tasks is that they do not provide the "at-a-glance" overview of interface understanding provided via Pathfinder networks. In addition, if the designer feels that users verbal reports are unlikely to be helpful (e.g., because as novices they do not understand the interface or as experts they no longer have enough insight into their use of the system; see Staggers & Norcio, 1993), then use of Pathfinder networks will provide much needed information about how the interface is being used.

The relatedness ratings that form the basis for the Pathfinder algorithm can be useful in their own right. Consideration of incorrect icon–function relatedness ratings revealed that those using abstract icons appear to have a "fuzzier" conception of icon–function relations when compared to those using concrete icons. This was not apparent in the Pathfinder networks. Interestingly, evidence from memory research suggests that conceptualization of abstract words may also be less clear. It has been argued that concrete words, referring to objects and people (e.g., *flower, computer*) are more likely to be recalled than abstract words (e.g., *theory, justice*). This is thought to be because concrete words are more definitive in their conceptualization than abstract words such as *theory* and *justice*, which are less easy to define (Marschark, Richman, Yuille, & Hunt, 1987). It may be that similar principles apply to concrete and abstract icons.

4.2. Performance Measures

Consistent differences in users' performance emerged as a result of the icon set they were asked to use and the experience they had gained with the interface. Users' accuracy, the effort required to solve problems, and the time that they took to do so all improved to asymptote after approximately seven blocks of trials (see Figures 2, 3, and 4). In general, performance was best for concrete icons and worst for arbitrary icons. Performance of participants using abstract icons tended to follow those using concrete icons more closely than those using arbitrary icons. However, differences in performance between icon sets tended to be short-lived, and once users had gained a little experience with the interface, performance differences between concrete and abstract icon users, and eventually even between concrete and arbitrary icon users, disappeared.

These findings suggest that concrete icons are most likely to be useful where users require an instant understanding of icons such as in public information signs or where icons are used infrequently (e.g., warnings). The efficacy of concrete icons in these contexts appears to stem from the close relations between icon and function (i.e., low semantic distance) at least as much as from the pictorial information they may contain.

4.3. Implications for the Use of the Visual Metaphor

The design of the modern graphical user interface often rests on the implicit assumption that the use of the visual metaphor carries considerable advantages. The visual metaphor presumes that depicting icons from the real world increases the strength of the relation between an icon and its function. As a result, icon concreteness is thought to be closely related to the semantic distance between an icon and its function. In our study we attempted to dissociate these two properties because it is evident that, in practice, abstract icons are often able to provide stronger icon–referent relations than is possible with concrete representations. In this study, both abstract and arbitrary icons were equally abstract, but abstract icons had closer icon–referent relations than arbitrary icons (see Appendix A). Because the concreteness of abstract and arbitrary icons were equally low, one might have expected users to have encountered difficulties with both of these icon sets if concreteness was the primary determinant of icon usability. On all measures, however, performance with abstract icons was better than for arbitrary icons. This suggests that the closeness of the icon–function relation, rather than the concreteness of the icons, is the most important determinant of icon comprehensibility and user performance.

The findings for concrete icons are less clear because these icons differed from the other icon sets on both concreteness and semantic distance (see Table 1). If we assume that measures of concreteness and semantic distance vary along a continuum (as suggested by McDougall et al., 1999), then there is every reason to expect that similar pattern of findings to those observed for abstract icons if comparisons were to be made between a set of arbitrary concrete icons and the concrete set included in this study. Further investigation, however, is required to obtain conclusive evidence.

Distinguishing an image from the image–function relation is not something that has been previously addressed in the icon literature. Its importance lies in the fact that it carries significant implications for the decision-making process involved in identifying icons suitable for interface design. If semantic distance, rather than concreteness, is the primary determinant of comprehensibility and performance, this challenges the assumption that metaphors on the interface need to be primarily concrete in nature. Although renditions of real-world items may help users in their initial encounters with an interface, forming strong systematic relations between icons and functions is more important, particularly where there are no pictorial alternatives for a given icon function.

ACKNOWLEDGMENTS

This research was part of the Effective Symbology Project supported by a grant from BAE SYSTEMS (SRC/UOS/060495).

A large proportion of the symbols printed in Appendix A were created specifically for this experiment or are within public domain and are not copyrighted. Other icons are reprinted with permission of the British Standards Institute and the North Atlantic Treaty Organization.

REFERENCES

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory and Cognition, 9,* 422–433.

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

Barsalou, L. M. (1989). Intra-concept similarity and its implications for inter-concept similarity. In S. Vosnaidou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge, England: Cambridge University Press.

Blankenberger, S., & Hahn, K. (1991). Effects of icon design on human–computer interaction. *International Journal of Man–Machine Studies, 35,* 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,* 11–44.

Byrne, M. D. (1993). Using icons to find documents: Simplicity is critical. *Proceedings of INTERCHI '93,* pp. 446–453.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human–computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chen, C. (1998). Generalized similarity analysis and pathfinder network scaling. *Interacting with Computers, 10,* 107–128.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Christ, R. E. (1975). Review and analysis of color coding. Research for visual displays. *Human Factors, 17,* 542–570.

Christ, R. E., & Corso, G. (1982). The effects of extended practice on the evaluation of visual display codes. *Human Factors, 25,* 71–84.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Cooke, N. J., Stout, R., Rivera, K., & Salas, E. (1998). Exploring measures of team knowledge. *Proceedings of the Human Factors and Ergonomics Society 42nd annual meeting* (pp. 305–309). Santa Monica, CA: Human Factors.

de Groot, A. D. (1965). *Thought and choices in chess.* The Hague, Netherlands: Mouton.

~~Erickson~~, A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press. Ericsson

Familant, E. M., & Detweiler, M. C. (1993). Iconic reference: evolving perspective and an organizing framework. *International Journal of Man–Machine Studies, 39,* 705–728.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12,* 395–~~247~~. 427

Gillan, D. J., Breedin, S. D., & Cooke, N. J. (1992). Network and multidimensional representations of the declarative knowledge of human–computer interface design experts. *Internal Journal of Man–Machine Studies, 36,* 587–615.

Goh, S. K., & Coury, B. G. (1994). Incorporating the effect of display formats in cognitive modeling. *Ergonomics, 37,* 725–745.

Goldsmith, T. E., & Davenport, D. M. (1990). Assessing structural similarity of graphs. In R. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 75– 87). Norword, NJ: Ablex.

Goldsmith, T. E., & Johnson, P. J. (1990). A structural assessment of classroom learning. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 241–254). Norwood, NJ: Ablex.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83,* 88–96.

Gomez, R., Hadfield, O. D., & Housner, L. (in press). Conceptual maps and simulated teaching episodes as indicators of competence in teaching elementary mathematics. *Journal of Educational Psychology.*

Gomez, R., Schvaneveldt, R. W., & Staudenmayer, H. (in press). Assessing beliefs about "environmental illness/multiple chemical sensitivity." *Journal of Health Psychology.*

Gonzalvo, P., Canas, J. J., & Bajo, M.-T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86,* 601–616.

Gray, P. D., Waite, K. W., & Draper, S. (1990). Do-it-yourself iconic displays: Reconfigurable iconic representations of application objects. In D. Diaper, D. Gilmore, G. Cockton, & B. Shackel (Eds.), *Human–computer interaction, INTERACT '90.* (p~~p. xx–xx~~). Amsterdam: Elsevier. 473 – 478

Green, A. J. K., & Barnard, P. J. (1990). Iconic interfacing: The role of icon distinctiveness and fixed or variable screen locations. In D. Diaper, D. Gilmore, G. Cockton, & B. Shackel (Eds.), *Human–computer interaction, INTERACT '90* (pp.~~xx–xx~~). Amsterdam: Elsevier. 457 – 462

~~Keiras~~, D. (1978). Beyond pictures and words: Alternative information processing models for imagery effects in verbal memory. *Psychological Bulletin, 85,* 532–544. Kieras

McDougall, S. J. P., Curry, M. B., & de Bruijn, O. (1999). Measuring icon and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity and semantic distance for 239 icons. *Behavior Research Methods, Computers & Instrumentation, 31,* 487–519.

McDougall, S. J. P., de Bruijn, O., & Curry, M. B. (2000). Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity and distinctiveness. *Journal of Experimental Psychology: Applied, 6,* 291–306.

McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology, 13,* 307–325.

Meyer, D. E., & Schvaneveldt, R. W. (1976). Meaning, memory structure, and mental processes. *Science, 192,* 27–33.

Microsoft Corporation. (1997). *Microsoft Visual Basic 5.0 Language Reference.* ~~XXXXX: Author~~ Redmond WA: Microsoft.

Morphew, M. E., Thordsen, M. L., & Klein, G. (1998). The development of cognitive analysis methods to aid interface design. In *Proceedings of the Human Factors and Ergonomics Society 42nd annual meeting* (pp. 305–309). Santa Monica, CA: Human Factors Society.

Neilsen, J. (1994). *Usability engineering.* Boston: Academic.

Nolan, P. R. (1989). Designing screen icons: Ranking and matching studies. In *Proceedings of the Human Factors Society 33rd annual meeting* (pp. 380–384). Santa Monica, CA: Human Factors Society.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement, 76,* 1–25.

Quillian, M. R. (1969). The teachable language comprehender. *Communications of the Association for Computing Machinery, 12,* 459–475.

Rogers, Y. (1986). Evaluating the meaningfulness of icon sets to represent command operations. In M. D. Harrison & A. F. Monk (Eds.), *People and computers: Designing for usability* (pp. ~~xx–xx~~ 586–598). Cambridge, England: Cambridge University Press.

Rogers, Y. (1989). Icon design for the user interface. *International Review of Ergonomics, 2,* 129–154.

Rowe, A. L., Cooke, N. J., Hall, E. P., & Halgren, T. L. (1996). Toward an on-line knowledge assessment methodology: Building on the relationship between knowing and doing. *Journal of Experimental Psychology: Applied, 2,* 31–47.

Saarilouma, P. (1994). Location coding in chess. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47A,* 607–630.

Schonfield, A., & Herrmann, D. (1982). Problem perception and knowledge structures in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning Memory and Cognition, 8,* 484–494.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., & Cooke, N. J. (1985). Measuring the structure of expertise. *International Journal of Man–Machine Studies, 23,* 699–728.

Smith, D. C., Irby, C., Kimball, R., & Verplank, B. (1982). Designing the STAR user interface. *Byte, 7,* 242–282.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174–215.

Staggers, N., & Norcio, A. F. (1993). Mental models: Concepts for human–computer interaction research. *International Journal for Man–Machine Studies, 38,* 587–605.

Stammers, R. B., & Hoffman, J. (1991). Transfer between icon sets and ratings of icon concreteness and appropriateness. In *Proceedings of the Human Factors Society 35th annual meeting,* (pp. 354–458). Santa Monica, CA: Human Factors Society.

Tullis, T. (1981). An evaluation of alphanumeric, graphic, and color information displays. *Human Factors, 23,* 541–550.

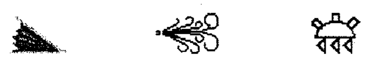
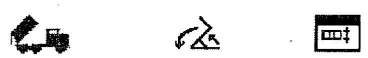
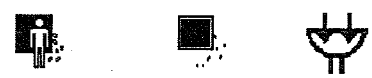
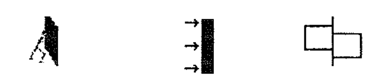
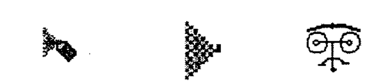
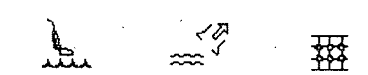
Tullis, T. S. (1983). The formatting of alphanumeric displays: A review and analysis. *Human Factors, 25,* 657–682.

Wilson, J. M. (1994). Network representations of knowledge about chemical equilibrium: Variations with achievement. *Journal of Research in Science Teaching, 31,* 1133–1147.

Ye, N. (1997). Objective and consistent analysis of group differences in knowledge representation. *International Journal of Cognitive Ergonomics, 1,* 169–187.

APPENDIX A

Icons and Functions for Pictorial, Abstract, and Arbitrary Icon Sets

	Pictorial Abstract Arbitrary
Accelerate	
Blow	
Blow up	
Burn	
Chop	
Dump load	
Go around	
Go left	
Illuminate	
Pay	
Push	
Repair	
Shovel	
Sound Horn	
Spray	
Suction	

APPENDIX B

List of Obstacles and "Functional" Solutions

Obstacle	Solution 1	Solution 2	Solution 3
Tree in road	'Push' off road	'Burn'	'Chop' up
Heap of sand in road	'Push' sand off road	'Shovel' sand off road	'Blow' sand off road
Hill in road	'Go round'	'Blow up'	'Dump' heavy load to get up hill more easily
Wall across road	'Go round'	'Blow up'	
Hole in road	Fill in with 'Shovel'	'Dump' load into hole	
Broken road bridge	'Repair' damage	'Accelerate' to get across damage	
Snow on road	'Burn' to melt snow	'Blow' snow off road	
Smog reduces visibility on road	'Blow' smog away	'Spray' smog away	
Sheep on road	'Go round'	Sound 'Horn'	
Car on road	Sound 'Horn'	'Go left'	
Damaged road tunnel	'Repair'		
Flood across road	Use 'Suction' to clear water		
In dark tunnel	'Illuminate'		
Road toll booth	'Pay'		
Tiger in front of car	'Go left'		
Tiger at the rear of car	'Accelerate'		