approach to consistency taken by other vendors such as Apple and DEC.

The final recommended action is to continue to seek vendors' feedback regarding consistency. The vendors realize that consistency can play a part in helping their customers reduce training and support costs. They appreciate IBM's willingness to seek their feedback regarding consistency issues. The effort can be included (at little cost) as part of the early ship program and through continuing work with ISVs.

About the Authors

Dr. Alan Happ works as an Advisory Human Factors Engineer in the Design Center at the Entry System Division of IBM US in Boca Raton, Florida. He joined IBM in 1982 and has worked on human computer interaction issues for personal and mid-range business systems. His research interests focus on the human factors required for the design and development of products that meet the users needs. He received M.A. and Ph.D. degrees in Experimental Psychology from Miami University, Ohio.

Dr. Karen C. Cohen has been at MIT for 15 years. She is currently Principal Research Associate at the Center for Cognitive Science and Project Athena. Known primarily for her work in academic and industry on evaluating the impact of computers and learning, she has published eight books and over 100 articles and monographs in the field. She has consulted to IBM on several projects. She received her B.A., Magna Cum Laude, in Social Relations from Harvard University and her M.A. and Ph.D. in Social Relations and Educational Psychology at The John Hopkins University.

The study was sponsored by the User Interface Technology Dept., Entry Systems Division, IBM U.S., Boca Raton, Florida. Special thanks are extended to H. Dulaney, manager of UIT, and Doretha Lippett, lead for the task force on consistency.

AN EMPIRICAL APPROACH TO THE EVALUATION OF ICONS

JAYSON M. WEBB, PAUL F. SORENSON, NIC P. LYONS

General Abstract

This poster provides a definition and taxonomy for iconic communication and describes the use of formal psychological tools and methods in the evaluation of icons. The methods that can be usefully applied include:

1. Psychophysics
2. Scaling
3. Recognition/Memory Testing
4. Statistical Modeling / Analysis

Examples of some of these approaches are provided from pilot studies currently under way at HP. Analyses used include Multi-Dimensional Scaling (MDS) and Cluster analysis. Results can be applied to development of metrics, standard methods, and design guidelines.

Detailed Abstract

Definition and Taxonomy

An icon is a pictogram which can be selected or otherwise interacted with by the user of a system interface, and which represents one or more of the following:

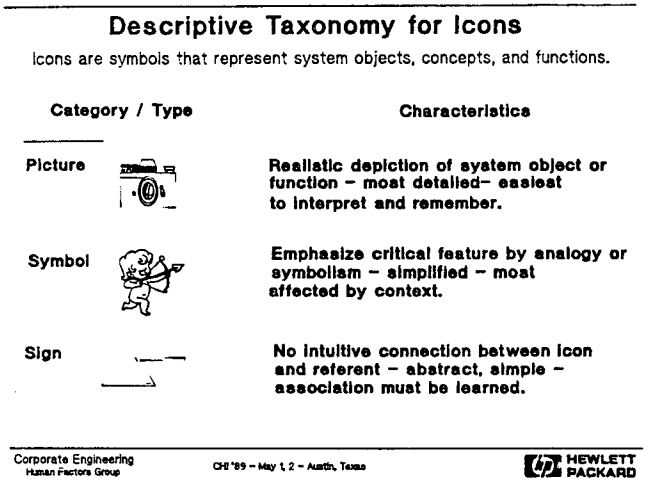
The functions of the computer system, The system objects upon which these functions act, Certain types of system status.

The user interacts with Icons in several ways, including: Selecting (Activation using mouse, or other input device), Moving, Copying, and Deleting.

Types of Icons

There are 3 types of icon (see 1st Figure), each of which conveys its meaning in a different way:

Figure 1



Pictorial: Realistic depiction of system object or function. Reference by resemblance. Have the most detail, are the most concrete, easiest to interpret and remember.

Symbolic: Depicts a critical feature of the referent object or function through analogy or symbolism. Reference by symbolism. Representation is simplified - most affected by context of presentation (e.g., system metaphor employed).

Sign: No inherent, intuitive connection exists between the icon and its referent. Relationship between icon and system object or function must be learned by rote. Reference by learned association. Simplest, and most abstract.

Interactive Attributes

1. Detectability (in a crowd, distinguishability)
2. Legibility
3. Interpretability

4. Recognizability
5. Preference

Characteristics of icons that may have an effect upon these attributes include: *Size, Contrast, Color* (for Detectability), *Complexity, Concreteness, Dynamism* (the extent to which an icon represents an object or an action), *Icon-Type, System Context*, and the user's *Past Experience* with icons.

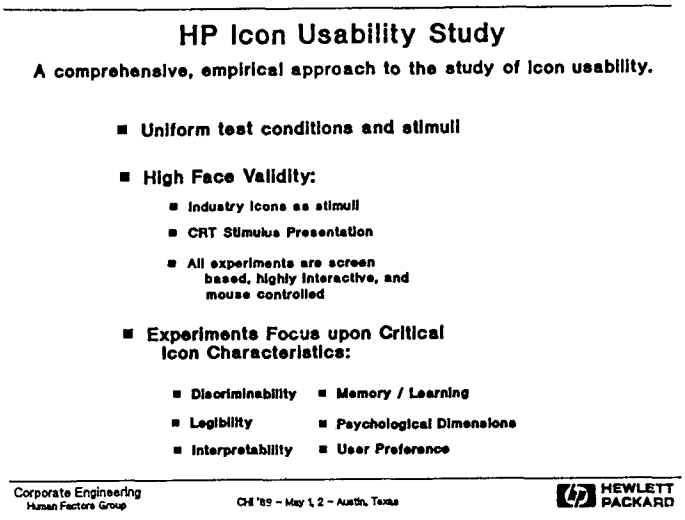
In order to discover which of these characteristics most affects user's performance with and preference for icons, controlled experiments must be conducted. This entails quantifying the characteristics, and measuring performance and preference for icons that vary in the characteristics of interest. Quantifying the characteristics in this way amounts to assigning numerical values according to "how much" of each characteristic the icon "has". For example, complexity can be quantified in many ways, including Number of component shapes, number of component angles, subsymmetry measures, symmenetropy measures, relative amount of contour, grouping, and closure.

Stimuli to be used in the experiments are analyzed to determine their values for each of the characteristics to be measured. These data can then be compared (using regression analysis) to subjects' scale ratings, performance scores, similarity judgements, and preference scores. Of particular interest are predictive relationships between these characteristics and subjects' results. These can be used as the basis of predictive metrics, standard evaluation tools, and design guidelines for icon usability and applicability for different applications.

Icon Attributes and Experimental Approach

The second Figure provides an overview of the experimental program on icons currently underway at Hewlett Packard Corporate Engineering. The series of experiments was designed to allow comparison of data across experiments, a coherent and realistic environment and presentation method, and to provide baseline comparisons across companies on both preference and performance measures. Icons from Xerox, Apple, HP, Ricoh, Toshiba, Interleaf and X-Windows were used as stimuli.

Figure 2



Detectability / Discriminability

Matrix Detection was chosen to assess this attribute of icons. While this implementation focuses more strongly on Discriminability *per se*, vision and perception literature provide ample models and data with which to predict psychophysical detectability for stimuli like icons. Stimulus characteristics like size, color, contrast, and so on adequately account for performance in such tasks. Of more direct interest to software engineers is how well icons can be discriminated from others displayed on the screen, and what characteristics of an icon's design might predict this.

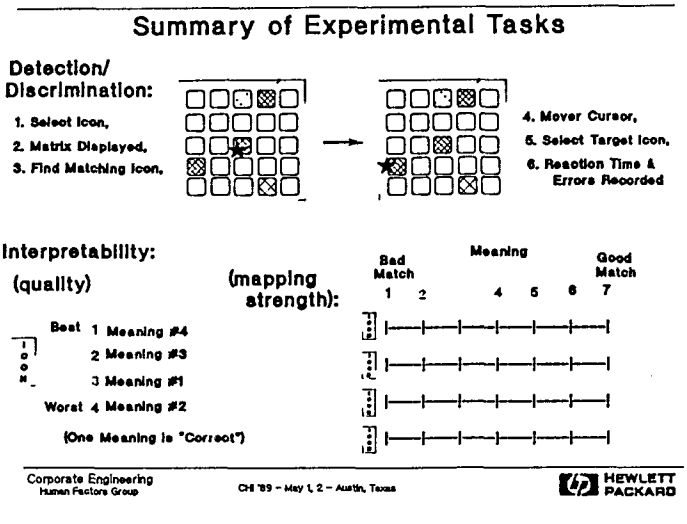
Legibility (not shown in Figures)

Refers to the perceptual quality of an icon's structural features. Legibility is central to an icon's capability to convey meaning. Characteristics that effect legibility are based upon principles of Gestal Psychology, including contrast boundary, continuity, closure and simplicity of figure. The task used to assess this must address not only whether subjects can correctly deduce the intended meaning of an icon (strictly speaking, this is interpretability), but what figural characteristics of the icon influence those deductions. When an error in guessing an icons' meaning is made, why is it made? Are subjects able to identify the structural components of the icon, but mistakenly apply the wrong metaphor to get to the meaning - or are the figural elements of the icon themselves misleading and the source of the error ?

Interpretability

Two tasks are shown in the 3rd Figure that address differing, but related aspects of interpretability. The first task is to rank-order 4 potential meanings for each icon (or 4 potential icons for each system function - both experiments are being performed, though only the first is shown). The second task is to rate the quality of the match between each meaning and the test icon (or, in the other case, the match of each icon to the meaning). Here, the subject is making an absolute judgement, while in the first experiment, it is a relative one. We have found that people can easily rank order meanings that they think are all bad matches to a given icon. Both types of information are useful.

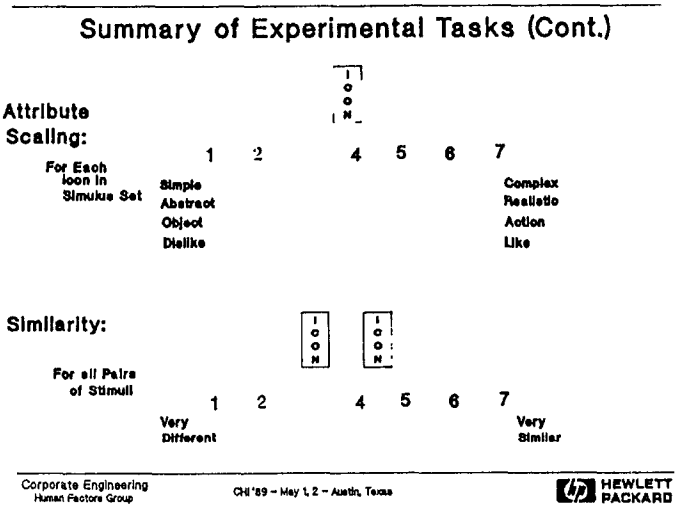
Figure 3



Attribute Scaling (4th Figure)

Each icon is rated along four 7 point scales, each representing a characteristic of icons. Data from this experiment will be used to construct user-type profiles, and to investigate the Multi-dimensional Spaces obtained in the Similarity judgement tasks (see below). Guidelines for icon design may be developed based upon these results as well, if clear differences on these judgements are found across user-type. For example, there are a number of analytical metrics for complexity that can be applied to the stimuli and the results correlated with the subjects' responses on the scale. If one of these analytical measures can be found to predict subjects' assessments of complexity, predictive guidelines can be written detailing how complex icons should be for different user groups, product lines, etc.

Figure 4



Similarity Judgements

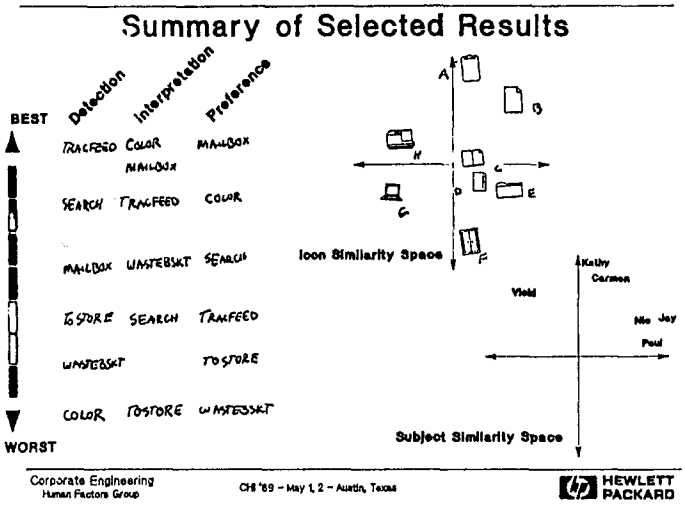
Judgements of the similarity of icons provide insight into how subjects think about and organize icons. The different "natural" attributes of icons according to which subjects categorize them are revealed in the analysis of this kind of data. Having subjects indicate how similar or different pairs of icons are on a seven point scale is one experimental technique used to collect similarity judgements (called the Paired Comparison Technique, PST). Scale values are transformed to euclidian distances, and the MDS and clustering analyses applied to map out subjects' "cognitive spaces" for icons. Such models depict the dimensions (attributes) that subjects use to think about icons as axes in an N-Dimensional space.

Another experimental technique often used to get similarity data is the Multiple Sort Technique (MST). In the application of the MST, subjects physically group stimuli on the basis of similarity. They can form as many groups as they desire from a set of stimuli, and they can use as many different bases of similarity as they want, performing one sort for each different basis. The data is then analyzed using MDS. The MDS analysis scales the similarity judgements along several dimensions. Again, each dimension reflects some salient aspect of the stimuli.

RESULTS

Representative icons (see Figure 5) from the stimuli used in our pilot testing are shown along a continuum from "Best" to "Worst" performance (detection and interpretation tasks) and preference. Note that the same icon appears in different relative positions depending upon the experiment being considered. These results suggest that speed of detection and interpretability are not highly positively correlated with subjects' preference.

Figure 5



Indeed, regression analyses of these preliminary data show r-squared values of 0.031 for detection speed with preference, and 0.12 for interpretation quality. The only attribute of icons in the pilot study that reliably predicted preference for icons was *Realism*: The more pictorially realistic an icon is, the higher its preference rating (r-squared = 0.78). This is not surprising, but it is reassuring to find our experimental tools accurately reflecting our intuitions about the world.

Individual Differences in MDS Spaces

Icon Space

Two different methods were used to qualify differences between individual MDS solutions in this study: cluster analysis and Individual Differences Multidimensional Scaling (INDSCAL).

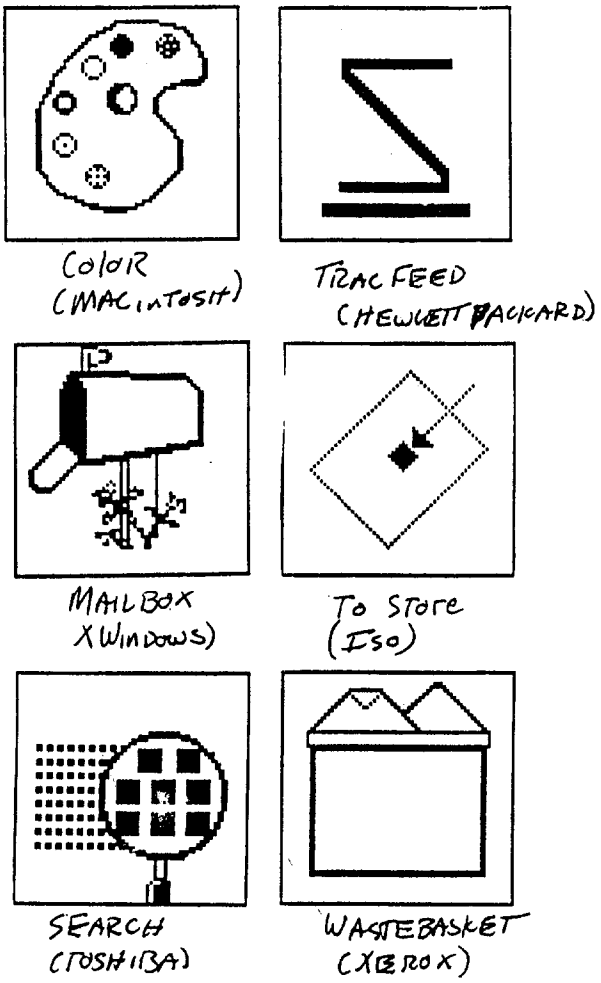
As noted earlier, subjects' similarity scores have to be transformed into distance scores for MDS analysis. The next step is to turn everyone's raw judgement data into Z-scores and average the data in each cluster. These average matrices are then each analyzed using KYST. The qualitative differences between them can be "eyeballed" and decisions can be based on that. The Figure shows an icon space from one of our subjects.

Dimension 1 (horizontal axis) separates the two icons that depict electronic devices from the icons that depict non-electronic devices. This is the most salient distinction

among the icons for this subject. The second dimension (vertical axis) scales the icons according to what appears to be a "multiplicity" factor. That is, icons are scaled along a continuum of increasing volume of paper that they contain or are made up of.

Further analyses to be performed on these data include regressing the attribute dimensions through the icon space to see if any high correlations can be found. This can often lead to identification of salient, *measurable and predictable* dimensions in the way subjects think about and use icons.

Subject Space



Data representing the individual icon spaces (as just described) can be considered as variables and analyzed themselves. This results in a "Subject Space", where similarities across subjects' icons spaces are revealed. For this sample, three groups are observable. This Figure is a plot of dimensions 1 and 2, but a total of 4 dimensions were salient for the subjects in this experiment. All four stimulus dimensions were interpretable. Subject 1 (Paul) loaded mostly on dimensions 1 and 4. Subject 2 (Jay) loaded mostly on dimension 1. Subjects 3 (Kathy) and 4 (Carmen) loaded on 2, and Subject 5 (Nic) loaded on dimensions 1 and 3.

Dimension 1 is the electronic vs. non-electronic dimension that we saw earlier in Paul's KYST solution. Dimension 2 scales the icons according to shape with wide things at one end of the scale and tall things at the other. Dimension 3 is the most difficult to interpret, but it appears to be a temporal dimension: The subject who loaded on this dimension reported that he thought the scanner was a printer. So, a progression can be seen that starts with the terminal, moves through the various icons, and ends up at the "printer".

Summary

Our analysis of the preliminary data from this set of pilot experiments indicates that these experimental methods are likely to provide a great deal of useful information about how our user population perceives, thinks about, and uses icons. The full experimental study is currently being run on subjects with a wide variety of computer backgrounds. Topics for future consideration using these experimental tools include the effects of color on discriminability and interpretability, the effects of context on icon interpretation, and whether icons truly do represent an "international Language".

BEACONS AND INITIAL PROGRAM COMPREHENSION

SUSAN WIEDENBECK, JEAN SCHOLTZ

Beacons are any surface features which typically occur in a program and strongly point to the program's function. In a sort the swap is a beacon because it is typically present and is the prime operation which moves the problem toward the goal of a sorted list. Our objective was to establish a causal connection between the presence of beacons and comprehension of high-level program function.

For our first experiment we developed two correct Shellsort programs which differed only in that one version contained a standard beacon-like swapping sequence, while the other disguised the swap so that it no longer was in its typical beacon form. Twenty novice and 20 advanced programmers studied each program then did two tasks: 1) described the program's function and 2) recalled the program.

Subjects were more accurate in determining the function of the non-disguised version than the disguised version ($F(1,76) = 7.03$, $p = .010$). Advanced subjects were better at determining program function than novices ($F(1,76) = 7.03$, $p = .010$). However, advanced subjects' performance was not aided more by the presence of the beacon than was novice subjects' performance (interaction of expertise and program version: $F(1,76) = 2.85$, $p = .096$). The recall measure showed a non-significant trend for subjects to recall the swap lines in the non-disguised program better than those in the disguised version.