Int. J. Human-Computer Studies (2000) **53,** 637–662
doi:10.1006/ijhc.2000.0412
Available online at http://www.idealibrary.com on IDEAL®

Evaluating visualizations: using a taxonomic guide

E. Morse† and M. Lewis

School of Information Sciences, University of Pittsburgh, 135 N. Bellefield St. Pittsburgh, PA 15260, USA. email: morse@nist.gov, ml@sis.pitt.edu.

K. A. Olsen

Box 308, N-6401, Molde College, Norway. email: kai.olsen@molde.no.

(Received 11 January 2000, and accepted in revised form 31 May 2000)

Although visualizations are components of many information interfaces, testing of these visual elements is rarely undertaken except as a part of overall usability testing. For this reason, it is unclear what role, if any, visualizations actually perform. Our method involves the creation of simple visual prototypes and task sets based on a visual taxonomy which allows testing of the visualization in isolation from the rest of the system. By defining tests using a visual taxonomy rather than customary tasks from the application domain, our method circumvents the problems of restricting evaluation of newer more capable systems to only those tasks which might be accomplished with older, less capable ones. This paper will discuss methods for exhaustively testing the capabilities of a visualization by mapping from a domain-independent taxonomy of visual tasks to a specific domain, i.e. information retrieval. Experimental results are presented illustrating this approach to determining the role visualizations may play in supporting users in information-seeking environments. Our methods could easily be extended to other domains including data visualization.

1. Introduction

Researchers in information retrieval (IR) have long searched for ways to make their systems more accessible to end-users and to develop new ways for users to explore data. Visualization techniques (computer methods for displaying large quantities of information graphically) appear promising as a means for achieving both goals. Information visualization can show multidimensional relationships that are difficult to extract from tabular data. However, unlike scientific visualizations, which are largely developed and used within elite specialties, IR visualizations are targeted toward guiding the public through newly accessible oceans of on-line information.

Users employ many strategies when engaged in information seeking, including bibliographical search, analytical search, search by analogy, empirical search, browsing and check routine (Pejtersen, 1988). Each of these activities might be augmented by using

† Current address: National Institute of Standards and Technology, 100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899, USA. email: morse@nist.gov.

1071-5819/00/110637 + 26 $35.00/0

visualizations, but browsing and analytical search are the strategies cited most frequently as benefiting from visual support (Marchionini, 1995). According to Lin (1997), browsing is a superior strategy when (1) there is good underlying structure so that items close to one another can be inferred to be similar, (2) users are unfamiliar with the contents of a collection, (3) users have limited understanding of how a system is organized and prefer the less cognitively loaded method of exploration, (4) users have difficulty verbalizing their underlying information need and (5) information is easier to recognize than to describe.

Work on IR visualization systems is at a relatively early stage. In the past 10 years, systems such as *Bead* (Chalmers, 1996), *InfoCrystal* (Spoerri, 1993), and *LyberWorld* (Hemmje, Kunkel & Willet, 1994), have been developed as visual information exploration tools to aid in retrieval tasks. Researchers at the University of Pittsburgh have contributed to the development of information visualization systems with VIBE (Olsen, Korfhage, Spring, Sochats & Williams, 1993), GUIDO (Nuchprayoon, 1996), and BIRD Kim & Korfhage, 1994).

IR visualization systems portray semantic information about documents through visible relations. Each type of visualization, by its inherent structure, emphasizes particular attributes and supports a particular set of visual tasks/inferences. In TileBars (Hearst, 1995), for example, a document retrieved through a multi-term query is represented by a bar divided by paragraphs along the x-axis and query terms in the y-axis (Figure 1). Each query term tile is shaded according to how well the paragraph matches the query term. By glancing at a document's TileBar a user can see which terms of the query are best represented, which sections of the document are most relevant, and the distribution and coincidence of topics throughout the document. By comparing TileBars between documents the user can make judgments involving the relative organization and treatment of topics by documents within the set.

In *VIBE* the degree of match between a document and query term is represented spatially instead of by shading (Figure 2). *VIBE* represents query terms as moveable circles with documents as variously sized rectangles suspended between them. The position of each document is determined by its weighting with respect to each of the terms, with higher weights causing the document to be located nearer that term. Thus, the display can show the relations between a document and each term of a multi-term query simultaneously. This is seen from the example in Figure 2 where the countries of the world are positioned with regard to six different query terms.

While TileBars makes the user visually aggregate shaded rectangles to find the strength of a document"s relation to a term, in *VIBE* this is revealed directly by the distance between the two. Other judgments such as lack of relation to a query term, which can be determined by inspection in TileBars, may require the user to rely on emergent features such as documents falling along a line between two terms in VIBE. Other IR visualization systems rely on additional visual relations to convey the degree of matching between query terms and documents. To restrict comparisons to common features or to treat current system capabilities as the standard for comparisons unfairly discriminates against newer more capable systems. In a turn of the century comparison of this sort between horses and cars, for example, the horse would win hands down by being able to plough fields and travel over irregular terrain yet today the car and tractor have replaced the horse for everything but pleasure. Our approach overcomes these

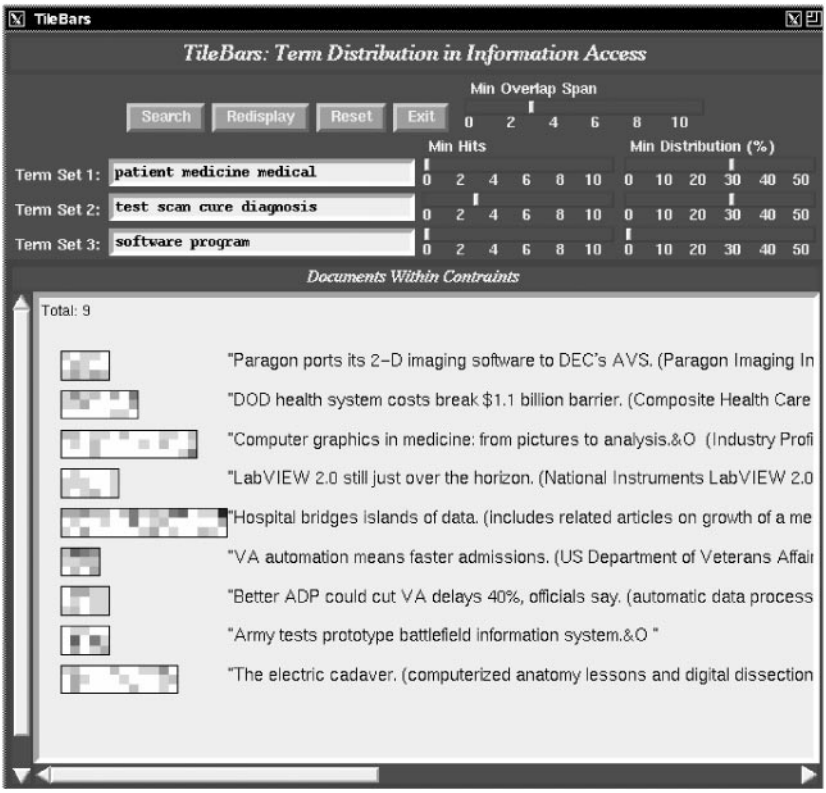


FIGURE 1. TileBars.

difficulties by considering the full range of tasks which might be performed. By casting this wide net, both customary tasks and novel tasks unsupported by current systems can be evaluated.

To investigate the usability of a visualization a researcher must demonstrate: (1) subject's ability to perform *visual tasks* such as aggregating shaded tiles or identifying clusters and (2) subject's ability to relate these visual features to problems in the task (IR) domain. IR visualization tools contain a variety of features and functions for retrieving, accessing and displaying text as well as manipulating visualizations. Confounding and interaction among multiple features is a recurring problem in evaluating intact systems. An earlier evaluation of VIBE (Koshman, 1997), for example, yielded equivocal results because of this variety of alternative explanations. The objective of our research is not to evaluate or compare extant IR visualization systems but instead to establish the degrees of usability of the visualizations on which they depend. We first pursued the idea by "defeaturing" interfaces so that users could learn the remaining core functions quickly (Morse & Lewis, 1997). Our preliminary usability evaluations demonstrated that training problems, comprehension problems, performance problems and "ratings" problems could be diminished for the de-featured interfaces.

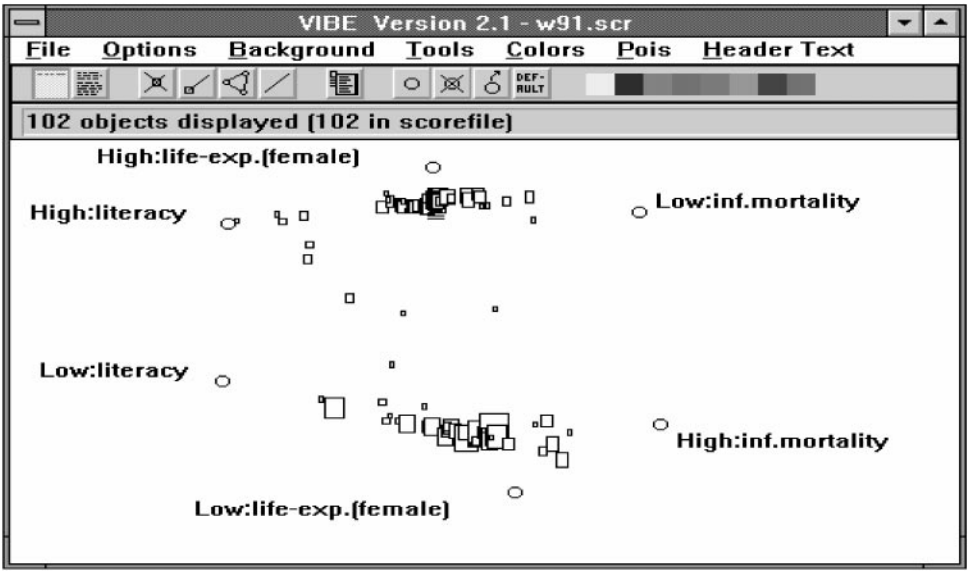


FIGURE 2. VIBE display.

An even more rigorous de-featuring is possible by eliminating features and functions until only the visualization remains. This "back to basics" approach guides the overall development of the current research including the preliminary studies based on Boolean data. By starting with the simplest instances, it becomes possible to test increasingly more complex situations and compare results within a level of difficulty as well as across levels. Our studies began by investigating 2- and 3-term Boolean queries. In Boolean retrieval, documents are scored for the presence or absence of a term so that they can be unambiguously placed in a small number of categories. A visualization for Boolean retrieval, therefore, defines some "meaningful" organization for the possible combinations of query terms. The 3-term query, (A, B, C), for example would require bins to accommodate documents containing one: (A), (B), or (C), two: (AB), (AC), (BC) or three: (ABC) of the query terms.

The next level of difficulty is the vector representation of documents, which underlies most modern retrieval systems. Documents in these systems are represented as vectors of term occurrences. Often adjustments of various types are applied to the vectors to account for document length or other factors, but each document is characterized as a collection of numeric values. A visualization for vector-based retrieval requires allocating a visible dimension to each of the query terms. Vector-based visualizations can show users all the documents that contain any of the terms and which documents have which terms just as Boolean visualization do, but in addition they also provide information about the weighting of the terms.

The results of the Boolean studies clarified the factors that influence the ability of subjects to make appropriate inferences from various presentations of document data. These factors include the type of display, the type of task, the order in which the display

types are presented, and the number of key terms embedded in the displays. In addition to measuring the ability of subjects to perform in the test environment, subjects were also probed for their preferences regarding the various displays.

2. Visual and domain tasks

Visual interfaces for information retrieval have taken two basic forms: *map systems* which attempt to organize the entire document set with respect to its dominant clusters and *dimensions and reference point systems* which organize a subset of documents with respect to a query. Morse (1999) identified nine IR visualization systems using reference points. Of these, only VIBE had been subjected to extensive user testing (Koshman, 1997). A test set of four visualization types accounting for seven of these nine systems was chosen for these experiments (Table 1). The library literature is replete with high-level models of user's strategies; however, there is no unanimity as to the set of elemental tasks required. The visual task taxonomy used to design these experiments defines a superset of elemental tasks encompassing those identified from a variety of IR strategy models.

From prior work, therefore, we find that a majority of reference point-based IR visualization systems use one of four basic visualizations, that minimal user-based evaluation of these systems has been performed and that the selected taxonomy of visual tasks subsumes the elemental tasks we identified in a review of IR user models. Our approach is to compare representative visualization types across tasks and difficulties defined by numbers of reference points (2 or 3) and scaling (Boolean or vector) to characterize their performance.

Several frameworks for information visualization have been proposed (Wehrend & Lewis, 1990; Rogowitz & Treinish, 1993; Kennedy, Mitchell & Barclay, 1996). Because we wanted to evaluate the effectiveness of visualization types for IR tasks we adopted a user-centered design approach to defining tasks. Applying this approach to visual displays we needed a taxonomy of tasks such as "distinguishing entities" which defined perceptual acts performed by a user rather than a data-oriented visualization taxonomy which might do things such as match levels of measurement to compatible display dimensions. By knowing the data that exist, the requirements of the interface,

TABLE 1
Visualization types

Visualization Type	System
Word	Ordered text such as search engine output
Icon list	TileBars, Cougar
Graph (Cartesian)	GUIDO, BIRD, InfoCrystal, Component State
Spring (physical analogue)	VIBE

and the goals of the user, it becomes possible to document visualization systems precisely.

2.1. WEHREND AND LEWIS

The task classification of Wehrend and Lewis (1990) is a low-level, domain-independent taxonomy of tasks that users might perform in a visual environment. Domain-independence allows generalizability. The Wehrend and Lewis classification consists of the following set of user actions.

- *Locate* this action can be applied to dependent as well as to independent variables. It covers interaction techniques that allow the user to find special data entries. Annotation techniques are covered by this action, for example an arrow marking the most interesting point of the display. Locate can also work like a filter, e.g. by highlighting data items that lie in a special range. Locate includes search for an object that the user already knows about.
- *Identify* identify is similar to Locate, but in this case, the user is being asked to describe an object that was not necessarily known previously.
- *Distinguish* this action allows distinguishing between different values of the same variable, e.g. for a user to know which objects have already been identified or interacted with. The interface might show different iconic representations for each object type.
- *Categorize* to define divisions that displayed objects can be sorted by. Examples of VIBE (Olsen *et al*., 1993) tasks that are categorizations are: (1) to define all the regions of a 3-POI (point of interest) Boolean display and (2) draw boundaries in a vector VIBE 3-POI display for each of the possible Boolean combinations of terms.
- *Cluster* the cluster task covers techniques that allow us to determine whether data entries are clustered or not. The ambiguity introduced by flattening the hyperdimensional spaces into two dimensions would be probed by this activity. It includes finding gaps in the display field (cluster of nothing).
- *Distribution* the distribution action is closely related to cluster in much the same way that locate and identify are related. To distribute, the user needs to describe the overall pattern while cluster merely asks that the set be detected.
- *Rank* ranking is only possible for scalar and ordinal data. Users could be asked to indicate the best and worst cases in a display. Since nominal data cannot be ranked, it is important that displays of nominal data be designed so that the user is discouraged from trying to perform such actions.
- *Compare within entities* this action describes tasks in which a user is called upon to decide something based on the attributes of similar objects.
- *Compare between relations* when different entities are used as the basis of comparison, the "compare between relations" operator is used. For instance, if a set of objects has been marked as seen and the remainder of the set is unseen, then the user might compare and contrast attributes of the sets.
- *Associate* the associate action calls upon the user to form relationships between objects in a display.
- *Correlate* if objects in a display have multiple attributes, it should be possible to discern which other objects share attributes. For instance, in a scatterplot in which the

marks have shape and color as well as their x and y position, the objects should be groupable by any of the attributes.

2.2. ZHOU AND FEINER

Zhou and Feiner (1998) have developed a visual task taxonomy. This taxonomy extends that of Wehrend and Lewis (1990) by defining additional tasks, by parameterizing the tasks, and by developing a set of dimensions by which the tasks can be grouped.

The major dimensions of visual tasks that they describe are visual accomplishments and visual implications:

> Visual accomplishments describe the type of presentation intents that a visual might help to achieve, while visual implications specify a particular type of visual action that a visual task may carry out. Zhou and Feiner (1998)

The structure that results from applying the visual accomplishments dimension is a hierarchy. The major branches describe tasks that "Enable" and tasks that "Inform". The former is further decomposed into exploration tasks and compute tasks, while the later is described as elaborate and summarize tasks. The breakdown along the line of visual implications seems that it might be useful in developing domain-dependent tasks. Zhou and Feiner propose three types of implications: (1) visual organization, (2) visual signaling and (3) visual transformations. The overall structure of the implication dimension of the visual taxonomy is shown in Table 2.

TABLE 2
Visual implications and related elemental tasks (*from Zhou and Feiner*, 1998)

Implication	Type	Subtype	Elemental tasks
Organization	Visual grouping	Proximity	Associate, cluster, locate
		Similarity	Categorize, cluster, distinguish
		Continuity	Associate, locate, reveal
		Closure	Cluster, locate, outline
	Visual attention		Cluster, distinguish, emphasize, locate
	Visual sequence		Emphasize, identify, rank
	Visual composition		Associate, correlate, identify, reveal
Signaling	Structuring		Tabulate, plot, structure, trace, map
	Encoding		Label, symbolize, portray, quantify
Transformation	Modification		Emphasize, generalize, reveal
	Transition		Switch

3. Methodology

The dependent measures of performance are number of correct answers and time-to-completion of a task set, where a set refers to all the tasks for a single display type. The measure of preference is the user"s rankings of each display. The independent variables are display type, order of presentation, individual task and scenario difficulty. Scenario difficulty is defined as the number of terms depicted in a display, i.e. 2- or 3-term. Subjects will perform the experimental tasks with a single level of difficulty.

Hundered and nintyfive subjects undertook the study. These were randomized to receive either the 2- or 3-term experimental study. Both studies were performed using the web. When a subject first accessed the URL for the study, he was randomized to one of the experimental conditions. Then specific instructions were shown for the first display type, followed by 10 pages that contained a question, a "Submit" button and a display configuration. The process of instruction and 10 displays was repeated for each display to be evaluated. A post-test questionnaire captured demographic and preference information.

The document set that was used for creating the various displays used in this study was selected from the AP 1989 newswire document set of the TREC collection. Raw frequency counts of word stems were gathered, while using a 443-term stop-list. The resulting list of $> 18\,000$ terms was trimmed by removing words that occurred in $> 95\%$ or $< 15\%$ of the documents. The remaining terms were subjected to term discrimination value analysis (Willett, 1985) using the cosine measure.

The displays that were used for this study are named "word", "icon", "table", "graph" and "spring" displays. The preliminary Boolean studies showed that prototype displays had varying abilities to support user task performance. The prototypes were based on plain text, tables, list of iconic representations, and, for 2-term studies, a graphical display (Figure 3). In addition, a display was created based on the VIBE positional formula (Olsen, Williams, Sochats & Hirtle 1992; Oslen *et al*., 1993) for portraying a set of documents in a space defined by a set of key terms (Figure 4).

Since VIBE positioning is based on the characteristics of the physics of a spring, the prototype has been referred to as a "spring" display. In two dimensions, i.e. when two key terms are chosen for display, the picture shows a line and in three dimensions, a triangle is shown. These elemental types can be used with minor modification to visualize data that has weighted or vector characteristics. The "graph" display was only tested in the 2-term condition so that the problems of 3-D displays could be avoided. Figure 5 shows typical examples of each of the chosen types for the 3-term test.

4. Tasks

The primary method presented is the procedure for mapping from the visual taxonomy to information retrieval domain (Figure 6). Elemental visual tasks were chosen from the taxonomy of Zhou and Feiner (1998). The full taxonomy contains approximately 50 tasks. In order to create a test that could be taken within a target of 1 h, it was necessary to prune the task tree. The rationale applied to the pruning was as follows.

- To sample as broadly as possible rather than deeply.
- To select tasks whose parameter lists varied significantly.

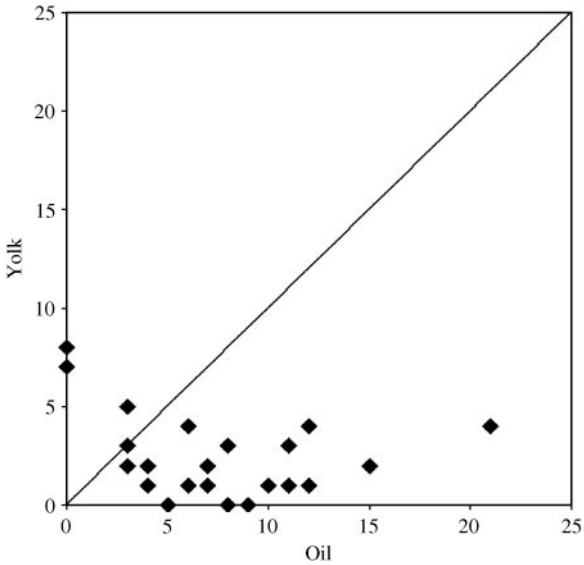


FIGURE 3. Graphical display.

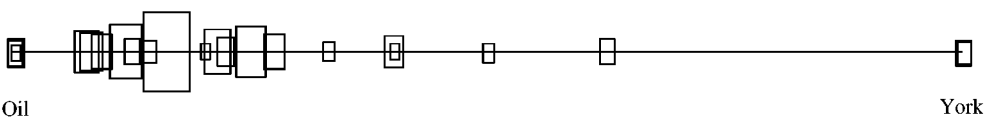


FIGURE 4. 2-term VIBE display.

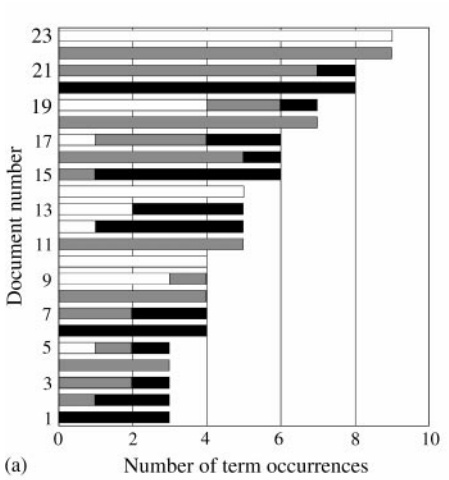
The result of the selection process is shown in Table 3, where the boldface entries are the tasks that were selected for inclusion.

Table 4 demonstrates the first stage in mapping from the above taxonomic categories to actual question or task formulations. These statements are quite general since specific document data was not considered.

Tables 5 and 6 show examples of the actual questions presented to the subjects. Table 5 refers to the 2-term test and Table 6 to the 3-term test.

The final step in creating and characterizing the tasks was to determine the number of parameters that were involved in each task instantiation. These data are presented in Table 7. The Compare, Associate, Distinguish, Locate and Identify tasks require only two parameters as defined in Table 3. In each Ranking task, subjects were asked to rank three documents according to a single criterion — a total of four parameters. Clusters contained three documents in the 2-term study and four documents in the 3-term study. Correlation and Categorization required judgments across the entire document set presented in the display, leading to parameter lists of the same cardinality as the size of the document set.

Performance was calculated by two measures—time to complete an activity and correctness of the answer. Data based on time was shown suitable for analysis by



1. Earthquake earthquake earthquake earthquake earthquake earthquake earthquake earthquake earthquake
2. Death death death death death death death death death
3. Death death death death death death death california
4. California california california california california california california california california
5. Earthquake earthquake earthquake earthquake death death california
6. Death death death death death death death death
7. Earthquake death death death california california
8. Death death death death death california
9. Death california california california california california
10. Earthquake earthquake earthquake earthquake earthquake
11. Earthquake earthquake california california california
12. Earthquake california california california california
13. Death death death death death
14. Earthquake earthquake earthquake earthquake
15. Earthquake earthquake earthquake death
16. Death death death death
17. Death death california california
18. California california california california california
19. Earthquake death california
20. Death death death
21. Death death california
22. Death california california
23. California california california

(b)

Doc#	Earthquake	Death	California	Total
1	9	0	0	9
2	0	9	0	9
3	0	7	1	8
4	0	0	8	8
5	4	2	1	7
6	0	7	0	7
7	1	3	2	6
8	0	5	1	6
9	0	1	5	6
10	5	0	0	5
11	2	0	3	5
12	1	0	4	5

(c)

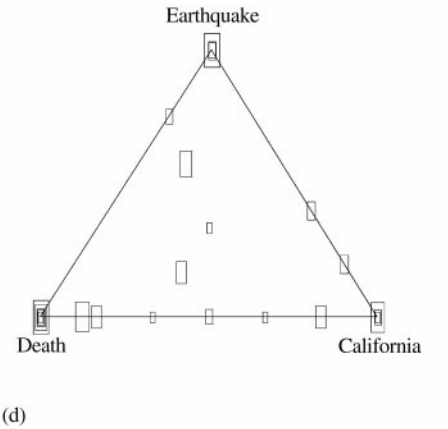


FIGURE 5. Examples of the icon (a), text (b), table (c), and spring (d) displays used in the 3-term study. □ Earthquake; ▒ Death; ■ California.

parametric statistical methods. When answers were pooled to form total scores for a particular display (10 answers) or for a whole experimental session (40 or 50 answers), the criteria for using parametric methods was met. When individual answers were inspected, however, the value was binary (right or wrong); this situation called for non-parametric methods.

5. Results

5.1. SUBJECTS

Demographic information collected in this study included gender, current educational level, area of study at a broad categorization of physical science, social science or

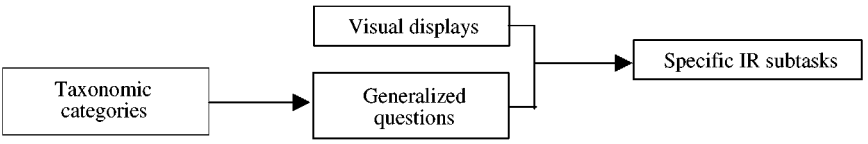


FIGURE 6. Generating experimental tasks.

TABLE 3
Comparison of taxonomic categories

Wehrend and Lewis (1990)	**Zhou and Feiner (1998)**
Associate	**Associate** $\langle ?x, ?y \rangle$
Correlate	**Correlate** $\langle ?x1, \ldots, ?xn \rangle$
Locate	**Locate** $\langle ?x, ?\text{locator} \rangle$
Distinguish	**Distinguish** $\langle ?x, ?y \rangle$
Rank	**Rank** $\langle ?x1, \ldots, ?xn, ?\text{attr} \rangle$
Categorize	**Categorize** $\langle ?x1, \ldots, ?xn \rangle$
Cluster	**Cluster** $\langle ?\text{cluster}, \ldots, ?xn \rangle$
Compare within entities	Compare $\langle ?x, ?y \rangle$
Compare between relations	
Identity	Identity $\langle x, ?\text{identifier} \rangle$
Distribution	Encode $\langle ?x \rangle$
	Background $\langle ?x, ?\text{bkg} \rangle$
	Emphasize $\langle ?x, ?x\text{-part} \rangle$
	Reveal $\langle ?\text{x}, ?x\text{-part} \rangle$
	Generalize $\langle ?x1, \ldots, ?xn \rangle$
	Switch $\langle ?x, ?y \rangle$

humanities, and native language. There were no statistically significant differences between the studies for any of these variables. The mean age of subjects in the 2- and 3-term studies was 23.2 and 23.6 years, respectively. Median ages were 20 and 21.

The data regarding the amount of time that the subject uses a computer currently, the length of time that she has used a computer and a self-assessment of her skill level show that the 2- and 3-term groups were well matched and that there were no statistically significant differences between the groups. The average subject uses a computer on a daily basis and has been using computer technology for over five years. By-and-large the subjects are computer-literate and, therefore, the on-line format of these studies should have been familiar to most of them

5.2. TIME TO COMPLETION

The results of the analysis of time to completion with respect to display type are shown graphically in Figure 7.

There are several important observations that can be made upon inspecting the data. First, for the 2-term study, there are significant differences among the displays with respect to performance times. Analysis of variance showed a p value < 0.001 for this

TABLE 4
First level mapping from taxonomic categories to generalized task statements

No.	**Type**	**Formulation**
1	Compare	Which key term has the most documents about ONLY it?
2	Associate	Which key term is associated with more documents?
3	Distinguish	One of the documents is unlike any of the others. Can you identify it? What is different about it?
4	Rank	Rank document *w*, *x*, and *y* with respect to the amount of term B that they contain.
5	Cluster	Which of the following sets are similar? What is the basis for your judgment?
6	Correlate	What significance do you attach to the indicated region (or point in a list)? [Region is a gap where no documents are found?]
7	Locate	If a new document was discovered that had these characteristics (*x*, *y*), where would it be placed in the display? [between which two labeled documents]
8	Categorize	What general category would you place the indicated documents in? [show documents that are related to a single point of interest]
9	Identify	Find a document that is likely to be about both terms in equal proportion.
10	Compare	If you definitely wanted to read documents that had BOTH [ALL] terms in them, which documents would you ignore?

TABLE 5
Second-level mapping from generalized to specific task statements—2-term questions

2-1. Are there more documents that contain ONLY the term Romania or ONLY the term Czechoslovakia?
2-2. Which is the most frequent key term in this set of documents? A. Oil; B. York
2-3. One of the documents is unlike any of the others. Can you identify it? Place the document number in the text box.
2-4. Rank documents A, B, and C with respect to the amount of term "Soviet" that they contain.
2-5. Which of the following documents are most similar with respect to the relative amount of the key terms?
2-6. What of the following statements is true?
A. There are no documents that contain roughly equal amounts for the two terms.
B. If a document talks about "Oil" then it also talks about "Texas".
C. "Texas" and "Oil" are not very highly related.
D. A and C
E. All of the above
2-7. Location
…

TABLE 6
Second-level mapping from generalized to specific task statements—3-term questions

3-1. Are there more documents that contain ONLY the term "earthquake" or ONLY the term "California" or ONLY the term "death"?
3-2. Which is the most frequent key term in this set of documents? A. Vatican; B. Embassy; C. Noriega
3-3. One of the documents is unlike any of the others. Can you identify it? Place the document number in the text box.
3-4. Rank documents A, B, and C with respect to the amount of term "Company" that they contain.
3-5. Which of the following documents are most similar with respect to the relative amount of the key terms?
3-6. Which of the following statements is true? A. At least one document contains all three terms. B. At least one document contains the terms "Arab" and "bomb". C. "Vatican" and "Arab" are not very highly related. D. B and C E. All of the above.
3-7. Location
…

TABLE 7
Parameter number for specific tasks

	Parameter number	
Taxonomic category	**2-term**	**3-term**
Compare $\langle ?x, ?y \rangle$	2	2
Associate $\langle ?x, ?y \rangle$	2	2
Distinguish $\langle ?x, ?y \rangle$	2	2
Rank $\langle ?x1, \ldots, ?xn, ?\text{attr} \rangle$	4	4
Cluster $\langle ?\text{cluster}, \ldots, ?xn \rangle$	5	4
Correlate $\langle ?x1, \ldots, ?xn \rangle$	20	30
Locate $\langle ?x, ?\text{locator} \rangle$	2	2
Categorize $\langle ?x1, \ldots, ?xn \rangle$	20	30
Identify $\langle ?x, ?\text{identifier} \rangle$	2	2
Compare $\langle ?x, ?y \rangle$	2	2

comparison. Using the "spring" as the pivot case, all of the other display types are shown to take a significantly longer time. Generally, for the 2-term displays, the Word display is slowest, the "spring" is fastest, and the other displays are intermediate.

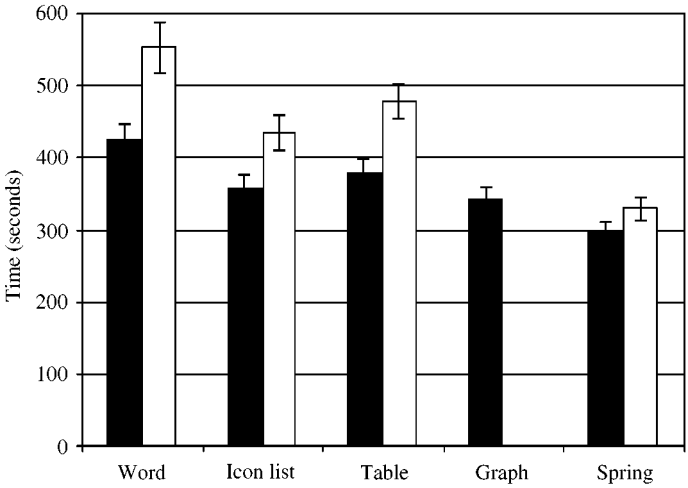


FIGURE 7. Comparison time to completion of 10-question task set vs. display type. □ 2-term; ■ 3-term.

TABLE 8
Within-subjects contrasts for 3-term display

Contrasted sets	F	Significance
Word vs. Spring	50.374	0.000
Icon vs. Spring	18.356	0.000
Table vs. Spring	34.234	0.000

The second major point to be made in this section pertains to the 3-term displays. The four displays were shown by ANOVA to be significantly different ($p < 0.001$). Within-subject contrasts, using the "spring" display as the pivot case, showed it highly different from each of the other displays (Table 8). Further analysis by pair-wise contrasts showed that the word and table displays were roughly equivalent in terms of speed of performance, while the icon display was faster and the "spring", once again, was the fastest.

The final major observation for the data is a comparison across study types. The statistical analysis is shown in Table 9. The data were analysed by repeated-measures ANOVA using study type as the Between-subjects factor. For the word, icon and table displays, the 3-term condition required more time for the subjects to complete than the corresponding 2-term condition. The results for the "spring" display, however, did not achieve significance ($p = 0.086$).

5.3. CORRECTNESS OF ANSWERS

Analysis of the second method of assessing performance, correctness of answers, is shown in Figure 8. The word displays shows a lower number of correct answers than the other

TABLE 9
Effect of display type on time to complete task set — 2-term vs. 3-term

Display type	F	Significance
Word	8.643	0.004
Icon	6.581	0.011
Table	10.126	0.002
Spring	2.979	0.086

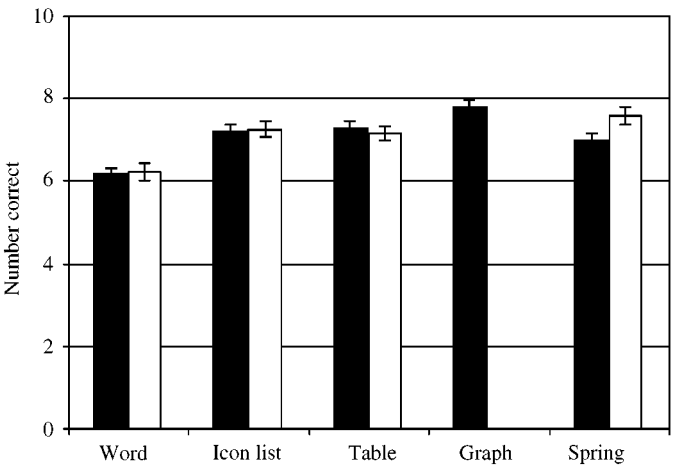


FIGURE 8. Mean score per 10-question task display type. □ 2-term; ■ 3-term.

displays (pair-wise comparisons all < 0.001). The statistical results confirm the visual impression that there is no difference in number of correct answers between the 2- and 3-term studies (repeated measures display vs. correct answers: $F = 0.236$, $p =$ NS).

5.4. ORDER OF PRESENTATION

The order of presentation was randomized. A complete block for the 2-term display series was comprised of 120 different orderings. The 3-term display series required only 24 orderings; each was administered three times. This design permitted analysis of display differences presented in the previous section without the need of worrying about the order. Order of presentation has been analysed in the preliminary and the current design supports dissection of this aspect of the overall plan of testing.

Since the types of tasks being presented to users may not be the types of tasks that they are used to performing, it might be possible that there is some general trend to learning

all displays. On the other hand, it might be the case that only displays that are difficult to use when presented early in the series are learned by using other displays, while some displays might be easy to use at first sight.

Performance as a function of time for the 2- and 3-term studies is shown in Figure 9. It is clear from the figures that each display type was associated with poorer performance when it was presented first in the series. There were progressive decreases in the time that it took the subjects to answer the full set of questions associated with each display. Statistical analysis of the effect of ordering showed that the first point was different from the others, but that subsequent presentations were not different from each other. This finding may seem contrary to the visual appearance of the figure; the later points appear to be steadily decreasing albeit at a slower rate than between the first and second points. It should be noted that the number of observations at each point is 24 rather than the full 120. That is, 24 subjects received one of the displays first, second, third, fourth and fifth. The standard error of the mean of these values was in the order of 10% of the mean. Such variation prevents detection of changes among the data points.

Time data for the 3-term study bear a striking resemblance to the 2-term data. The slopes of the lines, however, are initially steeper. For the word display, the time in the 2-term condition for a display seen first in sequence is 629 s which decreases to about 320 s if seen fourth; the same values for the 3-term study are 825 and 330, respectively. The "spring" display appears to be more flattened than the other curves in the 3-term study (490–270 vs. 410–280). Statistical analysis using multivariate ANOVA showed that the Word, Icon and Tables displays were significantly different between the 2- and 3-term study, while there was no difference between ordering effects for the "spring" displays. This indicates that increasingly complex data might be more amenable to visual treatment. There was no significant effect of order of presentation on performance as measured by number of correct answers.

The order of presentation has a notable effect on time-to-completion but none on number of correct answers. The key observations regarding the time effect are: (1) there is a steep drop in time required between the first and second display regardless of which displays are seen in these slots; and (2) the "spring" display is handled extremely rapidly in the 3-term condition; the "spring" display is the only display that is not influenced by

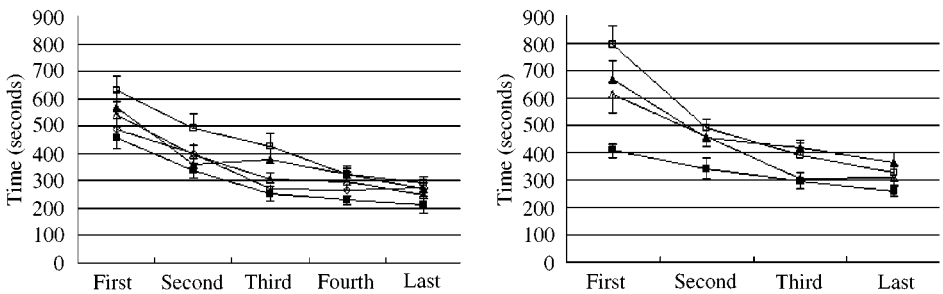


FIGURE 9. Time to completion with respect to order of presentation. Left panel shows the results of the 2-term study and right panel shows the 3-term data: —□—, word; —Δ—, icon; —▲—, table; —◇—, graph; —■— spring.

the increased complexity of the 3-term condition when compared with the paired 2-term display.

5.5. PERFORMANCE WITH RESPECT TO TASK TYPES

The role of individual tasks chosen from a visual taxonomy and implemented with a known number of parameters will be discussed in this section. We will first discuss question types independently of display type. Then the results of individual tasks vs. the different displays will be presented.

Analysis of the overall study design using a repeated measures analysis showed highly significant differences ($p < 0.001$) between subject performance and question type. Paired contrasts were performed to determine the source of these flagged differences. Arbitrarily, the first question was used as the pivot group. Both performance measures showed a significant difference for each pair of values, except for the "Distinguish" question for time and the "Rank" question with respect to correctness. This analysis was run with data pooled from both the 2- and 3-term studies.

Figure 10 shows the relationship between time and correctness for each question type, which are labeled at each data point. The data for this comparison ignore the values for the graph format since it was available only in the 2-term condition. Therefore, the maximum possible value on the y-axis is four (4). The error bars represent the standard error of the mean. For both the 2- and 3-term study, there is an inverse relationship between these measures. In general, questions that are answered quickly are also answered correctly and vice versa. On average across displays, each question takes longer to answer in the 3-term condition than in the 2-term one. On the other hand, average number of correct answers is not significantly different between the two studies.

The data from which Figure 10 was drawn are presented in Table 10. Variation is shown in the figure but has not been included in the table to clarify the presentation and support comparisons across columns. The standard error of both measures averaged 5% of the mean with a maximum departure of 10%. Inspection of Table 10 reveals that the Associate, Identify and Rank task were performed in very short time periods and were associated with a very high fraction of correct answers. The Cluster, Locate and some of the Compare tasks were prone to error and took significantly longer to perform.

In order to determine whether the number of parameters that specify a question determines its complexity, it is necessary to compare the rankings of the measures in Table 10 to the parameter number. The information from Table 7 is duplicated here to aid in the comparison. Clearly, there is no relationship between these pieces of data.

Repeated measures analysis of variance of the overall study design showed that there were significant differences between question types depending not only on type of study (2- or 3-term) but also with respect to individual display types. In order to dissect of this difference, the analysis was repeated using a question X display factor ordering instead of the display X question ordering used previously. This manipulation forces generation of the desired within-subjects contrasts. The observed power for this set of comparisons was lower than for the results presented to this point, which had been > 0.9.

The results showed that both Icon and/or "spring" displays were accompanied by significantly faster performance times than the base case (word) display. This difference

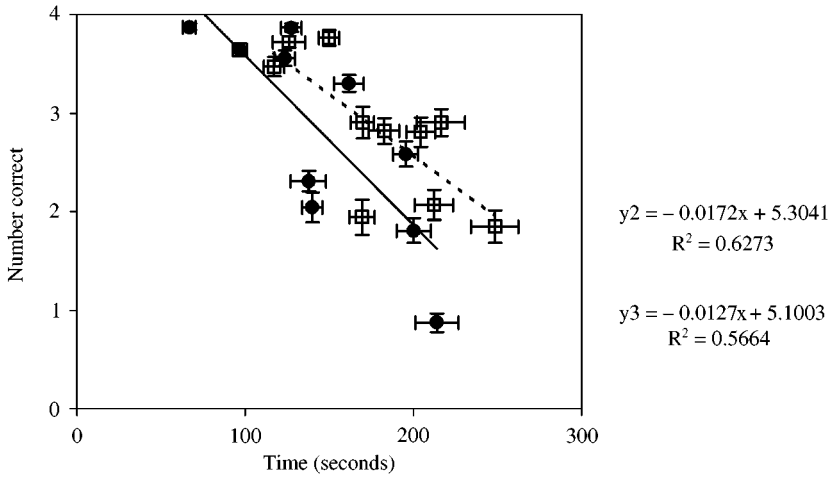


FIGURE 10. Relationship of time to number of correct answers: ● 2-term; □ 3-term; - - - - - 3-term; —— 2-term.

TABLE 10
Mean time and correctness for individual questions

Task type	2-term			3-term		
	Parameter no.	Time (sec)	Correct	Parameter no.	Time (s)	Correct
Compare 1	2	127.5	3.87	2	216.3	2.90
Associate	2	66.8	3.88	2	126.0	3.72
Distinguish	2	123.5	3.56	2	204.3	2.81
Rank	4	161.5	3.30	4	149.9	3.76
Cluster	5	213.9	0.88	4	212.1	2.07
Correlate	20	139.9	2.04	30	169.6	2.90
Locate	2	200.1	1.81	2	248.3	1.85
Categorize	20	137.6	2.31	30	169.5	1.94
Identify	2	96.8	3.64	2	117.2	3.47
Compare 2	2	195.4	2.58	2	182.4	2.82

was found for the Rank, Cluster, Correlate and Locate tasks ($p < 0.05$). The table, on the other hand, was relatively ineffective at producing fast response times and often was slower than the base case.

5.6. PREFERENCES

After performing tasks with each of the display types, subjects were asked to rank the displays. In addition, they were given a free choice area in which they could assign zero or more displays to categories such as "hard", "easy", etc.

Subjects ranked the displays after using all of them. The results are summarized in Figures 11 and 12 for the 2- and 3-term studies, respectively. Analysis showed that there was no relationship of these preference rankings and subject performance, when measured by time to completion. There was, however, a correlation between rankings and correctness for both the 2- and 3-term groups. In each case, the "spring" display was preferred by subjects who received high scores when using it. In the 2-term study, the same observation was made for Graph. However, overall the icon display was the most preferred, especially for the 3-term displays.

The rankings were tested for a relationship to the order in which the subject encountered the display type. Non-parametric analysis was used and the results showed no correlation between the position in which any display was seen and any positional ranking assigned by the subjects in either the 2- or 3-term study.

In order to compare the studies, the data were adjusted by removing references to the Graph presentation in the 2-term study. The Kruskal–Wallis test was applied to the resultant data and it showed that the rankings for best and for worst display were significantly different (Table 11). The inference than can be drawn from this data is that the "spring" display was preferred more often in the more difficult 3-term study than in the easier 2-term condition.

5.7. SUMMARY OF RESULTS

- Displays are different whether measured by time to complete or number of correct answers.
- Visual tasks can be created that vary widely in difficulty.
- There was no demonstrable correlation of task difficulty and taxonomic parameter number.

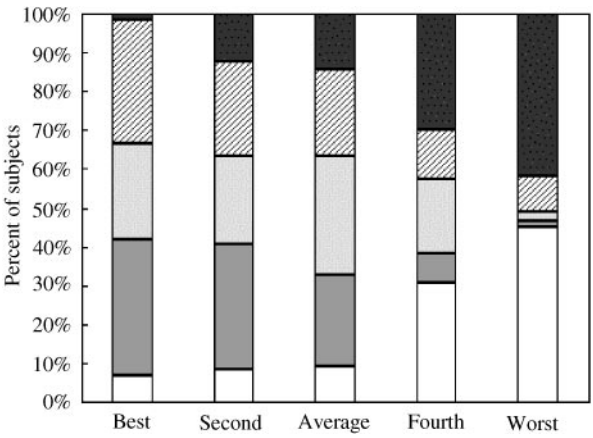


FIGURE 11. Preference rankings for 2-term displays: ■, spring; ▨, graph; □, table; ▩, icon list; □, word.

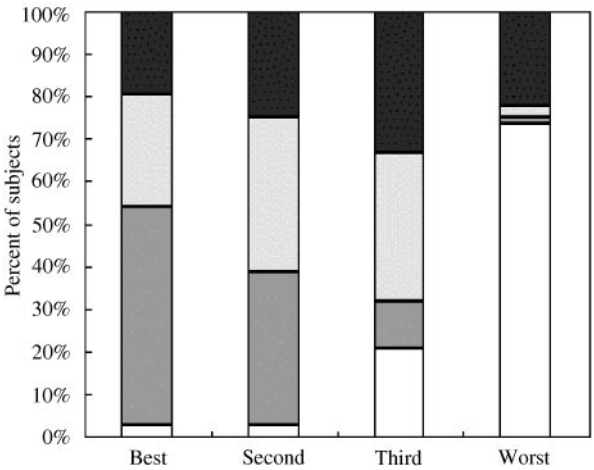


FIGURE 12. Preference rankings for 3-term displays: ■, spring; □, table; ▒, icon list; □, word.

TABLE 11
Results of Kruskal-Wallis analysis of ranking data with respect to study type

	Best	Second	Third	Worst
Chi-Square	6.308	1.389	2.187	26.746
Significance	0.012	0.239	0.139	0.000

- Three term displays were more difficult than their 2-term counterparts.
- As the complexity increased, the "spring" display was resistant to performance decrement.
- Subjects preferred the displays that made them perform better.
- Subjects were not influenced in their preferences by primacy or recency effects with respect to display presentation order.
- Overall, the icon display was the most preferred.

6. Discussion

This discussion will attempt to compare and contrast the current set of results with those of the preliminary studies reported earlier and, to the extent possible, to the pre-existing literature discussed in Section 2.

6.1. DISPLAY TYPES

In the multidimensional scaling of display types done by Lohse, Ruetor, Biolsi and Walker (1990) these displays could be ranked along the "data type" dimension as: Text ⇒ icons ⇒ graph ⇒ table ⇒ "spring". Using the "cognitive effort" dimension, the ordering would be: graph ⇒ text ⇒ table ⇒ icons ⇒ "spring". A third ranking of the

prototypes could be developed by rating them according to the type of symbols they contain; this contention would generate the following order: text ⇒ tables ⇒ graphs ∼ cons ⇒ "spring". Such an ordering captures an increasing amount of visual information as opposed to reading information. The symbols in text are the letters and words. In tabular presentations, numbers stand as surrogates for the words. Graphs use positioning as a substitute and icons use color and shape. The "spring" display uses size, shape, position and color.

Table 12 shows a summary of results with regard to display effectiveness. The shaded part of the table shows results from a set of preliminary tests (Morse, Lewis, Korfhage & Olsen 1998; Morse, 1999).

Clearly, this comparison shows that word and text displays were always associated with poor performance when time to perform a series of task was the measure. Just a clearly, the "spring" display is superior in producing quick responses. It is important to note that these finding are true regardless of the difficulty of the test or the type of data being rendered. Performance across trials with respect to correctness presents a less clear picture. In all cases, it appears that the questions being posed were of insufficient difficulty to elicit clear-cut performance differences among the groups of subjects. The mean score for all studies was about 85%, which is very high. In such situations, it is possible that the 15% error rate could be due to unintentional causes.

The fact that similar error rates were encountered with drastically different performance times indicates that the "spring", an instance of visual display, supported rapid extraction of information contained in that display at an optimal level of accuracy. If the "spring" display had been only superior with respect to speed but had led to more wrong answers, then the question would still be open about whether visual displays are superior. The only stronger case would have been if both measures had shown a positive effect.

Conclusions that might be drawn from analysing the effect of display type on performance include the following.

- Performance using correct answers as an indicator needs to be assessed in situations of increased difficulty in order to accentuate potential differences.

TABLE 12
Comparison of display effectiveness across all studies

No. of terms	Data type	Measure	Best	2nd	3rd	4th	Worst
2	Boolean	Correctness	Icons	Text	Spring = Table = Graph		
3	Boolean	Correctness	Icons	Table = Word		Spring	
		Time	Spring	Table	Icons	Word	
2	Vector	Correctness	Graph	Spring = Icons = Table			Word
		Time	Spring	Graph	Icons	Table	Word
3	Vector	Correctness	Spring = Icons = Table			Word	
		Time	Spring	Icons	Table	Word	

- The prototype displays used in these studies exhibit highly discernible characteristics that are associated with widely different times to process them to make inferences.
- The preliminary and current studies agree on the basic findings with respect to the prototypes.

6.2. TASKS CREATED FROM A VISUAL TAXONOMY

The tasks performed by subjects in the preliminary studies were implemented as Boolean tasks, that is, they were constructed as sequences of AND and OR segments. The vector studies presented a challenge, since Boolean tasks were not appropriate when the underlying data were not intrinsically Boolean. The visual taxonomy was therefore employed in developing a set of tasks for the 2- and 3-term vector studies. This decision, in retrospect, was a good one although there are some aspects of the implementation that deserve discussion.

The method used to map from the taxonomic categories to a test task set was a 3-step mapping. First, generalized questions that conformed to the descriptions of a taxonomic category were generated for the IR domain. For instance, a Locate task in the IR domain was interpreted to mean "where would a document with certain characteristics be found in the display". This process continued for each category. The second step matched particular displays with each generalized statement. And the final step was to create an actual test question based on the key terms used in the display and taking into account the features of the display, such as clusters, outliers or gaps.

At each stage in this process, the possibility existed of producing a task that might be ambiguous or misleading. The fact that the subjects performed with a very high degree of accuracy tends to suggest that the questions were well-formed by-and-large. The results also showed that there were very sharp differences among the various tasks with respect to the amount of time that it took to answer the questions. This difference in time is indicative of a level of difficulty.

A confounding factor that was not anticipated was the interaction with the actual format of a question. As alluded to several times in the Results section, a question that is posed to determine the level of a subject"s knowledge may be asked in many different ways. The notorious questions found on standardized exams that have answers, such as "A", "B", "C", "A and B", "A and C" or "All of the above", are very convoluted. In fact, the format of the question makes it very likely that a student would possess sufficient knowledge to answer many more straight-forward questions about the same material. The relevance to the current situation is that some of the final mappings produced tasks that were of a somewhat involved nature. These questions were less frequently answered correctly.

Although the use of the visual taxonomy failed at the lowest level, it succeeded in producing task sets that supported superior performance when visual displays were used. This finding differs from that of the preliminary Boolean studies, in which the "spring" display was associated with poorer performance than most of the other displays. That visual tasks are performed better with visual displays has not been demonstrated previously. When this statement is viewed in light of the conclusion regarding learnability of displays, it seems that the challenge lies in the development of interfaces that encourage subjects to ask questions in a visual manner. Before Boolean retrieval systems

existed, people would have had a hard time phrasing Boolean queries. Even today Boolean systems are difficult for many people to use. Systems that support visual inquiry will need to allow users to phrase questions visually. In developing such systems, the contention could be made that a visual taxonomy would provide a useful, if not necessary, set of guiding principles.

The major conclusions regarding the utility of a visual taxonomy are as follows.

- The visual "spring" display supported superior performance of the tasks in the vector condition. These tasks were mapped from a visual taxonomy.
- The visual "spring" display used in the Boolean study, although learnable, was not as good at supporting performance as the other display types.
- A visual taxonomy promises to be a useful guide for developing visual interfaces in general and IR interfaces in particular.

6.3. PREFERENCES

Of all the data gathered in these experiments, preference information is the most consistent. In all studies, subjects were asked to rank the displays that they had used. In the vector studies, subjects were also asked to rate the displays according to several qualitative categories — "hard", "easy", "fun" and "annoying". A third kind of preference information was collected from the optional comments that subjects could provide on the posttest survey form found in the vector format only. The three methods for gathering information in the vector studies produced the same results.

At first glance, the data presented in Table 13 might seem to indicate that the visual prototype used in this study was highly distasteful to the subjects. Closer examination shows clearly that the "spring" display was significantly more appreciated in the more complex 3-term situation than in the easier 2-term paired condition. Subjects not only ranked it less often as the "worst" display but also ranked it significantly more often as the "best" display. This observation is especially noteworthy when viewed in the context of how subjects performed on the tests in which they preferred the visual display. In these studies, there was a positive correlation between performance and preference. It appears, therefore, that subjects like to use things that make them successful. In the context of developing interfaces to assist users in exploring document spaces, it seems that making interfaces that can be used successfully will be met with acceptance by those users.

Icon displays were in each test scenario very highly values by the subjects. It is interesting to note that several of the interfaces that have been developed for IR systems use displays that incorporate icons of the sort embodied by the icon prototype of this study. TileBars (Hearst, 1995) and SIRRA (Aalbersberg, 1995) are clearly relatives of the prototype icon display. In addition, there are instances that are unattributed in various web search engine reports. These visual displays are based on bar graphs and/or histograms and are very familiar to the average user of systems, which contributes to their utility.

The most notable observation of the preference data is that "text" and "word" displays were extremely ill-preferred, regardless of performance. In the 2-term studies, performance with this display type was very good, yet user acceptance was very low. This is the normal anecdotal experience of users of Internet search engines and of library searchers.

TABLE 13
Subject preferences across studies (*percent of subjects*)

			Word	Icons	Table	Graph	Spring
Boolean	2-term	Best	14	33	15	9	29
		Worst	47	8	15	20	9
	3-term	Best	10	43	12		35
		Worst	52	6	16		26
Vector	2-term	Best	7	35	25	32	2
		Worst	45	2	2	9	42
	3-term	Best	3	51	26		19
		Worst	74	3	28		22

Text alone may be sufficient to allow users to solve problems when browsing but text alone is unsatisfying.

Conclusions regarding preference data include indications for interface design

- Visual interfaces are associated with enhanced performance as well as user preference.
- As situations become more difficult, visual presentations become more valued.
- The challenge to incorporating visual elements into IR systems is in designing easily learned interfaces. Acceptance is a function of perceived utility.

7. Conclusion

A study of text, table, icon, graphical and "spring" displays for presentation of vector data have been presented. The technique applied here is based on a back-to-basics strategy where the visualization techniques themselves, not the systems where they are implemented, are tested. This allows for simpler studies, that will give more accurate results.

A task set based on a visual taxonomy was developed, based on research reported in the literature. Based on this taxonomy a mapping process was employed where generalized questions that conformed to the descriptions of a taxonomic category were generated for the IR domain, then specific test question were created based on these general questions. Through this process the common error of presenting all retrieval question as Boolean statements was avoided.

The presented studies confirmed our belief that the "spring" visualization can provide an efficient and effective way to provide information about document sets. The "icon" visualization was well liked, and also proved effective for low-dimensional descriptions. In extending these methods to larger numbers of reference points, increasing visual

clutter and mental workload will be encountered for icon displays and increasing relational ambiguity for spring displays. While we hypothesize that the trend of improving performance for problems of increasing complexity will hold for "spring" visualizations this prediction remains to be tested.

The case has been made here for the utility of visualizations for supporting information retrieval activities. However, it is clear that not all tasks that an information seeker might need to perform can be satisfied with visual methods. The use of a visual taxonomy in these studies provided a way to deal with the complexity of visual tasks. It would be very desirable to have a parallel series of text-based tasks. With these two categorizations—visual and text-based tasks—work could proceed to delineate the characteristics of full IR systems. The integration of visual and non-visual components could be structured rather than being a matter of happenstance.

The application of the visual taxonomy described in this paper should be tested more rigorously. This plan should include attention to details of question formats. In fact, it would be highly desirable to perform this activity as the focus of a whole study rather than as a part of any larger work. The reasons for this suggestion include the breadth of the taxonomy itself. It is important to test as many of the groupings as possible. In addition, internal validity needs to be assessed by replicating question types with varying formats.

We believe that systematic evaluation of IR visualizations, themselves, as we have begun in these studies is needed to design the next generation of IR tools. Only by understanding the relationships among visual representations, complexity and IR tasks can we build tools that exploit human capabilities for visual inference effectively.

References

AALBERSBERG, I. J. (1995). Personal communication in Nuchprayoon (1996)

CHALMERS, M. (1996). A linear iteration time layout algorithm for visualising high-dimensional data. *Proceedings of IEEE Visualization '96*, pp. 127–132.

HEARST, M. A. (1995). TileBars: visualization of term distribution information in full text information access. *Proceedings of the CHI '95*, pp. 213–220.

HEMMJE, M., KUNKEL, C. & WILLET, A. (1994). LyberWorld — a visualization user interface supporting fulltext retrieval. *Proceedings of ACM SIGIR '94*, pp. 249–259, 3–6 July, Dublin.

KENNEDY, J. B., MITCHELL, K. J. & BARCLAY, P. J. (1996). A framework for information visualisation. *SIGMOD Record*, **25,** 30–34.

KIM, H. & KORFHAGE, R. R. (1994) BIRD: browsing interface for the retrieval of documents. *Proceedings of the IEEE Symposium on Visual Languages*, pp. 176–177, St. Louis.

KOSHMAN, S. (1997). *Usability testing of a prototype visualization-based information retrieval system.* Dissertation, University of Pittsburgh.

LIN, X. (1997). Map displays for information retrieval. *JASIS*, **48,** 40–54.

LOHSE, G., RUETER, H., BIOLSI, K. & WALKER, N. (1990) Classifying visual knowledge representations: a foundation for visualization research. *Visualization '90*: *Proceedings of the First Conference on Visualization*, pp. 131–138

MARCHIONINI, G. (1995). *Information Seeking in Electronic Environments*. New York: Cambridge University Press.

MORSE, E. & LEWIS, M. (1997). Why information visualizations sometimes fail. *Proceedings of IEEE International Conference on Systems Man and Cybernetics*, Orlando, FL, 12–15 October.

MORSE, E., LEWIS, M., KORFHAGE, R. R. & OLSEN, K. A. (1998) Evaluation of text, numeric and graphical presentations for information retrieval interfaces: user preference and task

performance measures. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1026–1031, 12–14 October, San Diego, CA.

MORSE, E. (1999). *Evaluation of visual information browsing displays*. Dissertation, University of Pittsburgh.

NUCHPRAYOON, A. (1996). *GUIDO: a usability study of its basic retrieval operations*. Doctoral Dissertation. School of Information Sciences, University of Pittsburgh.

OLSEN, K. A., WILLIAMS, J. G., SOCHATS, K. M. & HIRTLE, S. C. (1992). Ideation through visualization: the VIBE system. *Multimedia Review*, **3**, 48–59.

OLSEN, K. A., KORFHAGE, R. R., SPRING, M. B., SOCHATS, K. M. & WILLIAMS, J. G. (1993). Visualization of a document collection: the VIBE system. *Information Processing and Management*, **29**, 69–81.

PEJTERSEN, A. M. (1988). Search strategies and database design for information retrieval in libraries. In L. P. GOODSTEIN, H. B. ANDERSEN & S. E. OLSEN, Eds. *Tasks, Errors and Mental Models*, *Hampshire*, pp. 171–192. England: Taylor & Francis.

ROGOWITZ B. E. & TREINISH, L. A. (1993). An architecture for rule-based visualization. *Proceedings of IEEE Visualization '93*, pp. 236–243. San Jose, CA, October. Los Alamitos, CA: IEEE Computer Society Press

SPOERRI, A. (1993). InfoCrystal: a visual tool for information retrieval. *Proceedings Visualization '93,* pp. 150–157, San Jose, CA.

WEHREND, S. & LEWIS, C. (1990). A problem-oriented classification of visualization techniques. *Proceedings IEEE Visualization '90,* pp. 139–143

WILLETT, P. (1985). An algorithm for the calculation of exact term discrimination values. *Information Processing & Management*, **21,** 225–232.

ZHOU, M. X. & FEINER, S. K. (1998). Visual task characterization for automated visual discourse synthesis. *Proceedings of the CHI '98,* pp. 392–399