



## **Empirical studies of information visualization: a meta-analysis**

CHAOMEI CHEN AND YUE YU

*The VIVID Research Centre, Department of Information Systems and Computing,  
Brunel University, Uxbridge UB8 3PH, UK. email: chaomei.chen@brunel.ac.uk*

*(Received 23 June 2000, and accepted in revised form 2 August 2000)*

A meta-analysis is conducted on a set of empirical studies of information visualization. To be included in the meta-analysis, a study must meet a set of selection criteria. The meta-analysis synthesizes significant levels and effect sizes, tests the heterogeneity of findings from individual studies included and tests the linear trends over a range of information visualization features with ascending visual-spatial complexity. Recommendations for future experimental studies of information visualizations are included.

© 2000 Academic Press

### **1. Introduction**

Many innovative information visualization techniques and systems have been developed. The significance of empirical evaluations of these systems as well as specific features of visualization has been recognized and well understood. On the one hand, the number of empirical studies of information visualization features and systems is rapidly increasing. On the other hand, there is the urgent need for synthesizing various results across existing studies in the literature.

Similar issues have been traditionally addressed in a variety of disciplines using a quantitative synthesis method called meta-analysis. The greatest strength of a meta-analysis is that it can provide us a simplified and synthesized view and reveal any invariant underlying relations in the vast amount of complex, and often conflicting and confusing information in the literature (Hunter, Schmidt & Jackson, 1982; Hunter & Schmidt, 1990).

A number of fundamental issues must be addressed for the further development of the information visualization field. What is the central research question that most studies aim to address? What is the optimal task-feature taxonomy for information visualization design? What is the most commonly used experimental design? Is there any consensus that one can draw from the existing empirical findings in the literature? What is the most powerful visualization feature for a given task? To what extent are the current empirical findings consistent across different studies?

We conduct a meta-analysis of information visualization studies in order to capture the current theories and practices in empirical examinations of information visualizations. This meta-analysis focuses on three aspects of information visualization, namely users, tasks and tools.

Users refer to the role of individual differences, especially cognitive factors, in the context of a work environment supported by information visualization. Tasks refer to the design of experimental studies involving information visualization. Tools refer to the variety of information visualization design options adopted in a study.

The subsequent meta-analysis utilize the same methodology used in a meta-analysis of hypertext systems (Chen & Rada, 1996). In this meta-analysis, we apply the meta-analytical method to the field of information visualization. This is only the first step to build a task-feature taxonomy that can accommodate the majority mainstream information visualization technologies.

The rest of the article is organized as follows. First, we introduce the meta-analytical method to be used as well as the selection criteria for choosing appropriate studies for meta-analysis. Then, we present a subjective review of the studies identified and synthesize the most commonly used hypothesis, independent variables and dependent variables. The meta-analysis is based on a substantial encoding process. Finally, the results of the meta-analysis are reported and discussed.

## 2. Method

This meta-analysis focuses on experimental studies in which independent variables are related to one of the three contextual variables (namely, users, tasks and tools).

Measures to do with users include several cognitive factors such as *associative memory* (MA), *spatial ability* (VZ) and *visual memory* (MA). However, because of the small number of papers that directly address these cognitive factors, they are excluded from the meta-analysis.

Two types of dependent variables encoded are accuracy and efficiency measures. Accuracy measures typically include *precision*, *error rate*, *the average number of incorrect answers* and *the number of correct document retrieved*. Efficiency measures typically include *the average time to completion* and *the performance time*.

Tools refer to features of information visualization, including well-known visualization features such as cone trees, information landscape, associative networks and multidimensional scaling solutions.

According to Robert Rosenthal, there are six types of meta-analytic procedures, which include (1) comparing diffuse studies of significance testing, (2) comparing diffuse studies of effect size estimation, (3) comparing focused studies of significance testing, (4) comparing focused tests of effect size estimation, (5) combining studies of significance testing and (6) combining studies of effect size estimation.

Two broad strategies are commonly used in a meta-analysis: comparing and combining empirical findings. A meta-analysis usually focuses on two major aspects of a causal relationship: the size of the effect and the significance level of the effect.

In meta-analysis, there are two types of studies: diffuse and focused studies. Diffuse studies often investigate null hypothesis of interest, whereas focused studies typically have hypotheses in specific directions, often based on existing knowledge in the literature. In this meta-analysis, we include all the six types of analyses.

TABLE 1  
*Types of meta-analysis*

	Studies	Significance level	Effect size
Comparing	Diffuse	1	2
	Focused	3	4
Combining		5	6

### 2.1. A META-ANALYTICAL SYNTHESIS

The purpose of the study is to find invariant underlying relations suggested collectively by the empirical findings in order to form an overview of the design and practices of information visualization. This meta-analysis is based on 35 experimental studies published between 1991 and 2000. It focuses on the combined findings concerning following hypotheses.

Effects of users' cognitive ability.

- Users with stronger cognitive ability, i.e. higher psychometrics, tend to perform better with information visualization systems than users with weaker cognitive ability in terms of accuracy.
- Users with stronger cognitive ability tend to perform better with information visualization systems than users with weaker cognitive ability in terms of the speed of performance.

Effects of information visualization (Information visualization tools vs. none visualization tools).

- Users tend to perform better, in terms of accuracy or efficiency, with interfaces with visualization components than interfaces without such features.

According to Ben Shneiderman's data-structure-oriented taxonomy, information visualization can be classified according to the data types they use. His taxonomy includes seven classes, namely, one-, two- and three-dimensional, temporal, multidimensional, tree and network structures. These classes are not mutual exclusive.

For example, one can have a three-dimensional (3-D) network or a two-dimensional (2-D) tree.

In this meta-analysis, we concentrate on information structures visualized in the form of tree and network structures. In addition, we particularly restrict studies to information-retrieval tasks.

### 2.2. SAMPLING

This meta-analysis followed the procedures of sampling, coding and analysis developed in the social and behavioral sciences (e.g. Glass, McGaw & Smith, 1981; Rosenthal, 1987). Experimental studies were located from journals, conference proceedings and digital libraries such as the ACM Digital Library accessible on the web.

Each located study was examined against the following selection criteria.

- The study must include an experiment design.
- The study must include at least one experimental condition in which a visual-spatial component appears in the user interface.
- The study must include at least one dependent variable on accuracy or efficiency.
- The study must report its results in sufficient detail, including *F*-test, *t*-test, correlation coefficients or *p* levels.

Searching for studies in this meta-analysis heavily relied on on-line resources, notably the ACM Digital Library. Search queries were formed based on terms such as *visualisation*, *visualization*, *image*, *graphics*, *evaluation*, *empirical* and *experiment*. Studies were located from a number of journals.

*ACM Transactions on Computer-Human Interaction*

*Communications of ACM*

*ACM Transactions on Information System*

*ACM Transactions on Computer System*

*ACM Transactions on Design Automation of Electronic System*

*ACM Transactions on Database System*

*IEEE Computers*

*Journal of American Society for Information Science*

Conference proceedings searched include the proceedings of IEEE Information Visualization Symposium (1991–1999) and IEEE International Conference on Information Visualization. The search located 35 studies published between 1991 and 2000. Among them, 32 studies (91%) were published within the last 5 years, between 1996 and 2000, indicating that the empirical evaluation of information visualization is still in its early stage.

### 2.3. CODING INDIVIDUAL STUDIES

Coding individual studies is an important step in meta-analysis. The following information was coded for each study: independent variables, dependent variables, sample sizes, methods of assigning subjects, the background of the researchers, visual-spatial components used, the year of publication, tasks and statistics of significance tests. Figure 1 is a status diagram showing studies passed through each step of selection.

Independent variables turned out to be very diverse across individual studies. Many studies were excluded from the final meta-analysis. On the other hand, it was a worthwhile process because it improves our understanding of the current practices of empirical studies of information visualization. A detailed explanation of the process is given as follows.

Among the 35 studies, eight studies were excluded in the first round because they did not include information-retrieval tasks. For example, Mahmoud, Clayden and Higgins (1999) compared the acquisition of environmental cognitive knowledge in the real world and its VRML simulation. It focused on the effect of design background and gender on spatial cognition in both displays. Colin Ware and Glenn Franck (Ware & Franck, 1996) studied the benefits of presenting abstract data in 3-D. Their results showed that motion

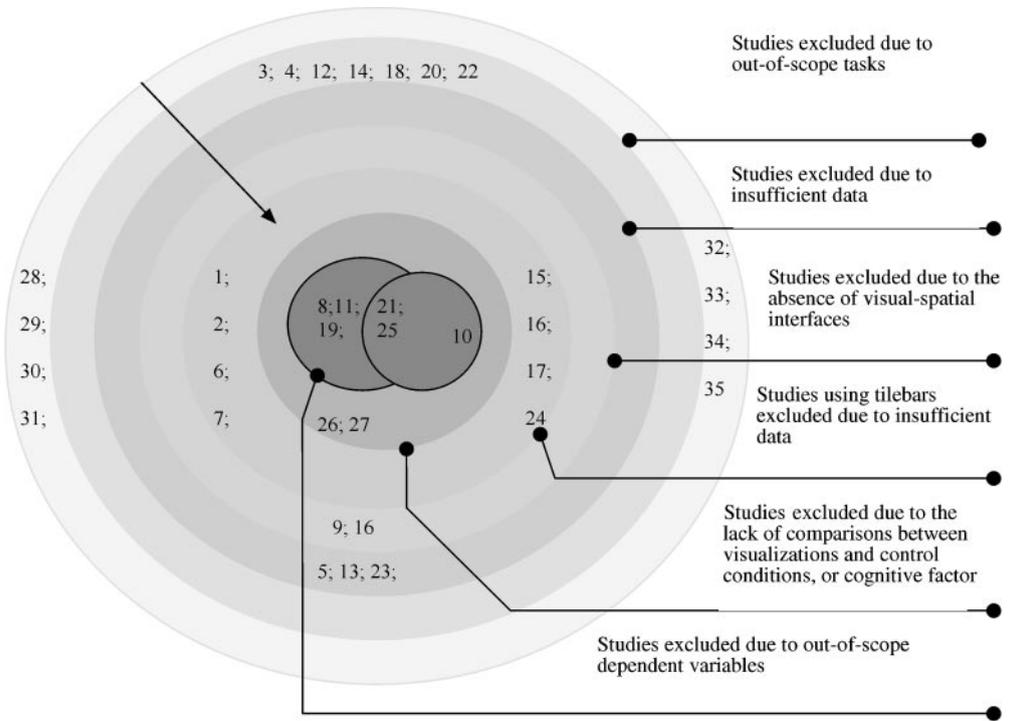


FIGURE 1. Status diagram of the selection process, starting from the outmost layer inward.

cues combined with stereo viewing can substantially increase the size of the graph that can be perceived, however the main aims of them are not relating to information retrieval.

Seven studies were further removed because they only reported standard deviations and means. For studies that only report group means and standard deviations, the significance levels can be calculated as paired *t*-tests. However, we decided to simplify the selection criteria and not to include such studies at this stage. For instance, Veerasamy and Belkin (1996) evaluated the use of a visualization tool for information retrieval and compared the effectiveness of the visualization tool to none visualization tools. However, they only reported the sample size, the mean and median as well as standard deviation in each condition. In practice, it is always possible for analysts to contact the original authors of such studies directly to obtain the necessary data. Twenty studies remained after this round.

According to our selection criteria, eligible studies must include visual-spatial components in user interfaces. We had to exclude three studies in which Scatter/Gather interfaces were used, which did not include visual-spatial displays.

Two studies compared TileBars-like visualization and none visualization versions of an information-retrieval system. Veerasamy and Heikes (1997) concluded that the graphically displaying document surrogate information enables set-at-a-time perusal of documents, rather than document-at-a-time perusal of textual displays. Whittaker,

	Study	Year	User	Task	Visual-spatial variable	
					Visualization	Control
1	Chen & Czerwinski	1997		Text retrieval	Associative network	N/A
2	Wiss & Carr	1999		Text retrieval	Information landscape (fastest) cam tree (fast) information cube	
3	Veerasamy & Belkin (study 1)	1996		Text retrieval	TileBars	
4	Veerasamy & Belkin (study 2)	1996		Text retrieval	TileBars	
5	Pirolli <i>et al.</i>	1996		Text retrieval	Scatter/Gather interface	Keyword search
6	Czerwinski <i>et al.</i> (study 1)	1999		Text retrieval	Data Mountain with implicit query	Implicit query off
7	Czerwinski <i>et al.</i> (study 2)	1999		Text retrieval	Data Mountain with implicit query	Implicit query off
8	Robertson <i>et al.</i>	1998		Text retrieval	Data Mountain	Internet Explorer 4.0 (IE4)
9	Veerasamy and Belkin	1997		Text retrieval	TileBars	No TileBars
10	Swan and Allan (study 1)	1998	Librarians/ general user	Text retrieval	Aspect Window	ZPRISE(GUI information retrieval system from NIST)
11	Swan and Allan (study 2)	1998	Librarians/ general user	Text retrieval	Aspect Window (AspInquery Plus)	ZPRISE(GUI information retrieval system from NIST)
12	Byrd	1999		Text retrieval	Scrollbars	
13	Jose <i>et al.</i>	1998		Image retrieval	Spatial query with EPIC (image retrieval system)	Textual query with EPIC
14	Plaisant <i>et al.</i>	1999		Query retrieval	EOSDIS(NASA's earth observing system data information system)	
15	Lohse	1991			UCIE (a program simulated graphical perception)	N/A
16	Whittaker <i>et al.</i>	1999		Speech documents retrieval	SCAN (Spoken Content-based Audio Navigation)	Control interface
17	Byrne	1993	Spatial ability	Icon retrieval		N/A

18	Card <i>et al.</i>	1994		Calendar retrieval	Spiral Calendar; Sun Calendar program CM	
19	Combs and Bederson	1999		Image retrieval	ThumbsPlusZoomable Image Browser LandScape PhotoGoRound	
20	Hertzum and Fokjær	1996	Dispersion ability	Text retrieval	TeSS (a text retrieval system)	
21	Sebrechts <i>et al.</i>	1999	Novice/expert	Text retrieval	NIRVE (globe)	Text
22	Rodden <i>et al.</i>	1999		Image retrieval	MDS	
23	Sears	1996	Access visualization ability	Text retrieval	State-transition diagram overlay	
24	Hascoët	1998		Icon retrieval	Spiral	Spring-embedder
25	Allen	2000	Cognitive style	Two-level search	MDS	Without MDS
26	Chen (study 1)	2000	VZ MA MV	Text retrieval	Associative network	
27	Chen (study 2)	2000	VZ MA MV	Search performance	Associative network	Text
28	Mahmoud <i>et al.</i>	1999	Cognitive style/gender		VRML world	Real world
29	Graham <i>et al.</i>	1998			With/without model for performance visualization	
30	Volbracht <i>et al.</i>	1997	Cognitive Style			
31	Watson <i>et al.</i>	1997			Two color LCD display	
32	Gribnau & Hennessey	1998		One/two hand operation	3D interface	
33	Ware and Rose	1999		Visual feedback		
34	Ware and Franck	1996		2D/Stereo head-coupled perspective		
35	Douglas and Kirkpatrick	1996		Interface with difference feedback	Color models	

FIGURE 2. Studies located for the meta-analysis.

Hirschberg, Choi, Hindle, Pereira and Singhal (1999) compared their SCAN interface with a visual tape recorder and found that a multimodal interface supporting local navigation helps relevance ranking and fact-finding. However, because they are the only two studies of this type in our sample, the number of studies is too small to run a meta-analysis. These two studies were subsequently removed from the dataset, which left eight studies. The remaining eight studies include information-retrieval tasks, have sufficient data for meta-analysis and, more importantly, include visual-spatial interfaces.

Studies #26 and #27 examined the role of individual differences in information retrieval through a visual-spatial interface, including an investigation of various cognitive factors such as spatial ability (VZ), associative memory (MA) and visual memory (MV). However, because they are the only studies of this sort, it is not sufficient to conduct a meta-analysis (see Figure 2).

Finally, we have six studies fully satisfied the selection criteria. Two broad types of causal relationships emerged.

1. Effects of visual-spatial interfaces on information retrieval.
2. Effects of cognitive ability of users on information retrieval.

Effects are measured by two categories of dependent variables: accuracy and efficiency.

#### 2.4. ANALYSIS

An effect size is the estimate of the magnitude of a specific relationship between two variables. Usually, one is the independent variable and the other is the dependent variable. Effect size  $r$  can be calculated from a given one-tailed  $p$  value and the corresponding sample size. Tests of significance alone are not informative enough for practitioners and designers of visualization systems to judge the usefulness of a visualization feature. This meta-analysis compares and combines effect sizes and significance levels in the form of Fisher's standard score  $z_r$  and the standard normal deviate score  $Z$ . An effect size  $r$  was transformed to Fisher's  $z_r$ . For instance, an effect size  $r$  of 0.30 corresponds to Fisher's  $z_r$  of 0.31.  $Z$  scores can be obtained from reported one-tailed  $p$  values of significance tests according to cumulative distribution functions, such as *TCDF* and *FCDF*.

Effect sizes in Fisher's  $z_r$  were combined according to standard formulae, which can be found in textbooks on meta-analysis (e.g. Rosenthal, 1987). The  $Z$  scores were combined according to Stouffer's method (see Rosenthal, 1987). These two procedures of combination are recommended for their computational simplicity. Finally, the results of the combination were converted back to a correlation coefficient  $r$  as the combined effect size and a one-tailed  $p$  value as the combined significance level.

For studies that only reported group means and standard deviations, the significance levels can be calculated as paired  $t$ -tests. However, we decided to keep our selection criteria simple and not to involve data that require additional processing at this stage. When results were reported as non-significant, a  $p$  value of 0.50 and a  $Z$  of 0.00 were coded.

The heterogeneity test addresses whether the grouping factor is theoretically sound (see Figure 3). A large heterogeneity usually suggests that the grouping factor may not capture the variance of a group of results. According to Rosenthal (1987), the

No.	Study	Year	User	Task	Visal-spatial condition	Control condition
8	Robertson <i>et al.</i>	1998		Text retrieval	Data Mountain	Internet Explorer 4.0 (IE4)
10	Swan & Allan (study 1)	1998	Librarians/ general user	Text retrieval	Aspect Window;	ZPRISE(GUI information retrieval system from NIST)
11	Swan & Allan (study 2)	1998	Librarians/ general user	Text retrieval	Aspect Window; AspInquery Plus	ZPRISE(GUI information retrieval system from NIST)
19	Combs & Bederson	1999		Image retrieval	ThumbsPlus; Zoomable Image Browser; LandScape; PhotoGoRound	1(- not good) 0
21	Sebrechts <i>et al.</i>	1999	Novice/expert	Text retrieval	NIRVE (globe)	Text
25	Allen	2000	Cognitive style	Two-level search	MDS	Absence of MDS

FIGURE 3. Independent variables.

heterogeneity of a set of effect sizes refers to fluctuations from the average of the group. It follows a distribution of  $\chi^2$  with  $K-1$  degrees of freedom, where  $K$  is the number of studies. The heterogeneity of significance levels has the same distribution.

As shown in Figure 4, a number of measures of accuracy were included, such as the average number of incorrect documents retrieved and recall. The measures of efficiency include the completion time, performance time, response time and search time. The majority of visualization systems used in this meta-analysis are research systems, such as Data Mountain, AspInquery Plus, ThumbsPlus, Zoomable Image Browser, LandScape, PhotoGoRound, NIRVE and MDS.

In comparing focused studies, we included a series of tests for linear trends. The purpose of such tests is to find out if the effect size is increasing over a range of visual-spatial design features. For example, contrast weighted effect sizes can be compared to determine whether the influence of explicit linkage display is substantial. One can assign contrast weights of  $-1$ ,  $0$ , and  $1$  to MDS, a minimum spanning tree, and a Pathfinder network respectively, and find out whether there is a linear trend.

### 3. Results

The results of the meta-analysis are presented in two parts corresponding to users and tools. In each part, empirical findings of individual studies are compared and synthesized in terms of effect sizes and significance levels.

No.	Study	Year	Effectiveness (accuracy)	Efficiency (time)
8	Robertson <i>et al.</i>	1998	Average number of incorrect pages	number of trails cannot be retrieved in time
10	Swan and Allan	1998	Recall	completion time
11	Swan and Allan	1998	Recall	completion time
19	Combs and Bederson	1999	Number of image for user satisfaction	performance time
21	Sebrechts <i>et al.</i>	1999	Complex of task and accuracy	response time
25	Allen	2000	Number of references viewed	search time

FIGURE 4. Dependent variables.

### 3.1. EFFECTS OF INDIVIDUAL DIFFERENCES ON ACCURACY

Individual differences refer to user's experience and their ability to use various visualization tools and their cognitive abilities in general. The synthesizing hypothesis states that users with stronger cognitive abilities, for instance, high VZ scores for spatial ability or high MA scores for associative memory, will benefit significantly more from visual-spatial interfaces than those with weaker cognitive abilities. The results from three studies were compared and combined (see Figure 5).

The combined effect size of cognitive abilities on accuracy is 0.60, which is usually regarded as a medium-to-large effect size. The combined significance level  $Z$  is 6.66 and this is statistically significant ( $p < 0.001$ ).

Comparing the significance levels of diffuse studies yielded a statistically significant  $\chi^2$  ( $\chi^2 = 15.99$ ,  $df = 2$ ,  $p < 0.001$ ). Comparing the effect sizes of diffuse studies was also found statistically significant ( $\chi^2 = 9.71$ ,  $df = 2$ ,  $p = 0.0078$ ). These results have confirmed that the meta-analysis hypothesis, i.e. users' cognitive abilities have effects on accuracy with visualization interfaces (see Figure 6).

Comparing focused studies did not find a statistically significant linear trend associated with the degree of visual-spatial features in interfaces. Contrast weights were assigned to MDS, Aspect window, and NIRVE (globe) on comparing focused tests are ( $Z_{\text{focused-contrast-linear-test}} = 0.9822$ ,  $p = 0.16$ ) and ( $Z_{r_{\text{focused-contrast-linear-test}}} = 0.5$ ,  $p = 0.28$ ).

### 3.2. EFFECTS OF USER'S COGNITIVE ABILITIES ON EFFICIENCY

The meta-analysis hypothesizes that users with stronger cognitive abilities will perform more efficiently than users with weaker cognitive abilities. This hypothesis was supported by all the results from studies #10, #21 and #25 (see Figure 7). Results supporting the hypothesis were assigned positive signs and the negative sign indicate findings in opposite direction.

Study	Sample size <i>N</i>	Probability <i>P</i> *	Effect size <i>r</i>	Fisher <i>Zr</i>	Significance test <i>Z</i>	Contrast Weights $\lambda$
10	24	0.0141	0.45	0.48	2.19	- 4
21	15	0.013	0.45	0.49	2.23	3
25	80	0.001	0.79	1.08	7.11	1
Combined			0.60	0.69	6.66	

FIGURE 5. The effects of users' cognitive abilities on accuracy with visual-spatial interfaces.

Comparing studies	Significance levels		Effect sizes	
	$\chi^2$ (2)	Significance value	$\chi^2$ (2)	Significance
Diffuse tests	15.99	0.000064	9.71	0.0078
Focused tests	Contrast test ( <i>Z</i> )	Significance	Contrast test ( <i>Zr</i> )	Significance
	0.98	0.1635	0.51	0.2843

FIGURE 6. The results of a meta-analysis of the effects of users' cognitive abilities on accuracy with visualization interfaces.

Study	Sample size <i>N</i>	Probability <i>P</i> *	Effect size <i>r</i>	Fisher <i>Zr</i>	Significance test <i>Z</i>	Contrast Weights $\lambda$
10	24	0.00075	0.65	0.77	3.17	- 4
21	15	0.223	0.20	0.20	0.76	3
25	80	0.001	0.67	0.81	5.99	1
Combined			0.59	0.53	5.73	

FIGURE 7. The effects of users' cognitive abilities on efficiency.

The combined effect size of users' cognitive abilities is 0.59, which is statistically significant ( $p = 0.005$ , one-tailed). The combined significance level *Z* is 6.66, which is also statistically significant ( $p = 0.005$ , one-tailed). The significance levels in comparing diffuse studies are heterogeneous according to the heterogeneity test ( $\chi^2 = 13.69$ ,  $df = 2$ ,  $p = 0.0002$ ), which means that this set of findings are essentially different from each other. On the other hand, the heterogeneous test of effect sizes in diffuse studies is not statistically significant ( $\chi^2 = 3.9$ ;  $df = 2$ ;  $p = 0.14$ ), which means, in terms of effect sizes, this set of findings are similar to each other. There is a statistically significant linear trend in terms of effect size across this set of studies. Effect sizes on visual-spatial interfaces towards the high end tend to be smaller than those on visual-spatial interfaces towards the lower end ( $Zr_{\text{focused-contrast-linear-test}} = -1.36$ ,  $p = 0.0043$ ). These results seem to suggest that given the same level of cognitive ability, users tend to perform better on less-sophisticated visualization interfaces. For example, users with MDS are likely to outperform their counterparts with NIRVE (globe) (see Figure 8).

Comparing studies	Significance levels		Effect sizes	
	$\chi^2 (2)$	Significance	$\chi^2 (2)$	Significance
Diffuse tests	13.69	0.000215	3.90	0.1420
Focused tests	Contrast test ( $Z$ )	Significance	Contrast test ( $Zr$ )	Significance
	- 0.87	0.088	- 1.36	0.0043

Note:  $\lambda = - 4; 3; 1$ .

FIGURE 8. The results of a meta-analysis of the effects of users, cognitive abilities on efficiency.

Study	Sample size $N$	Probability $P^*$	Effect size $r$	Fisher $Zr$	Significance test $Z$	Contrast Weights $\lambda$
8	32	0.013180	0.39	0.42	2.22	5
11	24	0.029300	0.39	0.41	1.89	3
19	30	0.000002	- 0.85	- 1.27	- 4.67	0
21	15	0.007500	0.63	0.74	2.43	- 3
25	80	0.002206	0.32	0.33	2.85	- 5
Combined			0.09	0.09	2.11	

Note: One-tailed  $p$  levels are used. If a study does not report specifically on this, we presume they are two-tailed and they will convert to one-tailed  $p$  levels.

FIGURE 9. The effects of visualization on accuracy.

### 3.3. EFFECTS OF VISUALIZATION ON ACCURACY

Five studies tested the effects of visualization. The hypothesis is that visual-spatial information-retrieval interfaces will enable users to perform better than traditional retrieval interfaces. This hypothesis was supported by four out of the five studies (see Figure 9).

The combined effect size is small ( $r = 0.089$ ) according to Cohen (1977), but this is not statistically significant ( $p = 0.234$ ). The individual effect sizes significantly differ from each other. The combined significance level ( $Z = 2.11$ ) is also not statistically significant ( $p = 0.05$ ).

Statistically significant discrepancies were found among both significance levels and effect sizes ( $\chi^2 = 39.89, df = 4, p = 0.000$  and  $\chi^2 = 64.12, df = 4, p = 0.000$ , respectively) (see Figure 10).

The results of linear trend tests in focused studies did not show a statistically significant linear trend across the range of visual-spatial interfaces ( $Z_{\text{focused-contrast-linear-test}} = - 0.464, p = 0.32$ , one-tailed). For example, users did not do increasingly better from MDS to the Data mountain.

### 3.4. EFFECTS OF VISUALIZATION ON EFFICIENCY

Three studies tested the efficiency of visualization. The hypothesis is that users using visualization interface in information retrieval will perform more efficiently than their

Comparing studies	Significance levels		Effect sizes	
	$\chi^2$ (4)	Significance value	$\chi^2$ (4)	Significance
Diffuse tests	39.89	0.0000	64.12	0.0000
Focused tests	Contrast test (Z)	Significance	Contrast test (Zr)	Significance
	- 0.46	0.3228	- 3.00	0.0013

FIGURE 10. The results of a meta-analysis of effects of visualization on accuracy.

Study	Sample size N	Probability P*	Effect size r	Fisher Zr	Significance test Z	Contrast Weights $\lambda$
8	32	0.027911	0.34	0.35	1.91	3
19	30	0.004187	- 0.48	- 0.52	- 2.64	1
25	15	0.000200	0.91	1.55	3.54	- 4
Combined			0.43	0.46	1.63	

FIGURE 11. The effects of visualization on efficiency.

Comparing studies	Significance levels		Effect sizes	
	$\chi^2$ (2)	Significance	$\chi^2$ (2)	Significance
Diffuse tests	20.50	0.000006	36.79	0.0000001
Focused tests	Contrast test (Z)	Significance	Contrast test (Zr)	Significance
	- 2.17	0.150	- 4.38	0.000005

FIGURE 12. The results of a meta-analysis of effects of visualization on efficiency.

using an none visualization interface. This hypothesis was supported by studies # 8 and # 19, but rejected in study # 19 (see Figure 11).

The combined effect size is medium large ( $r = 0.43$ ) according to Cohen (1977), but this not statistically significant ( $p = 0.05$ , one-tailed). The individual effect sizes differ significantly from each other. The combined significance level ( $Z = 1.63$ ) is statistically significant ( $p = 0.025$ ).

Statistically, neither effect sizes nor significance levels were found consistent across studies by the heterogenous tests ( $\chi^2 = 20.5$ ,  $df = 2$ ,  $p = 0.000$  and  $\chi^2 = 36.79$ ,  $df = 2$ ,  $p = 0.000$ , respectively). Linear trend tests did not find significant linear trends ( $Z_{\text{focused-contrast-linear-test}} = - 2.17$ ,  $p = 0.15$ , one-tailed) (see Figure 12).

#### 4. Conclusions

Major conclusions we can draw from this meta-analysis can be summarized as follows.

- Empirical studies of information visualization are still very diverse and it is difficult to apply meta-analysis methods.

- Individual differences, including a variety of cognitive abilities, should be investigated systematically in the future.
- Given the same level of cognitive abilities, users tend to perform better with simpler visual-spatial interfaces.
- The combined effect size of visualization is not statistically significant. A larger homogeneous sample of studies would be needed to expect conclusive results.

This is the first attempt in raising the awareness that it is crucial to conduct empirical studies concerning information visualization systematically within a comparable reference framework. As the number of studies on similar visualizations increases, we expect that regularly conducted meta-analyses would be particularly useful to help us to improve our understanding of the empirical aspect of the field as a whole.

In this meta-analysis, we have to reject many studies because they do not meet the conventional selection criteria for a meta analysis one way or the other. In order to improve the quality, clarity and comparability of experimental studies of information visualizations, future experimental studies of information visualizations should carefully take into account the following six aspects of an experimental design.

1. The use of standardized testing information.
2. The clarity of descriptions of visual-spatial properties of information visualizations.
3. The use of standardized task taxonomies for activities such as visual information retrieval, data exploration and data analysis.
4. The focus on the task-feature binding to be investigated in experimental studies.
5. The use of standardized cognitive ability tests.
6. The level of details in reporting statistical results.

Some of the resources are available and some are yet to be developed to enable us to carry out experimental studies at a larger scale of consistency and comparability. For example, many experiments have already made the use of the data collections prepared by NIST for the TREC Conference series<sup>†</sup>. These collections include not only documents but also pre-defined queries and relevance judgements given by domain experts. The *Kit of Factor-Referenced Cognitive Tests*<sup>‡</sup> has been widely used to measure individuals' cognitive abilities. Conventions of reporting statistical results should become a part of the standard instructions for authors in key journals and conferences in the field, for example, use  $p = 0.078$  rather than  $p < 0.1$ .

The more challenging issue is the design of realistic and practical tasks that can really put specific features of information visualization into test. The provision of more task-feature taxonomies is certainly desirable so as to widen the range of our options in designing experimental studies. The development of task-feature taxonomies relies on a better understanding of how users make use of given visualization functions. To a large extent this is an adaptation process between users, available visualization functions and their tasks at hand—there is always more for us to find out.

By following the above guidelines towards more consistent and comparable experimental studies, we will be able to utilize analysis and synthesis tools such as

<sup>†</sup><http://trec.nist.gov/data.html>

<sup>‡</sup><http://www.ets.org/aboutets/tsdirect/prog02.html#K010>

meta-analytical methods more effectively. We will be able to make sense of diverse and possibly conflicting empirical findings more confidently and systematically and we will be able to improve our knowledge of what makes an information visualization useful and how we can make it even better.

The authors would like to acknowledge the support of the EPSRC (grant number: GR/L61088).

## References†

- → ALLEN, B. (2000). Individual differences and the conundrums of user-centered design: two experiments. *Journal of the American Society for Information Science*, **51**, 508–520.
- BYRD, D. (1999). A scrollbar-based visualization for document navigation. *Proceedings of the Fourth ACM Conference on Digital Libraries*, Berkeley, CA, pp. 122–129.
- BYRNE, M. D. (1993). Using icon to find documents: simplicity is critical. *Proceedings of the Conference on Human Factors in Computing Systems*, Amsterdam, The Netherlands, pp. 446–453.
- CHEN, C. (2000). Individual differences in a spatial -semantic virtual environment. *Journal of the American Society For Information Science*, **51**, 529–542.
- CHEN, C. & CZERWINSKI, M. (1997). Spatial ability and visual navigation: an empirical study. *The New Review of Hypermedia and Multimedia*, **3**, 67–90.
- CHEN, C. & RADA (1996). Interacting with hypertext: a meta-analysis of experimental studies. *Human-Computer Interaction*, **11**, 125–156.
- COHEN, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (revised edn.). New York: Academic Press.
- → COMBS, T. A. & BEDERSON B. B. (1999). Does zooming improve image browsing? *Proceedings of the Fourth ACM Conference on Digital Libraries, '99*, Berkeley, CA, pp. 130–137.
- CZERWINSKI, M. *et al.* (1999). Visualizing Implicit Queries For Information Management and Retrieval. *Proceedings of the CHI 99 Conference on Human Factors in Computing Systems: the CHI is the Limit*, Pittsburgh, PA, pp. 560–567.
- DOUGLAS, S. & KIRKPATRICK, T. (1996). Do colour models really make a difference? *Conference Proceedings on Human Factors in Computing Systems*, Vancouver, Canada, p. 399.
- GLASS, G. V., MCGAW, B., & SMITH, M. L., (1981). *Meta-analysis in social research*. London: Sage.
- GRAHAM, J. H. *et al.* (1998). Evaluation of a prototype visualization for distributed simulations. *Proceedings of the 1998 Conference on Winter Simulation*, pp. 1469–1448.
- GRIBNAU, M. W. & HENNESSEY, J. M. (1998). Comparing single- and two-handed 3D input for a 3D object assembly task. *Proceedings of the CHI 98 Summary Conference on CHI 98 Summary: Human Factors in Computing Systems*, Los Angeles, CA, pp. 233–234. April 18–23.
- HASCOËT, M. (1998). Analytical versus empirical evaluation of spatial displays. *Proceedings of the CHI 98 Summary Conference on CHI 98 Summary: Human Factors in Computing Systems*, Los Angeles, CA.
- HERTZUM, M. & FOKJÆR, E. (1996). Browsing and querying in online documentation: a study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction*, **3**, 136–161.
- HUNTER, J. E., SCHMIDT, F. L., & JACKSON, G. B. (1982). *Meta-analysis: cumulating research findings across studies*. Beverly Hills, CA: Sage Publications.
- HUNTER, J. E., & SCHMIDT, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. London: Sage.
- JOSE, M. J. *et al.* (1998). Spatial querying for image retrieval: a user-oriented evaluation. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 232–240.
- LAZONDER, A. W. *et al.* Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society For Information Science*, **51**, 576–581.

†Studies with → are included in Figure 1. Studies with → → are used in the meta-analysis.

- LOHSE, J. (1991). A cognitive model for the perception and understanding of graphs, *Human Factors in Computing Systems Conference Proceedings on Reading through Technology*, New Orleans, LA, p. 137.
- MAHMOUD, A. H., CLAYDEN, A. & HIGGINS, C. A. (1999). A comparative study of environmental cognition in a real environment and its VRML simulation (virtual reality modelling language). *Information Visualization IV '99*, pp. 208–213.
- PALMQUIST, R. A. & KIM, K. S. (2000). Cognitive style and on-line database search experience as predictors of web search performance. *Journal of the American Society For Information Science*, **52**, 558–566.
- PIROLI, P. *et al.* (1996). Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. *Conference Proceedings on Human Factors in Computer Systems*, Vancouver, Canada, pp. 213–220.
- PLAISANT, C. *et al.* (1999). Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems*, p. 320.
- → ROBERTSON, G. *et al.* (1998). Data mountain: using spatial memory for document management. *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, San Francisco, CA, pp. 153–162.
- RODDEN, K., *et al.* (1999). Evaluating a visualisation of image similarity (poster abstract). *Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, pp. 275–276.
- ROSENTHAL, R. (1987). *Judgment studies: design, analysis, and meta-analysis*. Cambridge, England: Cambridge University Press.
- SEARS, A. (1996). Visualizing efficiency: a technique to help designers judge interface efficiency. *Proceedings of the CHI '96 Conference Companion on Human Factors in Computing Systems: Common Ground*, Vancouver, Canada, pp. 311–312.
- → SEBRECHTS, M. *et al.* (1999). Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces. *Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, pp. 3–10.
- STUART, K. C. *et al.* (1994). The cost-of-knowledge characteristic function: display evaluation for direct-walk dynamic information visualizations. *Conference Proceedings on Human Factors in Computing Systems*, Boston, MA, pp. 238–244.
- → SWAN, R. C. & ALLAN, J. (1998). Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 173–181.
- VEERASAMY, A. & BELKIN, N. J. (1996). Evaluation of a tool for visualization of information retrieval results. *Annual ACM Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 85–92.
- VEERASAMY, A. and HEIKES R. (1997). Effectiveness of a graphical display of retrieval results. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, Philadelphia, PA, pp. 236–245.
- VOLBRACHT, S. *et al.* (1997). How effective are 3D display mode? *ACM Conference Proceedings on Human Factors in Computing Systems*, Atlanta, GA, pp. 540–541.
- WARE, C. & FRANCK, G. (1996). Evaluating stereo and motion cues for visualizing information nets in three dimensions. *ACM Transactions on Graphics* **15**, 121–140.
- WATSON, B. *et al.* (1997). Managing level of detail through head-tracked peripheral degradation: a model and resulting design principles. *Proceeding of the ACM Symposium on Virtual Reality Software and Technology*, Lausanne, Switzerland, pp. 59–63.
- WHITTAKER, S., HIRSCHBERG, J., CHOI, J., HINDLE, D., PEREIRA, F. & SINGHAL, A. (1999). SCAN: designing and evaluating user interfaces to support retrieval from speech archives. *Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, pp. 26–33.
- WISS, U. & CARR, D. A. (1999). An empirical study of task support in 3D information visualizations. *Information Visualization IV '99*, London, UK, pp. 392–399.
- WARE, C. & ROSE, J. (1999). Rotating virtual objects with real handles. *ACM Transactions on Computer-Human Interaction*, **6**, 162–180.