Graphical Encoding for Information Visualization: An Empirical Study

Lucy Nowell
Battelle Pacific Northwest
Richland, WA 99352
(509) 372-4295
Lucy.Nowell@pnl.gov

Robert Schulman
Virginia Tech
Dept. of Statistics
Blacksburg, VA 24061
bs1@vt.edu

Deborah Hix
Virginia Tech
Dept. of Computer Science
Blacksburg, VA 24061
hix@vt.edu

Abstract

Research in several areas provides scientific guidance for use of graphical encoding to convey information in an information visualization display. By graphical encoding *we mean the use of visual display elements such as icon color, shape, size, or position to convey information about objects represented by the icons. Literature offers inconclusive and often conflicting viewpoints, including the suggestion that the effectiveness of a graphical encoding depends on the type of data represented. Our empirical study suggests that the nature of the users' perceptual task is more indicative of the effectiveness of a graphical encoding than the type of data represented.*

1. Overview of Perceptual Issues

In producing a design to visualize search results for a digital library called Envision [12, 13, 19], we found that choosing graphical devices and document attributes to be encoded with each graphical device is a surprisingly difficult task. By *graphical devices* we mean those visual display elements (e.g., icon color hue, color saturation, flash rate, shape, size, alphanumeric identifiers, position, etc.) used to convey encoded information. Providing access to graphically encoded information requires attention to a range of human cognitive activities, explored by researchers under at least three rubrics: psychophysics of visual search and identification tasks, graphical perception, and graphical language development. Research in these areas provides scientific guidance for design and evaluation of graphical encoding that might otherwise be reduced to opinion and personal taste. Because of space limits, we discuss here only a small portion of the research on graphical encoding that has been conducted. Additional information is in [20]. Ware [29] provides a broader review of perceptual issues pertaining to information visualization.

Especially useful for designers are rankings by effectiveness of various graphical devices in communicating different types of data (e.g., nominal, ordinal, or quantitative). Christ [6] provides such rankings in the context of visual search and identification tasks and provides some empirical evidence to support his findings. Mackinlay [17] suggests rankings of graphical devices for conveying nominal, ordinal, and quantitative data in the context of graphical language design, but these rankings have not been empirically validated [personal communication]. Cleveland and McGill [8, 9] have empirically validated their ranking of graphical devices for quantitative data. The rankings suggested by Christ, Mackinlay, and Cleveland and McGill are not the same, while other literature offers more conflicting viewpoints, suggesting the need for further research.

1.1 Visual Search and Identification Tasks

Psychophysics is a branch of psychology concerned with the "relationship between characteristics of physical stimuli and the psychological experience they produce" [28]. Studies in the psychophysics of visual search and identification tasks have roots in signal detection theory pertaining to air traffic control, process control, and cockpit displays. These studies suggest rankings of graphical devices [6, 7] described later in this paper and point out significant perceptual interactions among graphical devices used in multidimensional displays. *Visual search tasks* require visual scanning to locate one or more targets [6, 7, 31]. With a scatterplot-like display (sometimes known as a *starfield display* [1]), users perform a visual search task when they scan the display to determine the presence of one or more symbols meeting some specific criterion and to locate those symbols if present. For *identification tasks*, users go beyond visual search to report semantic data about symbols of interest, typically by answering true/false questions or by noting facts about encoded data [6, 7]. Measures of display effectiveness for visual search and identification tasks include time, accuracy, and cognitive workload. A more thorough introduction to signal detection theory may be found in Wickens' book [31].

Issues involved in studies that influenced the Envision design are complex and findings are sometimes contradictory. Following is a representative overview, but many im-

Proceedings of the IEEE Symposium on Information Visualization 2002 (InfoVis'02)
1522-404X/02 $17.00 © 2002 IEEE

portant details are necessarily omitted due to space limitations.

1.1.1 Unidimensional Displays. For unidimensional displays — those involving a single graphical code — Christ's [6, 7] meta-analysis of 42 prior studies suggests the following ranking of graphical devices by effectiveness: color, size, brightness or alphanumeric, and shape. Other studies confirm that color is the most effective graphical device for reducing display search time [7, 14, 25] but find it followed by shape and then letters or digits [7]. Benefits of color-coding increase for high-density displays [15, 16], but using shapes too similar to one another actually increases search time [22].

For identification tasks measuring accuracy with unidimensional displays, Christ's work [6, 7] suggests the following ranking of graphical devices by effectiveness: alphanumeric, color, brightness, size, and shape. In a later study, Christ found that digits gave the most accurate results but that color, letters, and familiar geometric shapes all produced equal results with experienced subjects [7]. However, Jubis [14] found that shape codes yielded faster mean reaction times than color codes, while Kopala [15] found no significant difference among codes for identification tasks.

1.1.2 Multidimensional Displays. For *multidimensional displays* — those using multiple graphical devices combined in one visual object to encode several pieces of information — codes may be either redundant or non-redundant. A redundant code using color and shape to encode the same information yields average search speeds even faster than non-redundant color or shape encoding [7]. Used redundantly with other codes, color yields faster results than shape, and either color or shape is superior as a redundant code to both letters and digits [7]. Jubis [14] confirms that a redundant code involving both color and shape is superior to shape coding but is approximately equal to non-redundant color-coding. For difficult tasks, using redundant color-coding may significantly reduce reaction time and increase accuracy [15]. Benefits of redundant color-coding increase as displays become more cluttered or complex [15].

1.1.3 Interactions Among Graphical Devices. Significant interactions among graphical devices complicate design for multidimensional displays. Color-coding interferes with all achromatic codes, reducing accuracy by as much as 43% [6]. Indeed, Luder [16] suggests that color has such cognitive dominance that it should only be used to encode the most important data and in situations where dependence on color-coding does not increase risk. While we found no supporting empirical evidence, we believe size and shape interact, causing the shape of very small objects to be perceived less accurately.

1.1.4 Ranges of Graphical Devices. The number of instances of each graphical device (e.g., how many colors or shapes are used in the code) is significant because it limits the range or number of values encoded using that device [3]. The conservative recommendation is to use only five or six distinct colors or shapes [3, 7, 27, 31]. However, some research suggests that 10 [3] to 18 [24] colors may be used for search tasks.

1.1.5 Integration vs. Non-integration Tasks. Later research has focused on how humans extract information from a multidimensional display to perform both integration and non-integration tasks [4, 26, 27]. An *integration task* uses information encoded non-redundantly with two or more graphical devices to reach a single decision or action, while a *non-integration task* bases decisions or actions on information encoded in only one graphical device. Studies [4, 30] provide evidence that object displays, in which multiple visual attributes of a single object present information about multiple characteristics, facilitate integration tasks, especially where multiple graphical encodings all convey information relevant to the task at hand. However, object displays hinder non-integration tasks, as additional effort is required to filter out unwanted information communicated by the objects.

1.2 Graphical Perception

Graphical perception is "the visual decoding of the quantitative and qualitative information encoded on graphs," where visual decoding means "instantaneous perception of the visual field that comes without apparent mental effort" [9, p. 828]. Cleveland and McGill studied the perception of quantitative data such as "numerical values of a variable...that are not highly discrete..." [9, p. 828]. They have identified and empirically validated a ranking of graphical devices for displaying quantitative data, ordered as follows from most to least accurately perceived [9, p. 830]: Position along a common scale; Position on identical but non-aligned scales; Length; Angle or Slope; Area; Volume, Density, and/or Color saturation; Color hue.

1.3 Graphical Language Development

Graphical language development is based on the assertion that graphical devices communicate information equivalent to sentences [17] and thus call for attention to appropriate use of each graphical device. In his discussion of graphical languages, Mackinlay [17] suggests three different rankings of the effectiveness of various graphical devices in communicating *quantitative* (numerical), *ordinal* (ranked), and *nominal* (non-ordinal textual) data about objects. Although based on psychophysical and graphical perception research, Mackinlay's rankings have not been experimentally validated [personal communication].

1.4 Observations on Prior Research

These studies make it clear that no single graphical device works equally well for all users, nor does any presentation work well for all purposes. Thus, the challenge for a user interface designer is to choose graphical devices to support the range of tasks users are likely to perform with an appli-

cation, while also supporting individual differences of the user population. The following section discusses our use of the Envision digital library in an empirical evaluation of the effectiveness of graphical encodings using icon color, shape, and size to convey nominal and quantitative information.

2. Test Bed: Envision

Named after Tufte's book [26], Envision [19, 20, 21] is a multimedia digital library of computer science literature, with full-text searching and full-content retrieval capabilities, serving computer science researchers, teachers, and students at all levels of expertise. Envision's Graphic View Window (see Figure 1), presents each document in a search result set graphically as an icon, with relevance ranking numbers shown as labels below the icons. The Graphic View resembles a number of other designs that use starfields [1] or other scatterplot-like displays for visualization of non-hierarchical data — including Bead [5], FilmFinder [1], and SemNet [10]. The Graphic View supports users in making decisions about which works to examine in potentially large sets of documents. The Envision user interface design has been subjected to extensive formative usability evaluation, detailed in [19, 20, 21].

2.1 Graphical Devices Used in Visualization

Because users' perceptual strengths vary and users' decision criteria reflect their current information needs, each graphical device in the Graphic View is user-controllable to represent different document attributes as a user desires. In Figure 1, the shape of each icon encodes document type (e.g., book, journal article, or proceedings article). Icon color encodes document relevance to the query, with bright gold for the most relevant, green for medium relevance, and pale blue for least relevant. Icon size is not used as a code in this example. A legend is at the top of the window, so users need not rely on memory in using the codes. Controls, which are part of the legend, allow users to change the graphical encoding on the fly.

2.2 Tasks Performed with the Graphic View

With Envision, users perform visual search tasks when they try to determine if a document icon meeting a specific requirement is present in the display; success is possible only if that requirement is represented by one of the graphical devices (e.g., icon position on an axis, icon color, etc.).

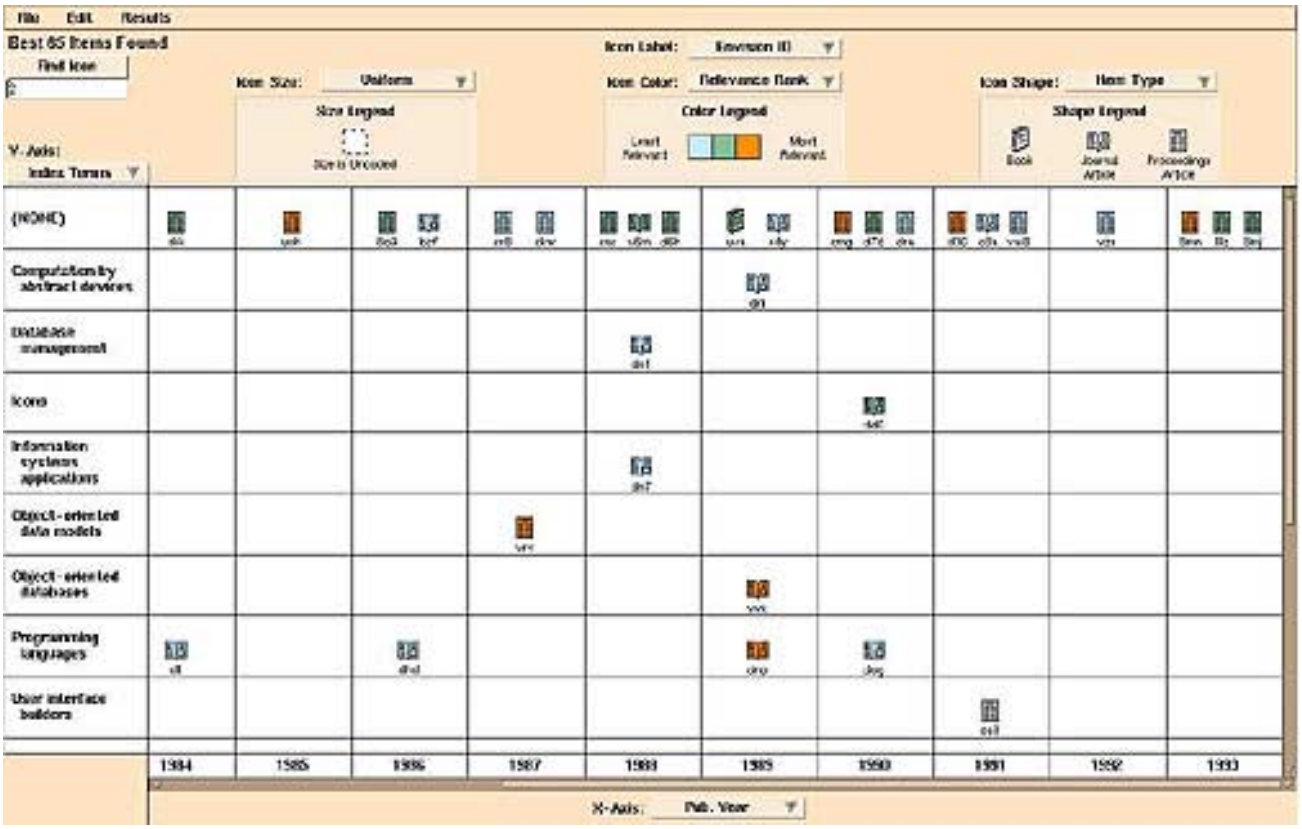

Figure 1. Envision Graphic View search result display.

Identification tasks are performed when a graphical device communicates nominal data such as document type about documents in the Graphic View and an Envision user attempts to locate documents of a particular type. Envision users perform graphical perception tasks when making comparative quantitative judgments about two documents based on icon attributes after those icons are located, as well as when accessing detailed information from the Graphic View about a specific document once its icon is located.

2.3 Icon Design Constraints

Our review of psychophysics and graphical perception suggests that some graphical devices produce slower response times and more errors when used to encode a particular data type (e.g., icon shape to show relevance or icon size to show document type). Accordingly, these visualizations were not originally planned for Envision. However, because a primary purpose of this research was to test such assumptions, those visualizations were included in the study. For encodings expected to perform poorly, we tried to devise the most effective code possible for use in the experiment, and we implemented these designs in a special version of the Envision client. Encodings in this group include use of icon shape to encode relevance and use of icon size to encode document type.

We used only three instances of each graphical device (e.g., 3 colors, 3 shapes, 3 sizes) because of the need to maintain discriminability among the individual instances of each graphical device when it represented nominal data (i.e., document type). The Envision design limits the maximum size of icons in order to maximize the number of icons that may be displayed simultaneously, supporting the goal of allowing users to perceive patterns in the collections and results set. The minimum size of icons is also limited by the need to allow colors and shapes of icons to be accurately perceived. Within these size limits, we observed that only three icon sizes could be readily distinguished. To avoid confounding due to variations in the number of instances comprising each graphical code, we also limited our color and shape codes to three instances each. Of necessity, therefore, icon color was limited to three colors to convey document type and three colors to convey document relevance, while icon shape was limited to three shapes to convey document type and three shapes to convey document relevance.

We also applied several other constraints to icon design, chiefly to avoid confounding by use of other graphical devices sometimes used in codes (e.g., orientation, texture, etc.). First, when size is uncoded, all icons for a shape code were required to be approximately the same area. When size is coded, all icons for a given size category (e.g., small, medium, or large) were required to be approximately the same area, though some variance was allowed because of differences in perceived area of various shapes. For example, if icon size conveys document type and icon shape conveys relevance, it is possible to have three sizes of each shape. Second, all icons were required to have a true vertical orientation, rather than allowing various degrees of slanting. A separate empirical study detailed in [20] confirmed the equal discriminability of the graphical codes used for the study described below.

3. Experimental Design

Using the Envision Graphic View, we conducted a within-subjects empirical investigation of the effectiveness of three graphical devices — icon size, icon shape, and icon color — in communicating nominal (document type) and quantitative (document relevance) data. We chose these graphical devices because of their widespread use and expected power in communication, combined with the uncertainty of their actual impact. We also wanted to work with graphical devices that can be combined in multi-dimenisional encoding. We used three experimental levels (i.e., representing nominal or quantitative data, or uncoded) for each of the three factors (e.g., color, size, and shape). Results for unidimensional and redundant encodings are below; results for the multi-dimensional non-redundant encodings are detailed in [20].

Participants were 20 graduate and undergraduate students at Virginia Tech and were evenly divided between men and women. They were also of varying races, nationalities, and academic disciplines. The recruiting message stipulated that participants would be self-reported to have normal or corrected-to-normal vision and normal color vision. However, we accepted one participant who volunteered the information that he had amblyopia (i.e., lazy eye). We also accepted one participant who said that he had a learning disability affecting use of symbols. We chose to accept these two participants because both were successful students and we believed their limitations existed in the Envision user community, so including them met our commitment to "real world" conditions. Each trial presented a search result display captured from Envision. Each subject was asked to count the icons representing documents that met given conditions, where the information about those conditions was graphically encoded in the display.

Throughout the experiment, the x-axis showed publication year (quantitative data) and the y-axis showed index terms (nominal data). Tasks for the experiment did not require use of either position encoding. Even though we did not use tasks requiring use of position encoding, this context was included because such context information is inherent in the Envision user interface, as it is in other complex visualization displays. Similarly, no tasks required use of icon labels, though these are always present in Envision. Typically, Envision icon labels show document relevance rank. To avoid confounding because of this possible redundancy with other relevance encoding, icon labels showed Envision document ID for all trials.

Trials were divided between training and measured trials. Because a given combination of graphical encodings may present multiple options for information extraction (e.g., using a single code out of several presented or some combination of the codes), tasks were balanced among the options, thus enabling us to study interaction of codes with one another. Objective measures were accuracy and time for task completion. Subjects were also asked to rate each condition for cognitive difficulty and for desirability as an information source.

4. Data Analysis and Results

Dependent variables were time for task completion and error rate. For time to task completion, we performed separate analyses of variance for each of two groups of experimental conditions: conditions 1-7, encoding only document type in the icon design; and conditions 8-14, encoding only relevance. For each group of conditions, we began with a repeated-measures one-way ANOVA, which was significant at the 0.0001 level in each case, even after reduction of degrees of freedom according to the Huynh-Feldt criterion.

Technically, multiple range tests to determine all pairwise differences should also employ reduced degrees of freedom, but in our analyses the degrees of freedom had very little effect on critical values. Accordingly, instead of reducing the degrees of freedom, we employed the more rigorous approach of using Tukey's Honestly Significant Difference (HSD) test [23] at the 1% level of significance. For each group of test conditions, this multiple range procedure tests each pair of means for equality, while allowing only a 1% chance of Type I error in the entire collection of pairwise comparisons.

For nominal data, with the risk of Type I error at 1% (alpha = 0.01), Tukey's HSD test showed all four codes involving color to require similar mean times to task completion but to be significantly faster than the mean times for all codes not involving color. The color code alone produced the fastest mean time, followed by the redundant code with both color and size as type, then the triply redundant code for type, and last the redundant code with both color and shape as type, though the differences among these four were not statistically significant. The remaining three codes were ordered with the redundant code using both shape and size as type first, followed by size as type, with shape as type last, though again the difference among these three was not significant.

For quantitative data analyzed in like manner, Tukey's HSD test produced three groupings by effectiveness, with the significantly faster first group again including all codes involving color and the slower second group consisting of both codes using shape but not color. Surprisingly, the third group contained only the unidimensional code using size, which produced faster results than shape among codes conveying document type.

The ordering of these groups was slightly different from that for the type codes. In the first group, the fastest mean time was from the redundant code with both color and size as relevance, followed by the code with color alone as relevance, then the redundant code with both color and shape as relevance, with the triply redundant code fourth, though differences among these mean times were not significant. In the second group, the code with shape alone as relevance led, followed by the redundant code with both shape and size as relevance, and the code with size alone as relevance was last.

For both unidimensional and redundantly encoded conditions, error rate was the proportion of trials in which the participant gave an incorrect answer in 12 trials for one experimental condition (i.e., the possible error rates were 0/12, 1/12,…12/12). Error rate was not continuous nor was it normally distributed, because many participants made no errors in a majority of conditions, so that analysis of variance was not possible.

As we did for measurements of mean time to task completion, we analyzed these measurements of error frequency in two groups: one group for all seven conditions representing only document type and one group for all seven conditions representing only document relevance. We began analysis with a Chi Square test for each group of proportions to determine whether the proportions were equal across the group. For the seven codes conveying only document type, the Chi Square test showed a significant difference in error frequency among the codes ($p = 0.001$). We then moved to Fisher's Exact Test with risk of Type I error at 1%. Fisher's Exact Test calculates the probability that two proportions differ by chance. We then used Fisher's Exact Test to establish the groupings shown below. Note that in these tests, there is a 1% risk of Type I error in each comparison, unlike the Tukey HSD Test, which distributes the risk of Type I error across all comparisons.

The ranking by error frequency of unidimensional graphical codes conveying type, established by Fisher's Exact Test, is as follows:

$$\text{Color} \leq \text{Shape} \leq \text{Size}$$

For redundant two- and three-dimensional codes conveying document relevance, the ranking produced is

$$\text{Color\&Shape} \leq \text{Color\&Size} \leq \text{Color\&Shape\&Size} < \text{Shape}$$

We analyzed data for conditions with quantitative data in like manner. For unidimensional codes conveying document relevance, the ranking by error frequency is

$$\text{Color} < \text{Shape} \leq \text{Size}$$

For redundant two- and three-dimensional codes conveying document relevance, the ranking produced is

$$\text{Color\&Shape} \leq \text{Color\&Size} \leq \text{Color\&Shape\&Size} < \text{Shape\&Size}$$

We compared the proportion of errors for all redundant codes conveying document relevance to that for all non-redundant codes conveying document relevance. The result showed a significant benefit to redundant codes ($p = 0.00000369$).

4.1 Rankings for Codes Conveying Nominal Data

We produced two rankings for codes conveying document type (nominal data) in the course of this analysis, one by mean time to task completion and the other by frequency of errors during trials.

These rankings, along with Mackinlay's [17] and Christ's [7], are shown in Table 1. Columns to the left of the vertical bar show rankings obtained from these studies; those to the right are rankings from other studies. In comparing these rankings, we note that we use color as a single icon attribute, as does Christ. Mackinlay, on the other hand, separates color into three icon attributes: hue, saturation, and density, or relative darkness on a gray-scale. In Mackinlay's rankings, each of these three icon attributes associated with color ranked ahead of shape and size.

Table 1 — Rankings for codes conveying nominal data

	Time	Errors	Mackinlay	Christ	Christ Ident.
Color	1	1	1	1	1
Shape	3	2	2	3	3
Size	2	3	3	2	2

Our rankings by error frequency correspond exactly to those of Mackinlay. Our ranking by mean time to task completion corresponds exactly to Christ's rankings by time for visual search tasks and by accuracy for identification tasks, while our ranking by time (and Christ's) agrees only partially with our rankings by error frequency (and Mackinlay's). Thus, we find that color is ranked as the most effective graphical device for conveying nominal data, across all rankings shown. The ordering of shape and size, however, differs among the rankings. Ranked by error frequency, as well as in Mackinlay's ranking, shape ranks second, with size third. However, both of Christ's measures and ours for time to task completion place size before shape in effectiveness.

4.2 Rankings for Codes Conveying Quantitative Data

We produced the same rankings for codes conveying quantitative data that we produced for nominal data. These rankings, along with those of Mackinlay [17], Christ [7], and Cleveland and McGill [8, 9] are shown in Table 2. Again, the vertical bar separates our results from those of other studies. And again, in comparing these rankings, we note that we use color as a single icon attribute, as does Christ. Both Mackinlay and Cleveland and McGill, on the other hand, separate color into three icon attributes: hue, saturation, and density. In their rankings, each of these three icon attributes associated with color ranked above shape and size, as we show for color in the table.

Table 2. Rankings for codes conveying quantitative data

	Time	Errors	Christ Search Time	Mackinlay	Cleveland & McGill
Color	1	1	1	2	2
Shape	2	2	3	3	3
Size	3	3	2	1	1

Our rankings by time to task completion and error frequency are the same. However, our rankings do not correspond exactly with those of any of Mackinlay, Christ, or Cleveland and McGill. Mackinlay and Cleveland and McGill agreed completely in their rankings, which suggest size as the most effective encoding for quantitative data, followed by color. Neither ranking suggested shape as an appropriate encoding for quantitative data, so we show it ranked third by both. Our rankings agree with Christ that color is the most effective graphical device for conveying quantitative data. However, we differ with Christ on the ordering of shape and size, in that we consistently rank shape as more effective than size, while Christ does the opposite.

The disagreement between our rankings and those of both Mackinlay and Cleveland and McGill does not necessarily signify an error in any of these rankings. Rather, we believe the disagreement reflects a fundamental difference in the nature of the tasks on which the rankings are based. The graphical perception task used in the experiments by Cleveland and McGill required making a determination about a specific quantitative value or comparing exactly two graphical items in a display that did not contain extraneous data. Our experiments, on the other hand, required users to search a display and count only those items that met specified criteria, among a varying number of distractors. This, we believe, is a realistic representative task that users of large, complex visualization displays (such as Envision) are likely to perform. The difference in tasks used in these experiments reflects a fundamental difficulty in conducting empirical research of this kind: there is more than one way to identify a measurable task that accurately reflects use of a graphical presentation.

The reason our rankings differ from those of Christ is less clear, because the tasks used in our experiments are more similar to the visual search and counting identification tasks he described. However, Christ's rankings were based on a meta-analysis of previous experiments, involving a wide variety of media. We believe the differences in our rankings may reflect differing degrees of discriminability between our graphical codes and those of earlier studies, including those re-analyzed by Christ. We have found only one such study, that of Smith and Thomas [25], that reports discriminability data. As noted in Section 2.3, our own study of discriminability is detailed in [20].

5. Recommendations To Designers

Mackinlay's [17] rankings suggest that designers' choices of graphical devices should vary, depending on the data represented. Cleveland and McGill [8, 9], on the other hand, are concerned only with graphical perception, which is, by definition, limited to quantitative data. Christ's [6] rankings ignore the type of data represented, focusing instead on the kinds of tasks performed.

Because variations in these choices lead to different conclusions about which rankings are best, we believe that the type of data represented should be less significant for designers choosing among rankings to guide design decisions than two other issues, namely

- the exact nature of user tasks to be performed, and
- the most significant measure of effectiveness.

As we have seen, no one study's ranking of graphical devices for conveying nominal data accurately predicts performance by all measures, though this is not the case for codes conveying quantitative data. For quantitative data, our rankings correlate imperfectly with those of Christ [6], whose research is from the vantage point of psychophysics, while our rankings differ more with those of Mackinlay [17] and Cleveland and McGill [8, 9].

As noted earlier, differences between our rankings and those of both Mackinlay and Cleveland and McGill do not necessarily imply that either is wrong. Where counting identification tasks is required, we believe our rankings are applicable. However, if a graphical perception task is required pertinent (e.g., where users must extract precise quantitative data or make precise numerical comparisons between two graphical objects), we suggest that Cleveland and McGill, along with Mackinlay, provide better guidance. We also note that one encoding may yield faster results than another but still be deemed more difficult to use and thus less likely to be used (see [20]). Thus, user preference may dictate choice of graphical devices where objective measures of performance alone are not the critical determinants of effectiveness.

Some designers are reluctant to use color codes, because of variability in computer displays and concern about color-impaired users. However, the power of color codes to improve both accuracy and speed of performance leads us to suggest that color codes should be used. The issue of computer display variability is a difficult one. One option is to allow users to select colors for themselves, but we also know research is underway to develop software tools that allow users to calibrate monitors so that colors displayed are true to originals, except where deviation is triggered to support user color impairments. Meanwhile, by treating color as a single variable, we believe it is possible to choose codes of a few colors (3-5) that can be distinguished by value alone, so that color-impaired users can perceive differences. We also believe it is appropriate to combine color with redundant use of another code, such as shape or texture, both to support color-impaired users and to improve results when the images must be printed on a monochrome or grayscale printer, or color printouts are subject to photocopying. We note that redundant encoding offers the additional advantage of improving accuracy.

Because decisions about every detail of graphical code design impact user performance, from the exact size of icons to barely perceptible variations in color attributes, we recommend that all complex information visualization displays be subject to frequent usability evaluation, exploring every level of design decision making.

6. Summary and Future Work

This study provides empirical evidence regarding the relative effectiveness of icon color, icon shape, and icon size in conveying both nominal and quantitative data. While our studies consistently rank color as most effective, the rankings differ for shape and size. For nominal data, icon shape ranks ahead of icon size on tests of accuracy, but the order reverses for time for task completion, which places shape behind size. For quantitative data, we found that encoding with icon shape is more effective than with icon size in terms of time to task completion and accuracy. For both nominal and quantitative data, we found significantly greater accuracy in responses when redundant codes are used. However, we conclude that the nature of tasks performed and the relative importance of measures of effectiveness are more significant than the type of data represented for designers choosing among rankings.

Further studies of this type are needed to empirically compare effectiveness of other graphical devices suggested by authors such as Christ [6], Cleveland and McGill [8, 9], and Mackinlay [17]. Other graphical devices that might be studied include texture, flash rate, letters and digits, and orientation or angle, etc. It also seems worthwhile to confirm the effectiveness of position encoding in the context of information visualization displays.

Additional studies are needed to determine the number of graphical devices that can be used simultaneously — that is, to determine how many non-redundant codes users can process at once to extract information. Usability evaluation of Envision has shown user success in integrating and filtering information from three-dimensional codes, where color and icon label redundantly represent relevance, and the x- and y-axes convey different document attributes, such as author name and publication year.

The range of further empirical studies in this area is virtually limitless. Such research will continue to produce results that inform the design of information visualization systems by empirically derived guidelines, rather than personal opinion.

Although we used a digital library as the test bed for experimentation, results are pertinent to any complex information visualization display involving large quantities of nominal and quantitative data, including weather displays, command and control information displays, air traffic control displays, and other target acquisition systems. Information regarding effectiveness of graphical devices is

broadly applicable to designers of statistical graphs and iconic displays in determining how to present data to users.

Acknowledgments

We gratefully acknowledge the Envision development team, especially Dr. Lenwood Heath, Dr. Robert K. France and Dr. Edward Fox. Envision was funded by the National Science Foundation (Dr. Maria Zemankova, monitor) and Virginia Tech, with additional support from the ACM

References

1. Ahlberg, C. & Shneiderman, B. (1994) Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proceedings of CHI '94*, Boston. ACM Press, 313-317, 479-480.
2. Bates, M. (1984) The Fallacy of the Perfect Thirty-Item On-Line Search. *RQ*, 42(1), 43-50.
3. Cahill, M. & Carter, R. (1976) Color Code Size for Searching Displays of Different Density. *Human Factors*, 18(3), 273-280
4. Carswell, C. M. & Wickens, C. D. (1987) Information Integration and the Object Display: an Interaction of Task Demands and Display Superiority. *Ergonomics*, 30(3) 511-527.
5. Chalmers, M. & Chitson, P. (1992) Bead: Explorations in Information Visualization. In *Proceedings of SIGIR '92*, 330-337.
6. Christ, R. E. (1984) Research for Evaluating Visual Display Codes: an Emphasis on Colour Coding. In R. Easterby & H. Zwaga (Ed.), *Information Design: The Design and Evaluation of Signs and Printed Material*, New York, NY: John Wiley, 209-228.
7. Christ, R. E. (1975) Review and Analysis of Color Coding Research for Visual Displays. *Human Factors*, 17(6) 542-570.
8. Cleveland, W. S. & McGill, R. (1995) Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, 229, 828-833.
9. Cleveland, W. S. & McGill, R. (1984) Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Stat. Assoc*., 79(387), 531-554.
10. Fairchild, K.M., Poltrock, S.E., & Furnas, G.W. (1988) SemNet: Three-Dimensional Graphic Representations of Large Knowledge Bases. In Guindon, R. (Ed.), *Cognitive Science and Its Application for Human-Computer Interaction*, Hillsdale, NJ: Lawrence Erlbaum,, 201-233.
11. Fox, E.A., France, R.K., Sahle, E., Daoud, A. & Cline, B.E. (1993) Development of a Modern OPAC: From REVTOLC to MARIAN. In *Proceedings of SIGIR '93*, Pittsburgh, pp. 248-259.
12. Fox, E.A., Hix, D., Nowell, L.T., Brueni, D., Wake, W.C., Heath, L.S., & Rao, D. Users, User Interfaces, and Objects: Envision, a Digital Library. *JASIS*, 44(5), 480-491.
13. Heath, L..S., Hix, D., Nowell, L.T., Wake, W. C., Averboch, G.A., Labow, E., Guyer, S.A., Brueni, D.J., France, R. K., Dalal, K., & Fox, E.A. (1995) Envision: A User-Centered Database of Computer Science Literature. *Communications of the ACM*, 38(4), 52-53.
14. Jubis, R.M.T. (1990) Coding Effects on Performance in a Process Control Task with Uniparameter and Multiparameter Displays. *Human Factors*, 32(3) 287-297.
15. Kopala, C. J. (1979) The Use of Color-Coded Symbols in a Highly Dense Situation Display. In *Proceedings HFS*, 397-401.
16. Luder, C.B. & Barber, P.J. (1984) Redundant Color Coding on Airborne CRT Displays. In *Human Factors*, 26, 19-32.
17. Mackinlay, J. (1986) Automating the Design of Graphical Presentations of Relational Information. *Trans. on Graphics*, 5(2), 110-141.
18. Merwin, D.H. & Wickens, C.D. (1993) Comparison of Eight Color and Gray Scales for Displaying Continuous Data. In *Proc. HFS, v. 2*, 1330-1334.
19. Nowell, L. T. & Hix, D. (1993) Visualizing Search Results: User Interface Development for the Project Envision Database of Computer Science Literature. In *Proc. HCI Int'l '93,* 56-61.
20. Nowell, L.T. (1997) *Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data.* PhD thesis for Dept. of Computer Science, Virginia Tech, Blacksburg, VA. http://scholar.lib.vt.edu/theses/available/etd-111897-163723/
21. Nowell, L.T., et al. (1996) Visualizing Search Results: Some Alternatives to Query-Document Similarity. In *Proc. of SIGIR 96,* ACM Press, 67-75.
22. Salton, G.; Wong, A.; & Yang, C. (1975) A Vector Space Model for Automatic Indexing. *CACM*, 18(11), 613- 620.
23. Schulman, Robert S. (1992) *Statistics in Plain English with Computer Applications.* New York: Chapman and Hall.
24. Smallman, H.S. & Boynton, R.M. (1990) Segregation of Basic Colors in an Information Display. *J. Optical Soc. America*, 7(102), 1985-1994.
25. Smith, L. & Thomas, D. (1964) Color Versus Shape Coding in Information Displays. *J. Applied P*sych., 48(3), 137- 146.
26. Tufte, E.R. (1990) *Envisioning Information.* Cheshire, CT: Graphics Press.
27. Umbers, I.G. & Collier, G.D. (1990) Coding Techniques for Process Plant VDU Formats. Applied Ergonomics, 21(3), 187-198.
28. Walker, P. (ed.) (1988) Chambers Science and Technology Dictionary, New York: Chambers/Cambridge.
29. Ware, C. (2000) Information Visualization: Perception for Design. New York: Morgan Kaufmann.
30. Wickens, C. & Andre, A. (1990) Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration. *Human Factors*, 32(1), 61-77.
31. Wickens, C. (1992) *Engineering Psychology and Human Performance, 2nd*. Ed. New York, NY: Harper Collins.