Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size To Convey Nominal and Quantitative Data

by
Lucille Terry Nowell

Dissertation submitted to the Faculty of
the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

Deborah Hix, Chair
Roger W. Ehrich
H. Rex Hartson
Lenwood S. Heath
Robert S. Schulman

November 7, 1997
Blacksburg, Virginia

Keywords: Graphical Encoding, Information Visualization, Visual Search, Digital Library

Graphical Encoding for Information Visualization : Using Icon Color, Shape, and Size To Convey Nominal and Quantitative Data

Lucille Terry Nowell

ABSTRACT

In producing a user interface design to visualize search results for a digital library called Envision [Nowell, France, Hix, Heath, & Fox, 1996] [Fox, Hix, Nowell, et al., 1993] [Nowell & Hix, 1993], 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., color, shape, size, position, etc.) used to convey encoded, semantic information.

Research in the areas of psychophysics of visual search and identification tasks, graphical perception, and graphical language development provides scientific guidance for design and evaluation of graphical encodings which might otherwise be reduced to opinion and personal taste. However, literature offers inconclusive and often conflicting viewpoints, suggesting a need for further research.

The goal of this research was to determine empirically the effectiveness of graphical devices for encoding nominal and quantitative information in complex visualization displays. Using the Envision Graphic View, we conducted a within-subjects empirical investigation of the effectiveness of three graphical devices — *icon color*, *icon shape*, and *icon size* — in communicating nominal (document type) and quantitative (document relevance) data.

Our study provides empirical evidence regarding the relative effectiveness of icon color, shape, and size for 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 by all measures except time for task completion, which places shape behind size. For quantitative data, we found, by all measures, that encodings with icon shape are more effective than with icon size. 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.

GRANT INFORMATION

Initial development of Envision was funded by the National Science Foundation (NSF) as "A User-Centered Database from the Computer Science Literature" at Virginia Tech for 1991-1995, Grant IRI-9116991 (Dr. Maria Zemankova, monitor). Principal investigators were Edward A. Fox, Lenwood S. Heath, and Deborah Hix. The Association for Computing Machinery also allowed free use of its publications.

Development of Envision continued with NSF CISE Institutional Infrastructure (Education) Grant CDA-9312611: Interactive Learning with a Digital Library in Computer Science for 1993-1997. Principal investigators were Edward A. Fox (project director), J.A.N. Lee, Cliff Shaffer, H. Rex Hartson, and Dwight Barnette. A supplement of $67,802 was received in 1997 for workshop, EI home page, digital library courseware.

DEDICATION

This work is dedicated to the memory of my mother and her mother before her, who showed me the way, and to my daughter, Jessie.

ACKNOWLEDGMENTS

This research would not have been possible without the support of many people, a few of whom I would like to thank here. First and foremost are the members of my committee. My friend and advisor, Dr. Deborah Hix, has stood by me throughout this long process, providing support when my morale was failing, as well as finding such scarce resources as a truly private lab for my experiments. She was there when the project that became Envision began and gave me the opportunity to design the user interface that has been the center of my work for over six years. Dr. Robert S. Schulman guided the design of this study and then led me through the maze of statistical analysis, patiently explaining the wonders of SAS and enduring hours of questions about the results. Envision, the test bed system, would not exist had Dr. Lenwood S. Heath not managed the project through a very difficult phase. Additionally, his words during a Project Envision team meeting inspired the design of Envision's Graphic View Window, which is the cornerstone of this research. Dr. H. Rex Hartson taught me much of what I know about user interface design and evaluation in his CS 5714 class and in meetings of the Human-Computer Interaction group, where he never let us forget to put the human first. Dr. Roger Ehrich kept me aware of the vagaries of computer monitors and the impossibility of providing truly consistent conditions for all participants in a study such as mine. His concerns led to our use of the color calibration system and to the discriminability experiment, both keys to eventual replication of these results.

Three other faculty members, who are not members of my committee, also have my thanks: Dr. Edward A. Fox, Dr. Verna Schuetz, and Dr. Helen Crawford. Dr. Fox, the third principal investigator for Project Envision, has been a steady supporter, providing equipment, encouragement, publication opportunities, and even allowing the use of his office for continued development of Envision. Dr. Schuetz, my first advisor at Virginia Tech who has since become a friend, has remained an unfailing source of practical guidance on the logistics and politics of graduate research. She often helped me regain my balance when I was losing it. Dr. Helen Crawford's course on Information Processing (PSYC 5354) gave me the background needed to comprehend many of the studies cited. She also been a friend and mentor, who encouraged me to pursue this path of research, far though it seemed from my roots in Computer Science.

Robert K. France has been central to the Envision effort. His search engine, designed and built for the Marian library catalog system, is at the heart of Envision. He has also put in many long hours programming additional visualizations for the Graphic View, long after grant funding ended. He has been a steadfast friend and supporter throughout the six and half years of our association.

All members of the Project Envision team have also contributed to this research. They include William Wake, Dennis Brueni, Kaushal Dalal, Scott Guyer, Stephen Moore, and Chris Kirmse, as well as Eric Labow, who programmed the first version of the Envision user interface.

Dr. Jock Mackinlay has graciously given permission to use Figure 2.1, which is adapted from one in his dissertation.

Finally, I would like to thank the members of my family who have endured this long journey with me. Tom Nowell, spouse and friend, has kept the home fires burning. Our daughter, Jessie, has managed to graduate from high school and college without her mom's constant attention, though never without her love. Last, there is my mother, Martha Terry, who died before I reached this point, but whose faith in my ability to achieve great things never failed, so she never let me stop believing in myself.

To all of these supporters I offer my heartfelt thanks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 MOTIVATION FOR RESEARCH

In producing a user interface design to visualize search results for a digital library called Envision [Fox, Hix, Nowell, et al., 1993] [Nowell & Hix, 1993], 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., color hue, color saturation, flash rate, shape, size, alphanumeric identifiers, position, etc.) used to convey encoded information. The result of assigning a new semantic value to a graphical device we term a *graphical code* or *encoding*. Graphical codes (or encodings) differ from graphical devices in that the former represent other objects or values, while graphical devices are simply visual elements without semantic content. We use the terms *graphical codes* and *encodings* interchangeably.

The challenge for a user interface designer is to choose devices to support the range of tasks users are likely to perform with an application, while also supporting perceptual and individual differences of the user population. Providing access to graphically encoded information requires attention to a range of human cognitive and perceptual 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 encodings which might otherwise be reduced to opinion and personal taste.

Especially useful are rankings of the effectiveness of various graphical devices in communicating different kinds of data (e.g., nominal, ordinal, or quantitative). Mackinlay [1986] provides such rankings, but they have not been empirically validated. Also, recent developments suggest changes to his rankings. Cleveland and McGill [1984] [1985] have empirically validated their ranking of graphical devices for quantitative data. However, we question the relevance of their methods for information visualization designs which resemble scatter plots (e.g., starfield displays [Ahlberg & Shneiderman, 1994], air traffic control displays, and other iconic representations of data) but often support tasks quite different from those for statistical graphs. Literature (see Chapter 2 Related Work) offers inconclusive and often conflicting viewpoints, suggesting that further research is needed.

1.2 CONTEXT OF WORK

Named after Tufte's book [1990], Envision [Heath, Hix, Nowell, et al., 1995] [Fox, Hix, Nowell, et al., 1993] 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. A unique characteristic of Envision, the Graphic View Window (see Figure 1.1) presents each document in a search results set graphically as an icon, while the Item Summary shows a textual listing of bibliographic information for documents whose Graphic View icons are selected by the user. The Graphic View supports users in making decisions about which works to examine in potentially large sets of documents. Since users' perceptual strengths vary and their 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.

Figure 1.1 — Envision's Search Results Display, Showing Small Query, Graphic View, and Item Summary Windows. (Link to page 3 — Fig1_1.gif — 33K.)

The Envision Graphic View Window is, we believe, the first design that allows users to customize the visualization by choosing the semantics of each graphical device while also constraining users from making choices psychophysical principles indicate are "bad." Envision is also the first visualization design for which psychophysical principles have served as guidelines in determining data attributes to be represented by each of the graphical devices used. Information encoding options for Envision's graphical devices were chosen after a thorough review of literature pertaining to visual psychophysics and graphical perception. (A detailed discussion of psychophysical issues in Envision's user interface design is in Chapter 3.) Thus the Envision user interface design suggests a number of experimental studies.

1.3 OBJECTIVES OF EMPIRICAL STUDY

The goal of this research was to determine empirically the effectiveness of graphical devices for encoding nominal and quantitative information in complex visualization displays. The studies (1) provide empirical evidence of the effectiveness of icon *color*, icon *shape*, and icon *size* in conveying both nominal and quantitative data, and (2) provide empirical evidence regarding the effectiveness of psychophysics, graphical perception, and graphical language guidelines for designing information visualization displays.

While much current research related to visualization focuses on exploring psychophysical issues and developing guidelines, it is not clear that following these guidelines does in fact lead to a more usable design, since there have been no studies of systems actually based on these principles. During discussion of related issues at the SIGIR '95 Workshop on Visual Information Retrieval Interfaces, one researcher specifically commented that it was unclear how research in psychophysics and graphical perception relates to information visualization interfaces. Furthermore, some researchers believe that the experimental conditions necessary to psychophysical research render the results meaningless in "real world" conditions. Since our data were obtained using an actual system in a normal user setting, rather than using only artificial presentations under conditions appropriate only to a psychophysics laboratory, we provide empirical evidence regarding the relevance of research in psychophysics and graphical perception to user interface design. Specifically, our results provide rankings for three graphical devices — icon color, shape, and size — for effectiveness by several measures. Such rankings are fundamental to developing effective graphical languages for information visualization. Comparison of our results to rankings suggested by other authors, along with comparisons among our rankings, provide guidance to designers in choosing among them.

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

1.4 APPROACH TO RESEARCH

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 for investigation because of their widespread use in graphs and other displays and their expected power in communication, combined with uncertainty about their actual impact. These graphical devices also have the advantage that they can be used in combination,presenting the possibility of encoding three different document attributes by using all three graphical devices, each encoding a different attribute. Almost all attributes available for visualization are either nominal or quantitative, as the remaining class — ordinal data — is really either quantitative data with lost precision or it is a special arrangement of nominal data. Since we also tested with each graphical device uncoded, the resulting experimental design required 20 different *design points,* or combinations of graphical encodings. (See table 4.1 for details.) A trial consisted of a single counting identification task performed with a single display.

Presented in SuperCard[1], each trial used a search results display captured from Envision. Trials were divided between training and measured trials. Since a given design point may present multiple options for information extraction (e.g., drawing on a single graphical code out of several presented, or some combination of codes), tasks were balanced among options, thus enabling us to study interaction of codes with one another.

Participants were undergraduate and graduate students at Virginia Tech. For each design point, 20 participants completed 15 counting identification tasks that required comprehension of the graphical code(s) for completion. Participants were instructed to perform tasks as quickly as possible without sacrificing accuracy. Dependent variables were error rate and time for task completion. Participants also were asked to rate each design point for ease of use and likelihood of choosing to use each visualization, providing two subjective measures. Trials for one design point were completed in five minutes or less, on average. A full description of the experiment is given in Chapter 4.

We also conducted a second experiment to document the discriminability of — or ease of distinguishing between — individual elements (i.e., single colors, shapes, or sizes) from other elements of that code, as well as to document the discriminability of the sets of graphical devices from one another. The design of the discriminability experiment was identical to that for the first experiment, except that it involved only 10 participants and the trials required simple visual search tasks, with no decoding needed. This experiment is also detailed in Chapter 4.

[1] SuperCard is a registered trademark of Allegiant.

1.5 CONTRIBUTIONS OF THE RESEARCH

Results of this research provide the following contributions:

- Empirical evidence regarding effectiveness of icon size, icon shape, and icon color in conveying nominal and quantitative data, including interactions among them.
- Rankings of icon color, icon shape, and icon size by several different measures of effectiveness.
- Guidelines on choosing among rankings, both those produced by this study and those resulting from other studies.

CHAPTER 2. RELATED WORK

2.1 PSYCHOPHYSICS OF 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" [Walker, 1988]. *Visual search tasks* require visual scanning to locate one or more targets [Christ, 1975] [Christ, 1984] [Wickens, 1992]. With information visualization displays resembling scatter plots, 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 [Christ, 1975] [Christ, 1984]. Measures of display effectiveness for visual search and identification tasks include time, accuracy, and cognitive workload. Studies in the psychophysics of visual search and identification tasks, rooted in signal detection theory pertaining to air traffic control, process control, and cockpit displays, point out significant perceptual interactions among graphical devices used in multidimensional displays suggest rankings of graphical devices [Christ, 1975] [Christ, 1984] described below and A more thorough introduction to signal detection theory is given in Wickens' book [1992].

2.1.1 Unidimensional Displays

Unidimensional displays — those involving a single graphical device — convey a single kind of information, or use one graphical code. Christ's [1975] reanalysis of 42 prior studies revealed that color codes yielded search times as much as 40% less than size, 43% less than either brightness or alphanumeric codes, and 60% less than shapes. Pairwise comparisons were not done, but these results suggest the following ranking of graphical codes by effectiveness, where < signifies faster performance:

color < size < brightness or alphanumeric < shape

Later experiments with experienced subjects showed that color is the most effective graphical mechanism for reducing display search time, followed by shape, and finally letters or digits [Christ, 1984]. Other studies have confirmed that color is more effective for reducing search times than shape [Smith & Thomas, 1964] [Jubis, 1990], with benefits of color increasing for high-density displays [MacDonald & Cole, 1988] [Louder & Barber, 1984] [Christ, 1975] [Copal, 1979]. Use of shapes too similar to one another actually increases search time [Smith & Thomas, 1964].

Measuring accuracy on identification tasks with unidimensional displays, [Christ, 1975] found that alphanumeric coding gave results as much as 48% better than color coding, while color coding was as much as 32% better than brightness, 176% better than size, and 202% better than shape. Although pairwise analysis was not performed, these results suggest the following order of graphical devices by effectiveness, where < signifies fewer user errors:

alphanumeric < color < brightness < size < shape

In later experiments with experienced subjects, Christ [1984] found that digits gave the most accurate results, but that colors, letters, and familiar geometric shapes all produced equal results. However, Jubis [1990] found that shape codes yielded faster mean reaction times than color codes. Luder and Barber [1984] found no significant difference between the codes for identification tasks.

2.1.1.1 COLOR CODING

Perhaps because numerous studies have shown color to be the most effective graphical device for reducing visual search time [Christ, 1975] [Cahill & Carter, 1976] [Carter, 1982] [Bundesen & Pedersen, 1983] [Treisman & Gormican, 1988], we have found far more studies pertaining to color than to any other graphical device used to encode information. These studies focus on the ability to locate and sometimes to count target items in a display, when cued by color name or sample, with or without distracters of other colors present. Some have sought to establish the value of color coding in comparison to other graphical codes (e.g., [Christ, 1975] [Christ, 1984] [Cleveland & McGill, 1985] [Jubis, 1990]), while others have aimed to determine the optimum number of colors and precisely which colors should be used ([Cahill & Carter, 1976] [Carter, 1982] [Smallman & Boynton, 1990]). Other studies [Bundesen & Pedersen, 1983] [Treisman & Gormican, 1988] [Cavanagh, Arguin & Cavanagh, 1990] [Nagy & Sanchez, 1990] [D'Zmura, 1991], not reported in detail herein, focus on identifying the cognitive mechanisms underlying visual search with color coded targets.

2.1.1.1.1 CAUTIONS ABOUT LITERATURE ON COLOR RESEARCH

Literature pertaining to research in color perception presents some special problems. Vocabulary is perhaps the most important problem: while the computer graphics field has standardized terminology for color characteristics [Foley, van Dam, Feiner, & Hughes, 1990], much pertinent research predates the standards and some of it was conducted without use of computers. Thus, depending on the context and author's background, the words "value," "brightness," and "saturation" may all be used to describe the amount of black or white present, or the placement of a color sample relative to a gray-scale. These words may also have other usage, however, and it is seldom clear in literature which meaning is intended. We have also encountered artists (and others) who use "saturation" to refer to the variable we call "chroma."

Further complications relate to the fact that color itself has three inherent variables: *hue*, *saturation*, and *brightness*. *Hue* is the pure color which is typically named: red, blue, yellow, green. *Saturation* refers to the relative darkness or lightness of a color, in comparison to a gray scale. *Brightness*, used in reference to objects that emit light, refers to the perceived intensity of the object. This term is similar to the use of *lightness* to describe the intensity of reflecting objects. [Foley, van Dam, Feiner, & Hughes, 1990] [Light Source, 1995] We may also use the term *chroma* to describe color purity, or the presence or absence of multiple hues. Further, we presume that "hue" is meant when "color" is described, but it is seldom clear in a report that experiments have adequately controlled for all three color variables and that hue is indeed the subject of the investigation.

Studies such as that by Smallman and Boynton [1990] make it clear that separation in color space is also an issue in any test of accuracy, but few studies reported below describe any steps taken to ensure adequate separability. Additionally, it is difficult to determine just which colors were actually used in studies, given the multiplicity of color identification standards and the absence of information relative to any standard in many articles. Finally, studies have been conducted using a variety of color media (colored papers, colored inks or paints, colored lights, colored light filters for projections, color monitors with computers) and it is not at all clear that results generalize from one to the others.

2.1.1.1.2 HOW MANY COLORS TO USE

The number of instances of each graphical device (e.g., how many colors or shapes are used in the encoding) is significant since the number of elements in the code constrains both the number of item categories and the precision of quantitative information that can be communicated via the encoding. That is, if only four colors are used, then only four classifications of objects may be identified by color; if the information conveyed by color coding is quantitative, then the range of values must be broken into at most four sub-ranges. Many studies of color coding in visual search have focused on the number of colors that may be successfully searched or decoded and the necessary differentiation of those colors in color space [Cahill & Carter, 1976] [Carter, 1982] [Bundesen & Pedersen, 1983] [Treisman & Gormican, 1988] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Nagy & Sanchez, 1990] [Smallman & Boynton, 1990] [D'Zmura, 1991] [Wickens, 1992]. Also, there is evidence that search occurs preattentively and in parallel when the target color is sufficiently different from background/distracter color(s). Exactly what qualifies as "sufficiently different" is a matter of debate and is discussed further below.

Numerous studies have suggested that, under appropriate conditions of color separation and with the number of colors within the cognitive limit, visual search tasks are performed preattentively and in parallel to the point of locating items of the target color, which are then searched serially [Cahill & Carter, 1976] [Carter, 1982] [Bundesen & Pedersen, 1983] [Treisman & Gormican, 1988] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Nagy & Sanchez, 1990] [Smallman & Boynton, 1990] [D'Zmura, 1991] [Wickens, 1992]. The number of colors in the code thus affects the number of items to be serially searched and therefore the time required.

The benefit obtained by addition of color to a display is limited in part by the number of "noise" or background items which are the same color as the target(s) [Christ, 1975] [Carter, 1982] [Treisman & Gormican, 1988] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Nagy & Sanchez, 1990]. If more than 70% of the items of the target color are not truly search targets, visual search is slower than it would be in the absence of color coding [Christ, 1975]. It seems that parallel search continues only up to the point of identifying all items of the target color, and then items of that color are apparently examined serially to identify the target, as search time has been shown to increase linearly with the addition of such items [Carter, 1982] [Treisman & Gormican, 1988] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Nagy & Sanchez, 1990]. Addition of background items not of the target color have been shown to have little or no impact on search time, provided their color is not similar to the target color [Carter, 1982]. If the color of added background items is similar to the target color, search time is increased significantly by their addition [Carter, 1982] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Nagy & Sanchez, 1990]. (Treisman and Gormican [1988] developed a controversial theoretical model to explain the relationship between serial and parallel cognitive processing of color, the details of which are beyond the scope of this paper.)

Thus the issue is how many colors can be present in the code and still yield acceptable discrimination among them, so that the cognitive advantages associated with color coding are optimized. That some such optimal number exists is evident from the fact that at the opposite extreme from a monochromatic display is one in which every item is uniquely color coded, in which case no color code is operative [Cahill & Carter, 1976].

Conventional wisdom holds that no more than five or six colors should be used in coding a display [Christ, 1975] [Carter, 1982] [Cleveland & McGill, 1985] [Shneiderman, 1987] [Wickens, 1992]. Such guidance is based on some evidence that the added benefits of parallelism in search begin to reverse with the use of more than six colors, so that search time increases with each color added beyond six. Also, Smith [1962] found that most people can identify five to eight colors. However,

other studies suggest that larger codes many be effective. Cahill and Carter [1976] concluded that "On low-density displays (10-20 items) as many as 10 colors can be used, as did Chapanis and Halsey [1956] with trained subjects. Even for high-density displays, probably as many as eight or nine colors could be used without risk" [Cahill & Carter, 1976, p. 280]. Smallman and Boynton [1990] have reported improved performance with up to 14 colors and have evidence to suggest that as many as 18 colors may be processed in parallel for visual search tasks. An early study by Bishop and Crook [1960] reported success using up to 30 different colors.

2.1.1.1.3 COLOR SEMANTICS

The semantic values of color codes have been left largely to convention [Umbers & Collier, 1990] [Rice, 1991] [Wickens, 1992]. Kosslyn [1985] summarized and reviewed five books on the design of graphs, evaluating them for psychological validity, clarity of communication, and so forth. Kosslyn asserts that colors should not be used as codes for quantitative differences. He observes that "Differences in color are more like differences on a nominal scale; shifting from red to green does not result in 'more of something' in the same way as shifting from a small dot to a large one does. Indeed, the best depiction of the psychological similarity-space for colors is not a single line, but a circle...Thus it is difficult to use progressive differences in color to stand for progressive increases or decreases in some quantity...Color is not psychologically ordered along a continuum..." [Kosslyn 1985, p. 503].

2.1.1.2 OTHER GRAPHICAL DEVICES

Smith and Thomas [1964] compared three shape codes (i.e., aircraft shapes, geometric forms, and military symbols), each using five distinct shapes. They found significant differences among codes in both accuracy and completion time on a counting task, with military symbols producing fewest errors, followed by geometric forms, and finally aircraft shapes. These effects increased with increases in display density. They observe that while color codes produce better results than shape codes at the code size used (five elements), humans can distinguish far more shapes than colors. Thus, when data require more than the small number of distinguishable colors, shape codes may yield better results.

Perlman and Swan [1993] found that for visual search tasks, texture coding produced reaction times nearly identical to uncoded displays. In a subsequent study, eliminating confounding factors and using a simpler task, Perlman and Swan [1994] found nearly equal search times for color, texture, and density coding (e.g., the value or darkness of the fill color). They note that for the latter experiment, colors used were of equal saturation and equal brightness, implying that the difference in relative effectiveness of the color code in the first experiment resulted from varying saturation and brightness redundantly with varying hue.

In their study of luminance and flash rate, Nagy and Sanchez [1992] found that it is possible to produce luminance codes that yield search rates as fast as those for color. However, they concluded that it is difficult to produce effective luminance codes because the maximum luminance that can be produced on most monitors is quite restrictive. That is, the difference in luminances required to establish a reliably discriminable difference varies between 23% and 33% of the total range, so that luminance coding is probably impractical for use on most color monitors.

Studying four graphical devices (i.e., text style, frame color, border shape, and spatial offset), Swierenga, Boff, and Donovan [1991] found no significant difference in accuracy among spatial

offset, text style, and frame color, but all produced more accurate performance than border shape. Their results for search time were similar.

2.1.2 Multidimensional Displays

For *multidimensional displays* — those using multiple graphical devices combined in one visual object to encode one or more pieces of information about one target — graphical codes may be either redundant or non-redundant. A graphical code is *redundant* if the information conveyed by that encoding can also be extracted from another code, while a *non-redundant* code provides the only means of accessing information thus encoded. A redundant code using color and shape to encode the same information yields average search speeds even faster than non-redundant color or shape encodings [Christ, 1984]. Used redundantly with other encodings, color yields faster results than shape, and either is superior as a redundant code to both letters and digits [Christ, 1984]. Jubis [1990] confirms that a redundant or partially redundant code involving both color and shape is superior to shape coding, but approximately equal to non-redundant color coding. For difficult tasks, using redundant color coding may significantly reduce both reaction time and errors [Kopala, 1979]. Furthermore, benefits of redundant color coding increase as displays become more cluttered or complex [Kopala, 1979]. A study using the graphical devices of text style, frame color, border shape, and spatial offset [Swierenga, Boff & Donovan, 1991] found consistently reduced search times and improved accuracy for all redundant codes except border shape, which reduced performance.

2.1.2.1 Interactions Among Graphical Devices

Significant perceptual interactions among graphical devices complicate issues for multidimensional displays. Studies suggest that color coding may interfere with all achromatic codes: evaluating for accuracy, color coding reduces effectiveness of size codes by as much as 29% [Christ, 1975] [Umbers & Collier, 1990], alphanumeric codes by as much as 14% [Christ, 1975 and 1985], and shape codes by as much as 43% [Christ, 1975]. Increasing the number of colors in a display from three to nine decreased the accuracy of identification of digits, as compared to a monochromatic display of digits [Christ, 1975]. Similarly, increasing the number of colors in a display decreased the accuracy of identification of size coded targets. With later experiments of his own, Christ found that "when colours were added to an achromatic display, the subject's performance in identifying and locating achromatic targets may decrease relative to when the display did not contain colours. This interference effect of colours can occur whether the colours are relevant (i.e. part of the target's unique code) or irrelevant" [Christ, 1984, p. 218].

Indeed, Luder and Barber [1984] suggest 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. Two factors led them to this recommendation: the interference of color with perception of achromatic codes and the presence of monochromatic backup systems, which operators accustomed to color displays are unable to use effectively. Luder and Barber go so far as to suggest that color may have "psychological precedence over shape," so that "The observer may be unable to inhibit the processing of color attributes even when it would be an advantage to do so (i.e., because color is currently inappropriate to the task requirements)" [Luder & Barber, 1984, p. 31]. Where it must be used, they urge that color coding always be used for information that is most directly relevant to the most important task in progress, to minimize the impact of interference.

Interestingly, size and color interact perceptually, so that smaller objects yield less accurate perception of color [Umbers & Collier, 1990]. While we have found no supporting experimental evidence, we believe it is likely that size and shape also interact, causing the shape of very small objects to be perceived less accurately.

2.1.2.2 INTEGRATION VS. NON-INTEGRATION TASKS

Later research [Carswell & Wickens, 1987] [Carswell & Wickens, 1990] [Wickens & Andre, 1992] [Wickens, 1992] [Quinlan & Humphreys, 1987] [Treisman & Gormican, 1988] [Cavanagh, Arguin, & Treisman, 1990] [Treisman, 1991] has focused on how humans extract information from a multidimensional display to perform both *integration* and *non-integration* tasks. An *integration task* requires use of information encoded non-redundantly using two or more graphical devices to reach a single decision or action, while a *non-integration task* based decisions or actions on information encoded in only one graphical device, though multiple graphical codes are present. Thus, some tasks may require integration of data from multiple graphical encodings while also requiring filtering of other graphically encoded data that are irrelevant to the tasks.

When several attributes of a single perceptual object present information about multiple characteristics or sources, the presentation is an *object display* [Carswell & Wickens, 1987]. Several studies [Carswell & Wickens, 1987] [Goettle, Wickens & Kramer, 1991] [Jones, Wickens & Deutsch, 1990] [Wickens & Andre, 1990] support the *principle of compatibility of proximity,* which has two components: 1) object displays facilitate integration tasks, especially where graphical devices all convey information relevant to the task at hand; however, 2) object displays hinder non-integration tasks, as additional effort is required to filter out unwanted information communicated by the objects.

2.1.3 Impact of Short-Term Memory

Use of a scatter plot display is affected by short-term memory of users to the extent that users avoid constant reference to a legend, if present. For example, if the display uses a shape code, users probably attempt to recall shapes of interest rather than comparing every possible target to the legend. Research in human memory commonly makes use of a Sternberg task in experiments [Ashcraft, 1989] [Wickens, 1992]. A *Sternberg task* is a test of human memory capacity. The participant is typically shown a small set (i.e., five or six items) of visual objects (i.e., letter, digits, shapes, colors, etc.) called the *memory set*, and asked to commit them to memory. The participant is then shown a number of *probe* items and asked to indicate whether or not each probe item belongs to the memory set. By varying the number, type, and discriminability of items in the memory set, researchers gain insight into characteristics of human memory — in particular, differences in ability to recall various types of information and the speed with which different types of information are recalled [Sternberg, 1966] [Ashcraft, 1989] [Wickens, 1992].

Studies by Cavanagh [1972] and by Schneider and Shiffren [1977], with experiments using Sternberg tasks, found differences in cognitive processing speed for various graphical coding devices. Cavanagh [1972] established the following order, from most to least rapidly processed:

digits < colors < letters < words < nonsense syllables < geometric shapes < random forms

The study also found that codes with faster processing speeds had larger memory spans. Schneider and Shiffren [1977] confirmed the ordering, although with fewer coding devices; they also found a practice effect, so that as experience with the coding devices increased equally, performance differences decreased. Tan [1990] experimented using divided-attention tasks (a tracking task and a Sternberg task, performed simultaneously) and found a difference in ordering by processing speed:

digits < words < letters < colors < shapes

Subjective workload was also monitored and established another order, from least demanding to most demanding code to process [Tan, 1990]:

digits < letters < words < colors < shapes

Tan speculates that the difference in orderings produced by his study is due to the distracting primary tracking task, which placed demands on visuospatial memory [1990]. All these studies report large individual differences in performance, producing different orderings for different users. They also suggest that subjective assessment of difficulty in using a coding scheme may be an important measure in determining usability.

2.1.4 Ranges of Graphical Devices

Psychophysics also provides guidance on the number or range of colors, sizes, shapes, and other devices that may be used in graphical codes. As described above in the discussion of color coding, the number of instances of each device is important, since that number limits the precision, range, or number of values encoded with that device. The conservative recommendation is to use only five or six distinct colors or shapes [Cahill & Carter, 1976] [Carter, 1982] [Nagy & Sanchez, 1990] [Treisman & Gormican, 1988] [Moraglia, Maloney, Fekete & Al-Basi, 1989] [Umbers & Collier, 1990], presumably because that number is within the capacity of most humans to memorize [Miller, 1956] and most humans can reliably distinguish five to eight distinct colors [Smith, 1962] [Umbers & Collier, 1990]. However, more recent research suggests that from 10 [Cahill & Carter, 1976] to as many as 18 [Smallman & Boynton, 1990] colors may be used for search tasks.

We have found no studies that establish a maximum number of shapes or sizes that can be used effectively. We suspect the limiting factors will prove to be related to human memory and to the number of readily discriminable shapes and sizes that can be developed within other limitations imposed on a given system.

2.1.5 Legends

Surprisingly, we have found no discussion of the impact of using legends, either on usable code size or on search speed. Some guidance may be found in studies of human memory for various coding devices. In a discussion related to the number of colors which may be used in a display, Kosslyn [1985] expresses concern about the amount of information conveyed in a key, which the user must memorize or constantly consult. In our opinion, the possible involvement of working memory in the use of graphs and iconic displays potentially explains results which limit color code size to five or six. However, we know of no evidence to explain the use of working memory in visual search or identification tasks.

2.2 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" [Cleveland & McGill, 1985, p. 828]. Cleveland and McGill stipulate that they are concerned with the perception of such quantitative data as "numerical values of a variable, such as frequency of radiation or gross national product, that are not highly discrete; this excludes categorical information, such as type of metal and nationality..." [Cleveland & McGill, 1985, p. 828]. They identify a number of *elementary perceptual tasks* associated with graph comprehension. These elementary perceptual tasks are not *user tasks*; instead, they are a

fundamental part of the cognitive process of using a display, akin to decoding information represented in a graphical code. Where ">" signifies greater perceptual accuracy and items in parentheses yield equal accuracy, these perceptual tasks are [Cleveland & McGill, 1985, p.830]:

position along a common scale > position on identical but non-aligned scales
> length > (angle, slope (with theta not too close to 0, ∏/2, or ∏ radians])
> area (volume, density, color saturation) > color hue

Density refers to the amount of black that is present, or to gray-scale value. In this context, color saturation appears to mean the amount of hue versus white that is present in a display which is not black and white. We note that this list differs by one task between two publications, with color saturation included in the 1985 publication in lieu of curvature, which appeared in the 1984 article. Cleveland and McGill have empirically verified this ranking through experiments in which subjects were shown two geometric objects of the same type and were asked to determine what percentage the smaller magnitude was of the larger. For the ordering shown above, tasks higher in the list were performed with greater accuracy than those lower in the list. The error of judgment reported is the absolute difference between the judged percent and the true percent of difference.

We find two problems with these experiments. First, in our experience, graphs and graphical displays rarely call for judgments between just two objects. Rather, a large number of objects must typically be scanned to select objects of interest for attentive study. Patterns within the display as a whole may also be of interest. Secondly, we have found no details of Cleveland and McGill's experiments pertaining to color hue or color saturation, as the published data focus on position, length, and area, to the exclusion of color. We question that adequate precautions were taken to use colors well separated in color space. We are also struck by the difference in the ranking of color hue relative to area and density in comparison with the results presented by Christ's [1975] studies of visual search and identification tasks. There is also a discrepancy between Christ's data on brightness relative to color and Cleveland and McGill's ranking of color saturation as more accurately perceived than color hue.

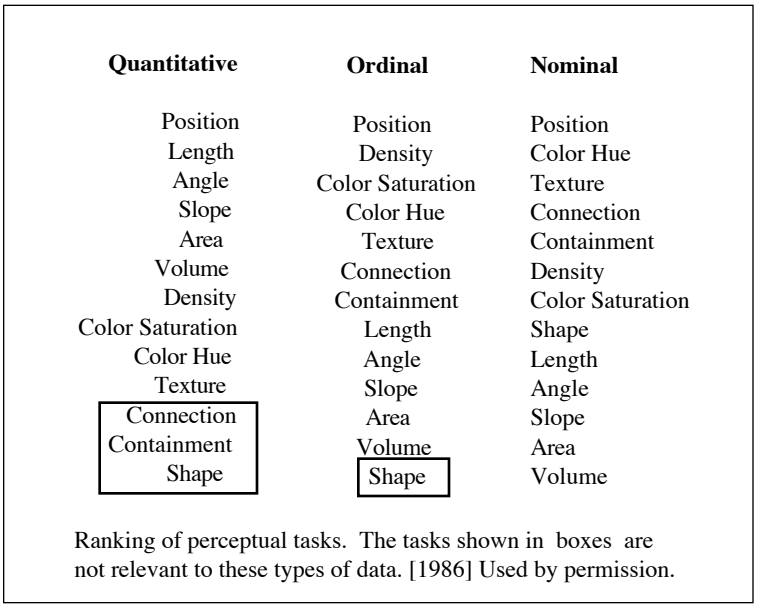
In a meta-analysis (see her reference list for sources on techniques) of 39 earlier studies of graphical perception, Carswell [1992] evaluates the basic task model for predicting graphical efficiency. Like Cleveland and McGill [1984] [1985], Carswell focuses solely on quantitative data. Her study also yields statistical validation for Cleveland and McGill's ordering of perceptual tasks. Where ">" signifies greater perceptual and items in parentheses yield equal accuracy, Carswell found the following pattern (p. 550):

position (aligned) > position (non-aligned) > length > (angle, slope)
> area > (volume, density, saturation)

2.3 GRAPHICAL LANGUAGE DEVELOPMENT

Graphical language development is based on the assertion that graphical devices communicate information equivalent to sentences [Mackinlay, 1986a & b], with attention to appropriate use of each graphical device. In his discussion of graphical languages, Mackinlay [1986a] presents rankings of the effectiveness of various graphical devices in communicating quantitative, ordinal, and nominal data about objects of interest (see Figure 2.1). Mackinlay concluded that position is the most effective graphical device for communicating nominal data, followed in order by color hue, shape, length, angle, slope, area, volume, and other attributes for this purpose. For presentation of ordinal data, his ranking of graphical devices is color saturation or density, color,

length, angle, slope, area, volume, and shape. For expressing quantitative data, Mackinlay ranks length, angle, slope, area, volume, density, color saturation, and color hue in this order. Although based on studies of psychophysical research and graphical perception, Mackinlay's classification and ordering scheme has not itself been experimentally validated [Personal communication at CHI'97].



Quantitative	**Ordinal**	**Nominal**
Position	Position	Position
Length	Density	Color Hue
Angle	Color Saturation	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color Saturation
Color Saturation	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Ranking of perceptual tasks. The tasks shown in boxes are not relevant to these types of data. [1986] Used by permission.

Figure 2.1 — Mackinlay's Rankings of Graphical Devices

2.4 OBSERVATIONS ON RESEARCH OVERVIEW

Rankings of the accuracy with which various graphical devices are perceived and of the kinds of information effectively conveyed by graphical devices are significant. As Cleveland and McGill [1985] observe, "...data should be encoded so that the visual decoding involves tasks as high in the ordering as possible, that is, tasks performed with greater accuracy" [p.828]. However, we find the theories presented by Mackinlay [1986] and by Cleveland and McGill [1984] [1985], particularly as they relate to color hue, to be at odds with our expectations based on experimental evidence from studies of visual search and identification tasks.

What guidance we have on the types of information best represented by various graphical devices is largely conjectural, with experimental verification [Cleveland & McGill, 1985] that relies on questionable assumptions about use of graphs and graph-like displays. However, the value of a classification system or hierarchy of accuracy, such as those presented in [Mackinlay, 1986] and [Cleveland & McGill, 1985] is evident, in that both human and automated layout artists should use the most appropriate, most accurately perceived graphical devices to communicate information.

Literature makes it clear that no single graphical code 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 codes which support the range of tasks users are likely to bring to a system, while also supporting individual differences of the user population. In particular, points of disagreement among studies discussed above offer both design challenges and opportunities for further research. The next chapter discusses how we used graphical devices suggested by these studies in a user interface design for user-controlled visualization of Envision search results and our use of the Envision system as a test bed for empirical studies.

CHAPTER 3. GRAPHICAL ENCODING ISSUES IN DESIGN OF THE ENVISION USER INTERFACE

3.1 TEST BED: ENVISION DIGITAL LIBRARY

The test bed for this proposed research was the Envision multimedia digital library of computer science literature [Fox, Hix, Nowell, Brueni, Wake, Heath, & Rao, 1993] [Heath, Hix, Nowell, Wake, Averboch, Labow, Guyer, Brueni, France, Dalal, and Fox, 1995] [Nowell 1997] [Nowell & Hix, 1993] [Nowell, France & Hix, 1997] [Nowell, France, Hix, Heath, & Fox, 1996]. Named after Tufte's book [1990], Envision provides full-text searching and full-content retrieval capabilities and serves computer science researchers, teacher, and students at all levels of expertise.

To determine user needs for Envision, we conducted intensive interviews with potential users, all established researchers in computer science and information science. We encouraged "blue sky" thinking to discover system capabilities interviewees long for that are not readily available in modern libraries, on-line or otherwise. While it has been shown that the "perfect 30-item on-line search" is not a realistic goal [Bates, 1984], interviewees nonetheless told us they fear being overwhelmed by vast results sets. Users wanted to home in on the document space relevant to their immediate interests and explore that space in many ways. Accordingly, Envision's vector-space retrieval system [Fox, France, Sahle, et al., 1993] [Salton, Wong & Yang, 1975] uses statistical pattern matching to determine document similarity to queries, returning relevance-ranked results sets containing only documents with highest relevance to each query.

Beyond ready access from their offices, chief among interviewees' wishes was the ability to identify and explore patterns in the literature. Some asked for visual representations, while others wanted ways to see connections not visible with current tools. Thus, while initial planning for Envision did not call for development of an information visualization system, we turned to visualization in attempting to meet user needs. We have limited our efforts to two-dimensional visualization so that Envision is accessible on most current desktop systems. Development proceeded iteratively, with extensive usability evaluation involving a wide range of participants throughout Envision's development life cycle.

3.2 GRAPHICAL DEVICES USED IN VISUALIZATION

Envision's visualization design is the *Graphic View Window*, shown in Figure 3.1 with its related *Item Summary Window* and small *Query Window*. The Graphic View resembles a number of other designs — all using displays resembling a scatter plot for visualization of non-hierarchical data — including Bead [Chalmers & Chitson, 1992], FilmFinder [Ahlberg & Shneiderman, 1994], VIBE [Olsen, Korfhage, Sochats, et al., 1993], and SemNet [Fairchild, Poltrock & Furnas, 1988]. Each document in an Envision search results set is shown in the Graphic View as an icon. In Figure 3.1, the circular "bubbles" in the Graphic View represent single documents, with relevance ranking numbers shown as labels below the circular icons. For example, in Figure 3.1, see icon numbers 44 and 45 in the top two rows.

Figure 3.1 — Envision's Search Results Display, Showing Small Query, Graphic View, and Item Summary Windows. (Link to page 17 — Fig3_1.gif — 33K)

Elliptical icons represent sets of documents whose icons collide, or fall at the same location in the graph, with the number of documents represented shown inside the ellipse. The labels below these "cluster icons" show rankings of the two most relevant documents in the cluster. For example in Figure 3.1, see the elliptical icon in the second row, where author is Ralph P. Hill. This icon represents three documents of equal relevance, the most relevant of which are ranked 32 and 33. The Item Summary, near the bottom in Figure 3.1, shows a textual listing of bibliographic data for documents whose Graphic View icons are selected by the user, indicated by bold outlines as shown in Figure 3.1, e.g., documents with icon number 1, in the third row. Item Summary text lines are related to their icons by icon label. Because graphical encoding issues were not a concern in the design of Envision's Query Window, it is not discussed herein.

The Graphic View design gives users control over the semantics of six graphical devices:
- icon position along the x-axis and y-axis,
- the alphanumeric label associated with each icon,
- icon color (saturation, for gray-scale monitors),
- icon shape, and
- icon size.

Since users' perceptual strengths vary and users' decision criteria reflect their current information needs, each graphical device is user-controllable to represent different document attributes as a user desires. The flexibility of Envision's Graphic View enabled us to experimentally validate a rich variety of graphical devices. Our design decisions regarding these graphical devices and their relationship to graphical encoding issues are described in Section 3.4.

3.3 DOCUMENT ATTRIBUTES FOR VISUALIZATION

Envision provides access to several document attributes which might be graphically encoded:
- author names,
- publication year,
- index terms,
- estimated relevance to query (a probability expressed decimally as a percentage),
- relevance rank (an integer), signifying where the document would fall in a list ordered by relevance,
- Envision database document ID, and
- document type (e.g., book, journal article, proceedings article, video, hypermedia, etc.).

All attributes available for visualization are either nominal or quantitative, as the remaining class — ordinal data — is really either quantitative data with lost precision or it is a special arrangement of nominal data. Table 3.1 summarizes our decisions about which document attributes were encoded with each graphical device in Envision, indicating which were originally implemented and which were included explicitly for our study (as discussed in Section 3.5).

Table 3.1 — Document Attributes Available for Visualization, With Chosen Graphical Devices

Document Attribute / **Graphical Device**	Position	Icon Label	Icon Color	Icon Shape	Icon Size
Author Name	•				
Publication Year	•				
Index Terms	•				
Est. Relevance	•		*	*	*
Relevance Rank		•	•		
Document ID		•			
Document Type	•		•	*	*

• Implemented * Added for Study

3.4 VERSION 1 OF ENVISION

In the paragraphs that follow, we detail design constraints and decisions implemented in the first version of Envision. These choices are represented by solid bullets in Table 3.1. Because shape and size encodings were not implemented at the time of Envision usability evaluation, we defer discussion of those graphical devices to section 3.5, where we also discuss use of color to encode document type.

3.4.1 Position Encoding

Largely because of its power as an encoding device [Cleveland & McGill, 1984] [Cleveland & McGill, 1985] [Mackinlay, 1986], position of a document icon relative to the x- or y-axis represents more document attributes than any other graphical device in the Graphic View: author names, publication year, index terms, estimated relevance, or relevance rank. Figure 3.1 shows author names on the y-axis. Relevance to the query is usually the most important attribute of each document in a results set. Since position is the most accurately perceived encoding [Cleveland & McGill, 1984] [Mackinlay, 1986], it is natural to display relevance in terms of position on the x- or y-axis, or redundantly on both. Captions on the ends of each axis, such as "Most Relevant" and "Least Relevant," clarify the semantics of position in these cases. (See x-axis in Figure 3.1.) Other quantitative attributes of documents, such as number of sources cited, times cited, or document size, might also be readily encoded as position on an axis, but these values are not currently reported by the Envision database.

Some important attributes of documents, such as author names, index terms and publication years, cannot be readily represented by graphical devices such as color or shape because these encodings allow too few instances to be displayed at once. Textual icon labels could convey author names or index terms, but the length of these text strings requires greater spacing between icons, reducing the number of icons visible at once and thus limiting the ability to show patterns in the display. Variability of string length also greatly complicates the task of automating graph layout. Mackinlay's classifications suggest that position is the most effective graphical device for

conveying all types of information. Since research in graphical perception also shows that position on a common scale is the most accurately perceived graphical device (and psychophysical evidence offers no contraindications), Envision can display author names and index terms as x-axis or y-axis headers in the Graphic View. Figure 3.1 shows author names on the y-axis.

3.4.2 Alphanumeric Label Encoding

Since the icon label serves as a link to related textual descriptions in the Item Summary, icon labels must be unique to avoid confusion. Thus, some document attributes that might otherwise seem appropriate for labels are not acceptable for that purpose. For example, publication year is a consistently short string that could be displayed as an icon label, but publication year is highly likely to be non-unique, so it is available for display only as x-axis or y-axis headers. Figure 3.2 shows publication year on the x-axis. Within a single search results set, relevance rank is unique for each document and is the default icon label (the small numbers below the circles in Figure 3.1). A document's Envision database ID, while not inherently interesting to users, provides a convenient means to compare results of two searches. Since document ID is a short, unique character string, we allow the user to choose it as an icon label. Because of our requirement for uniqueness, the only document attributes used as icon labels are relevance rank and Envision database identification number (ID).

3.4.3 Color Encoding

As discussed in Chapter 2, researchers differ widely on the impact of color coding. Mackinlay [1986] ranks color hue very low relative to area and density for conveying quantitative and ordinal data, while Christ [1975] [1984] affirms the power of color coding to facilitate visual search and identification tasks. Christ's [1975] [1984] data on brightness relative to color also disagrees with Cleveland and McGill's [1984] [1985] ranking showing color saturation to be more accurately perceived than color hue. Further, while some researchers assert that color does not form a psychological continuum and should not be used to convey quantitative information [Cleveland & McGill, 1985] [Kosslyn, 1985] [Wickens, 1992], others find color works well as a continuum for continuous variables [Umbers & Collier, 1990]. Given the ongoing debate, our design decisions on use of color coding follow. We began with a decision to treat color as an icon single attribute, rather than as either a composite of hue, saturation, and brightness, or as varying percentages of red, green, and blue, as used in some research [Ware & Beatty, 1988]. This decision allowed us to work within a wider range of stimuli and reduce reliance on hue alone, so that color-impaired users are better supported.

3.4.3.1 COLOR AS A CONTINUUM

Color coding is regularly used as a quantitative continuum on displays such as television and newspaper weather maps that convey temperature variations. Christ [1984] reports that increased experience and familiarity with any graphical encoding increases its effectiveness. Since color codes yield the most rapid visual search and relevance is a likely document characteristic for provoking search tasks, Envision uses a color continuum to show ranges of document relevance. Eventually, color will also be used to show ranges of other quantitative attributes — document size, number of sources cited, and times cited.

3.4.3.2 SEMANTIC VALUES OF COLOR

The semantic value of color has been largely left to convention [Umbers & Collier, 1990] [Rice, 1991] [Wickens, 1992]. The warm-to-cool color continuum is familiar because of its conventional use in weather maps, where warmer colors represent higher temperatures. Furthermore, childhood games use "warmer" to mean being closer to a search target, while "colder" means being farther away. Also, recent research [Merwin & Wickens, 1993] suggests that users have less difficulty understanding a blue-green-yellow scale than other color patterns for encoding continuous data. Accordingly, Envision's warmer colors represent higher relevance values and cooler colors signify less relevance. Envision uses a warm bright orange to represent more-relevant documents, a dull "temperature neutral" green to represent those of medium relevance, and cool pale blue to represent less-relevant documents (see Figure 3.1.). When users mark a document relevant, its icon color changes to a vibrant warm red, while documents marked not relevant are represented in white, the color of the background, to attract less attention (see Figure 3.3). Separation of these colors was tested informally with several participants prior to any use of Envision for this experimental study. To accommodate users with color impairments, the value of colors in Envision is adjusted so that a scale from darkest to lightest is evident and a gray-scale version is available.

Mackinlay [1986] suggests that color hue is second only to position in conveying nominal data, when the number of possible attribute instances is small enough. Psychophysical evidence agrees, while the issue is irrelevant to graphical perception. Envision uses color to show document type, a relatively small attribute range (e.g., journal article, proceedings article, and book). However, the number of colors in the code is currently limited to between five and ten, as suggested by literature [Cahill & Carter, 1976] [Christ, 1984] [Wickens, 1992]. This number is too small for color to represent wide-ranging nominal attributes such as author names or index terms.

3.5 VERSION 2 OF ENVISION — ENHANCEMENTS TO SUPPORT EMPIRICAL STUDY

Our review of psychophysics and graphical perception (see Chapter 2) 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, since a primary purpose of this research was to test such assumptions, those visualizations were included in the empirical study. For encodings expected to perform poorly, we made every effort to devise the most effective code possible for use in the experiment and these designs were implemented in a separate version of Envision. Encodings in this group include use of icon shape to encode relevance and use of icon size to encode document type. We also designed three other encodings that were originally planned for Envision but were not implemented when the project ended; these included use of icon size to encode relevance, color to encode document type, and shape to encode document type. Discussion of these designs follows.

As we added visualizations, other changes were needed, as well. In particular, the layout of the Graphic View control/legend space changed to accommodate added legends for icon shape and icon size. The modified legend may be seen in Figure 3.2, which shows the Graphic View Window from version 2 of Envision, designed specifically to support this research. Another change that may be seen in Figure 3.2 is the absence of cluster icons. Because cluster icons are not subject to either shape or size encoding, it was necessary to remove this feature for the research version of Envision. Without the cluster icons, only the most relevant document of a group that would all appear at the same point on the graph is shown. The less relevant documents from the group are simply "dropped" from the results set, which is not an acceptable condition for actual use.

Figure 3.2 — Graphic View Window from the Research Version of Envision. (Link to page 23 — Fig3_2.gif — 33K.)

3.5.1 Constraints in Design

We chose to use only three instances of each graphical device because of the need to maintain distinguishability among 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. After some trial-and-error experimentation, we concluded that, within size limits, 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.

Several other constraints were applied during icon design, chiefly to avoid confounding by use of other graphical devices sometime used in codes (e.g., orientation, texture, etc.). First, when size is uncoded, we made all icons for a shape code approximately the same area. When size is coded, we made all icons for a given size category (e.g., small, medium, or large) 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, we made all icons have a true vertical orientation, rather than allowing various degrees of slanting, which is sometimes used as another graphical device.

3.5.2 Shape Encoding

Icons designed for using shape to convey document type were bitmaps and had additional constraints: equal numbers of black pixels were required for all three shapes and likewise for the area that would show fill color. When icon size is used to encode relevance, the three bitmaps of each size were the same area and had the same constraints on dark and light pixels.

Designing mnemonic icons to convey document type was unexpectedly difficult because the three document types by far the most prevalent in Envision are books, journal articles, and proceedings articles — all predominantly text. The icon we designed for books presents the spine of the book to the user, with the cover rotated into the third dimension so that the pages are turned away. Dark lines down the spine and across the cover suggest text for title and author. For journal articles the icon is essentially a rectangle that is much wider than it is tall, suggesting an open journal. The "pages" shown include dark lines representing text and open boxes to represent plates or figures. A rectangle of the same size, but turned 90 degrees to sit vertically, is used for the proceedings article icon. It has bold dark lines to signify a title and two columns of lighter lines to represent text. Figure 3.3 shows an enlargement of one Graphic View cell, in which all three shapes are shown. Figure 3.4 shows the Graphic View legend, including the legend for shape as document type. The icons we designed are thus more similar in shape than might be possible in a more varied database, but we suspect this problem is common in many settings (e.g., aircraft cockpit displays, target acquisition systems, etc.) where a variety of closely related, similar objects must be presented.

Figure 3.3 — Enlargment of Envision Graphic View Cell, Showing Colors and Shapes for Document Type. (Link to page 25 — Fig3_3.gif — 17K.)

Figure 3.4 — Envision Graphic View Legend, Showing Shape, Color, and Size Codes Used to Convey Document Type. (Link to page 26 — Fig3_4.gif — 17K.)

Shape has been said to be inappropriate for encoding quantitative data because shape does not form a natural continuum [Cleveland & McGill, 1984] [Kosslyn, 1985]. However, we believed that shape might be an effective code for conveying quantitative data if a continuum of shapes could be established in the legend. Accordingly, we used five-pointed stars for documents in the top 30 percent by relevance, diamonds for the middle 30 percent of documents by relevance, and upward pointing triangles for the 40 percent of documents of least relevance. Figure 3.5 shows an enlargement of a Graphic View cell, in which all three shapes for relevance are shown.

Figure 3.5 — Enlargment of Envision Graphic View Cell, Showing Colors and Shapes for Document Relevance. (Link to page 27 — Fig3_5.gif — 17K.)

Figure 3.6 shows the Graphic View legend, including the legend for shape as document relevance.

> Figure 3.6 — Envision Graphic View Legend, Showing Shape, Color, and Size Codes Used to Convey Document Relevance. (Link to page 29 — Fig3_6. gif — 17K.)

The breakdown into these relevance categories was selected to ensure a reasonably equitable distribution of the various shapes to support our experiments. This distribution also reflects relative user disinterest in documents of lower relevance ranking within a set and is used in all encodings of relevance in the experimental version of Envision. Thus, we used easily distinguishable geometric shapes with decreasing numbers of points and sides as a code for conveying relevance (quantitative data). Informal evaluation of these design decisions always resulted in comments by observers such as "Stars are best, aren't they?"

For the shapes used to encode relevance, areas varied by no more than five percent — a variability necessary to achieve perceptual equality in area and because of deviations caused by floating point round-off. All three geometric shapes were the same height, except when size was used as a code. In this case, the small geometric shapes were all the same height and approximately the same area. The same was true for the middle size and the large size, though the height and area increased for each of these.

One pre-test observer commented that the upward pointing triangle used to convey least relevance seemed wrong to him, in that its upward thrust gave it a meaning of increasing value. In response, we tried inverting the triangle. The resulting downward pointing triangle seemed both larger and more powerful, though its size remained the same. Also, the inverted triangle seemed unstable and was therefore distracting. Accordingly, we returned to the upward pointing triangle, relying on the legend to clarify the meaning for users.

3.5.3 Color Encoding

We required colors used to represent document types to meet three conditions: 1) they had to be easily distinguished from each other, 2) they could not be colors used elsewhere in the Envision design, and 3) they could not be so saturated as to make it impossible to perceive internal detail if used as fill colors with the icons used for document type. Thus pale blue, mossy green, bright orange, red, and white were eliminated because of use in the relevance color code and for marked icons, while colors such as wine red, royal blue, and purple were too dark to use as fill colors. We arbitrarily selected bright pink for books, French blue for journal articles, and lime green for proceedings articles. Figure 3.3 shows the colors used to encode document type. Figure 3.4 shows the Graphic View legend, including the legend for color as document type.

3.5.4 Size Encoding

Documents have a size attribute and use of that size seemed the reasonable approach to assigning sizes to document types for a size-as-type code. Thus, books, which are usually longer than both journal articles and proceedings articles, were assigned the largest size. Journal articles are commonly longer than proceedings articles and were assigned the middle size. And proceedings articles, typically the shortest of these publications, were assigned the smallest size. Figure 3.7 shows a enlargement of a Graphic View cell, with size encoded and shape uncoded. The same three sizes are used for both type and relevance. Figure 3.6 shows the Graphic View legend, with size as document type.

Figure 3.7 Enlargment of Envision Graphic View Cell, Showing Icon Sizes Used to Convey Document Type or Relevance. (Link to page 31 — Fig3_7.fig — 17K.)

In designing for use of size to convey relevance, we assumed that documents of most relevance are those users most want to locate, so the largest sized icons were assigned to the top 30 percent of documents by relevance. The next 30 percent of documents by relevance received the middle shape, and the bottom 40 percent received the smallest shape. Figure 3.6 shows the Graphic View legend, with size as document relevance. This is the same partitioning used when shape or color encodes relevance.

During pilot testing, discussed further in Chapter 4, all three participants commented that it was easy to distinguish between the small and middle sized circles used when shape is uncoded, but that it was difficult to distinguish between the middle sized and large circles. We responded by increasing the diameter of the largest circles by two pixels, the maximum allowed by our icon density and cell size constraints. Interestingly, participants in the experiments remarked that it was easy to distinguish between medium sized and large icons, but that it was now very difficult to see the difference between small and medium icons.

3.6 TASKS PERFORMED WITH THE GRAPHIC VIEW

Thus far, we know of two efforts to classify tasks performed using an information visualization interface to a document database or library system such as Envision, both presented at workshops. The first of these, by Dubin [1995a], centers on the VIBE interface [Dubin 1995b] [Olesen, Korfhage, Sochats, Spring, & Williams, 1993] and does not generalize well to other systems, including Envision. More recently, the FADIVA (Foundations for Advanced Information Visualization) workshop at the ACM SIGIR '96 (the Association for Computing Machinery's Special Interest Group on Information Retrieval) conference in Zurich, Switzerland, began developing a list of tasks supported by the visualization interfaces familiar to the group (personal communication from Mark Rorvig, workshop co-chair, July, 1997), which is thus far unpublished.

Using Envision, *visual search tasks* are performed when users attempt to determine if a document icon meeting a specific requirement is present in the results display; success is possible only if that requirement is represented by one of the graphical devices. *Identification tasks* are performed when a graphical device (e.g., icon shape or color) communicates nominal data about documents in the Graphic View and an Envision user attempts to locate items of a particular type. Envision users perform *graphical perception tasks* when making comparative 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.

The Graphic View supports users in making decisions about which works to examine in potentially large sets of documents, as well as providing support for detecting patterns in a results set. In particular, the design facilitates integration tasks that are difficult to accomplish using the textual results lists provided as search results by most digital libraries and online public access catalogs (OPACs). For example, usability evaluation (discussed further in Section 3.7) has shown that users of Envision can readily locate the three most relevant works by a given author, the number of highly relevant works published in a given year, or the distribution over publication years of works indexed by a particular term.

Since users' perceptual strengths vary and their decision criteria reflect their current information needs, each graphical device in the Graphic View represents different document attributes upon user choice. To better support non-integration tasks (see Section 2.1.2.2) such as finding the most relevant document, three graphical devices — color, size, and shape — may be effectively neutralized by making them uniform for all icons. All graphical devices may be used to redundantly express relevance, the document characteristic we anticipate to be of greatest interest. Encoding information non-redundantly using different graphical devices creates an object display, supporting integration tasks such as finding the most relevant document by a specific author.

Figures 3.1, 3.2, and 3.8 through 3.12 show results of a single Envision search in a variety of layouts, revealing different characteristics of the set in each layout. The query used for each of these figures was "user interface management system; UIMS" in the title field.

Researchers interested in comparing publication patterns among authors might choose the layout in Figure 3.8, showing *authors* on the *y-axis* and *publication year* on the *x-axis*. This display is fairly sparse, reflecting the reality of publication patterns in the topic and collection, but it presents much information about each document. Using this layout, in which *icon shape* and *color* encode *document type*, the user who believes that journal articles contain more significant work than proceedings, or that proceedings articles are more likely to contain cutting-edge research, can distinguish these items from books, which might contain more in-depth coverage. The layout uses

icon size and *label* to show *document relevance*. Redundant encodings of this kind aid in quick, reliable perception of important features [Carswell & Wickens, 1987] [Wickens & Andre, 1990]. There are thus a total of four characteristics revealed for each document included in the figure (e.g., author, publication year, document type, and document relevance), yet the display remains aesthetically pleasing and uncrowded.

Figure 3.8 — Envision Graphic View with Author on the Y-Axis and Publication Year on the X-Axis. Color and Size Encode Document Relevance, While Shape Encodes Document Type. (Link to page 34 — Fig3_8.gif — 33K.)

A user seeking more terms to use in query revision might choose the layout in Figure 3.9, with assigned *index terms* on the *y-axis* and *publication year* on the *x-axis*. *Document relevance* is encoded redundantly in *icon color* and *size*. Clustering of relevant documents by different index terms may reveal relationships among the categories. Pairing index terms with either author or publication year in the Graphic View (not shown) can reveal other commonalities among indexed topics.

Figure 3.9 — Envision Graphic View with Index Terms on the Y-Axis and Publication Year on the X-Axis. Color and Size Encode Document Relevance, While Shape is Uncoded. (Link to page 35 — Fig3_9.gif — 33K.)

The utility of visualizing index terms obviously depends on the quality of indexing. Envision currently visualizes only index terms or keywords that have been assigned by authors or editors — clearly a major limitation, especially since our vector-space search system does full-text searching. Furthermore, both prevalence and quality of assigned index terms vary widely among segments of the collection, from copious to completely absent, and from controlled descriptors through ordinary language to cryptic abbreviations. Additionally, since Envision currently visualizes only one index term per document (the first listed), neither the full range of assigned terms nor the true amount of overlap among them is available to users. Visualizing multiple index terms per document presents a number of usability problems, discussed in Section 3.8 and Chapter 6.

—

In a different layout, putting *publication year* on the *y-axis* and estimated *relevance* on the *x-axis*, as in Figure 3.10, creates a graphic picture of increasing research within a topic area. Icon *color* again shows document *relevance*, so that icons for the most relevant documents are orange and further right than other icons. In this display, the user has marked the documents represented by icons 1 and 3 are *useful*, turning the icons *red*, while icon 23 is *white*, signifying the document has been marked *not useful*.

Figure 3.10 — Envision's Graphic View with Publication Year on Y-Axis and Relevance on the X-Axis. Color Encodes Relelvance, While Shape and Size are Uncoded. (Link to page 37 — Fig3_10.gif — 33K.)

Finally we present two configurations that allow the user to view the entire results set without scrolling. In the first (Figure 3.11), both *x-* and *y-axes* have been set to show estimated *relevance* This layout is particularly useful for studying relative relevance of documents in the set. The display reveals drop-offs in the estimates, giving an information retrieval researcher insight into performance of the underlying search engine on the query, and allowing typical end-user to pick a highly ranked subset to examine.

Figure 3.11 — Envision's Graphic View with Relevance on Both Axes. Color Encodes Relevance, While Shape and Size are Uncoded. (Link to page 38 — Fig3_11.gif — 17K.)

Figure 3.12, which also shows the entire results set, presents *document type* on the *x-axis* and estimated *relevance* on the *y-axis*. *Relevance* is also encoded redundantly in *icon color, shape*, and *size*. Putting relevance on the y-axis rather than the x-axis invokes a different metaphor: that the most relevant items, like cream, are rising to the top. Giving users control over layout allows them to choose comfortable metaphors. This may be one reason that users report high satisfaction with the Envision interface.

Figure 3.12 — Graphic View with Document Type on the X-Axis and Estimated Relevance on the Y-Axis. Color and Size Encode Relevance, While Shape Encodes Document Type. (Link to page 39 — Fig3_12.gif — 33K.)

We have described user tasks with the Graphic View in terms of perceiving patterns in a results set. There are many more patterns users might seek with different combinations of Envision's graphical codes, as any combination may be meaningful to some user. Of course, users will also use Envision to obtain detailed information about one or more documents and to compare items of interest, using additional information available in the Item Summary Window and by viewing documents.

3.7 ENVISION USABILITY EVALUATION

In our early cycles of usability evaluation of the Envision user interface, we were specifically concerned with proof of concept: Did users understand relationships among the windows and the graphical objects within them? Did users make sense of the complex display, based on graphical devices used in the design, and find this a desirable way to view search results?

Using a SuperCard prototype for our earliest usability evaluations, participants performed benchmark tasks (e.g., finding three works published by a given author in a specified year, locating the title and author of the most relevant work) and then responded to 28 subjective questions. Results were excellent for most of the 42 objective measures of usability. Average task completion times equaled or bettered our planned usability goals for 13 of 16 measures of time. For 14 counts of errors and 12 counts of Help usage, all measures met our usability goals. For 26 of the 28 subjective questions, the average user response was positive, on a scale of -3 to +3. Both negative responses pertained to the Help system, which was present in only rudimentary form at the time of our usability evaluation. Further, all participants were strongly positive in their evaluations of the design concept. They liked the variety of information presented visually in the Graphic View and the constrained flexibility offered for customizing search results layout. All commented positively on the power the Envision interface provides to the user. No participants had any difficulty in recognizing relationships among the windows and objects in them. Minor changes to the Graphic View design resulted from the evaluations, such as changing one label from plain text to bold and revising the layout of control buttons. Additional details of that evaluation are in [Nowell & Hix, 1993].

For a later second cycle of usability evaluations, we used the X-Windows implementation of Envision. In addition to evaluating changes resulting from earlier prototype evaluations, we focused on features that were not fully implemented in the prototype: controlling the number of items in the results set, changing icon attribute semantics, and more extensive exploration of relationships among the windows. Five computer scientists (one faculty, two graduate students, and two undergraduates) each were given a one-page "Getting Started" handout and were allowed ten minutes to explore Envision's features before performing 11 tasks. A typical task required the participant to create a query meeting specified criteria, have Envision complete the search, and then use Envision's search results display to locate documents fulfilling various requirements. To ensure that participants used various aspects of the Graphic View, some tasks required participants to change the semantics of icon attributes (e.g., "change the x-axis setting to show publication years"), while for other tasks the various icon semantics were left to participant discretion. A description of a sample task is shown in Figure 3.13. Use of the Item Summary window was required for some tasks, but many others could be completed using only the Graphic View. Upon completion of all tasks, users were given time for additional free exploration of the system. Throughout each evaluation session we recorded verbal protocol and critical incidents.

Four of the 11 tasks were designated as benchmark tasks for objective measurement of user performance. Three benchmark tasks focused on initial use of a design element, while the fourth studied learning curve. Performance measures included task completion time, number of errors, and number of questions asked (since the on-line Help system was not yet implemented). For task completion time, our goal was that mean participant time should not exceed the time required for one of the interface designers to complete the same task. For the initial task using a particular feature, we aimed for a mean error count and a mean number of questions equal to 0.2 — allowing for only one of the five participants to experience difficulty. Using the Graphic View, participants made no errors, asked no questions, and all required less time than expected to

Still working with your second set of search results,

a. Change the graphic view so that the x-axis label shows the years in which works were published.

b. Point to the icon for the most relevant work by Shneiderman.

c. Tell me in what year the work was published.

d. Change the Graphic View so it again shows relevance on the x-axis.

Figure 3.13 — Sample Task Using the Graphic View

complete the benchmark tasks—to our delight, surpassing the performance of an Envision designer.

Upon completion of the 11 tasks, each participant completed a questionnaire of 14 questions, with a scale of -3 (least satisfactory) to 3 (most satisfactory). A typical question is shown in Figure 3.14. The lowest mean response was 1.6 for a single question. Mean responses of 1.8 and 2.2 were each received for 2 questions, while mean responses of 2.4 and 2.6 were given for 4 questions each. The question shown in Figure 3.14 received a mean response of 3 for the main question, 2.6 for 7.a, and 2.4 for 7.b. The mean rating over all questions was 2.3.

7. The Graphic View Window and its iconic representation of search results are	useless -3 -2 -1 0 1 2 3 NA	valuable
7.a	overwhelming -3 -2 -1 0 1 2 3 NA	empowering
7.b	confusing -3 -2 -1 0 1 2 3 NA	illuminating

Figure 3.14 — Sample Item from Usability Evaluation Questionnaire

3.8 CURRENT STATUS, LIMITATIONS, & FUTURE DEVELOPMENT OF ENVISION

The Envision database currently contains approximately 100,000 bibliographic records, 700 full-text articles, and 16,000 scanned pages. One hypertext and one video document are also available for experimentation. In addition to basic retrieval tasks, the user interface provides capability to explore patterns in the collection, visualizing such information as number of works published annually by an author or the number of works associated with an index term.

We plan further usability evaluation to determine the desirability of allowing users to change the document characteristic represented by an icon attribute during use of the display. That is, what happens to user performance when users are allowed to change layout semantics? Given results of our latest usability evaluation and other studies [Carswell & Wickens, 1987] [Wickens & Andre, 1990], we expect some temporary loss of speed and accuracy in use of the display immediately after users change the layout.

Issues of scalability pertain to the size of results sets the Graphic View can display. We have tested the current version with results sets as large as 500 documents. We found that for some icon attribute settings (e.g., authors on the y-axis and index terms on the x-axis, not shown), the display is quite sparse, reflecting the need for a "zoom" feature that is planned but not yet implemented. Zoom will allow users to see a larger area of the scatterplot in less detail or a smaller area in greater detail. Ideally, the Graphic View could then be used as a browser for the entire collection, allowing users to zoom in on selected areas of interest.

Full use of Envision's Graphic View requires access to a number of document characteristics that are infrequently available in a bibliographic database or library system, such as document size, the number of citations contained in a work, and the number of times a document has been cited in other works. Even when these characteristics are represented in the database in some form, a visualization may be difficult or misleading. Visualizing document size appears to be a straightforward matter, dependent on page count, word count, or storage required. However, in a multimedia database, none of these is a consistent indicator of time required to use different types of works. For example, a video that can be viewed in five minutes may occupy more storage space than a book, and may have no word or page count. Some means of converting raw size values to a meaningful common scale is needed.

Visualizing "times cited by others" also presents challenges. We are developing a database of citation links for Envision that will ultimately provide not only the number of citations but hypertext links among related documents. Even so, our database will only provide information about citation links among documents in the database — a small percentage of the total number of documents about computer science. Since a visualization of "times cited by others" will show only citations from works in our collection, works heavily cited by publications not in the collection may appear to be less significant than they are. Accessing a citation index might be a solution to this problem.

One of the more interesting issues we are exploring is presentation of multi-sets — those instances when a single document belongs to multiple categories on either axis. For example, a document frequently has more than one author and is usually assigned more than one index term. Yet presenting multiple icons for one document has the potential to greatly increase display clutter, and we have questions about such a display: How will users respond when selecting or marking one

icon causes several others to highlight or change color because they represent the same document? What about a document that occurs both as a single icon and as part of a cluster?

During usability evaluation and demonstrations of Envision, users have told us they especially like the flexibility and power of the Graphic View, and they want many more visualizations. For example, we have been asked to reveal who cites whom by placing citing author on one axis and cited authors on the other — thus depicting communities of discourse, as users requested during our initial interviews. Musicians want to visualize by genre, style, and instruments required. For a medical collection, visualizations might present key symptoms, effectiveness of medications suggested per symptom, risk of drug interactions, etc. This user feedback, even more than success in formal usability evaluation, convinces us that library systems have much more to visualize than query-document similarity or semantic content and that the Envision Graphic View is a powerful, flexible design for increasing the range of characteristics visualized by a retrieval system.

CHAPTER 4. EMPIRICAL STUDY

4.1 OVERVIEW

Using the Envision Graphic View, we conducted a within-subjects empirical study of the effectiveness of three graphical devices — *icon size*, *icon shape*, and *icon color* — in communicating both nominal (document type) and quantitative (document relevance) data. These are the major independent variables in the experiment. We chose these graphical devices for investigation because of their widespread use and expected power in communication, combined with uncertainty about their actual impact. It is also possible to combine these graphical devices within a single icon, creating a multi-dimensional code that conveys more information about each document. Prior studies [Mackinlay 1986] [Christ, 1984] [Cleveland & McGill, 1985] [Carswell, 1992] support the conclusion that other graphical devices used in Envision — position on the x-axis and y-axis, and the alphanumeric icon label — yield highly accurate responses and rapid reaction times. Other graphical devices (e.g., flash rate, luminance or brightness, texture, sound, angle, etc.) were not tried in the Envision user interface design because of evidence of limited effectiveness, as described in section 2.1.1.2.

Because the ability to distinguish among the elements of a code affects the time required to complete tasks, we also conducted a second experiment to measure the discriminability of code elements. For this experiment, five sets of graphical devices were tested: the set of colors used for document relevance, the set of colors used for document types, the set of geometric shapes used for relevance, the icons used for document type in the shape code, and one set of sizes used for both document relevance and type. For each set, other icon attributes were used in their "uncoded" value, so that circles were used to test colors and sizes, and the medium green used for the "uncoded" color was the fill color used for both type and relevance shapes and for the multiply sized set.

4.2 EXPERIMENTAL DESIGN

Table 4.1 (Graphical Encodings for Each Design Point, with Distribution of Tasks) summarizes the experimental design and data collection. Each of the three graphical devices was used three ways — uncoded (i.e., held constant), representing document type, and representing document relevance. In all cases, the semantics of the x-axis, y-axis, and icon label remained constant: the x-axis showed publication year (quantitative data) and the y-axis showed index terms (nominal data), while icon label showed Envision database ID. Specifically, we studied the following *design points,* or combinations of graphical encodings:

- six unidimensional graphical codes. (See design points 1-3 and 8-10 in Table 4.1; i.e., icon color representing document type with icon shape and size uncoded, icon shape representing document type with icon color and size uncoded, icon size representing document type with icon color and shape uncoded, and in like manner with each graphical device representing document relevance.)
- six redundant two-dimensional graphical codes. (See design points 4-6 and 11-13 in Table 4.1; i.e., color and shape both representing document type with size uncoded, color and size both representing document type with shape uncoded, shape and size both representing document type with color uncoded, and in like manner with each pair of graphical devices representing document relevance.)

- two redundant three-dimensional graphical codes. (See design points 7 and 14 in Table 4.1; i.e., icon color, shape, and size all representing either document type or document relevance)
- six non-redundant two-dimensional graphical codes. (See design points 15-20 in Table 4.1; i.e., all possible combinations of one graphical device uncoded, with one graphical device representing document type and one representing document relevance).

Thus we have 20 different design points. For example, design point 1 uses icon color to represent document type, with icon shape and icon size both uncoded, as shown in Table 4.1**.**

We limited our graphical devices to three instances for each graphical code:

- three icon sizes (small, medium, and large),
- three icon shapes (for quantitative data: triangles, squares, and five-pointed stars; for nominal data: mnemonic icons based on selections from [Horton, 1994]), and
- three icon colors (for quantitative data: bright yellow-orange, green, and pale blue; for nominal data: bright pink, blue, and lime green).

4.2.1 Method: Experiment 1

4.2.1.1 PARTICIPANTS

For the first experiment, participants were 20 graduate and undergraduate students at Virginia Tech. Participants were recruited by email messages sent to two mailing lists, including a request that the message be forwarded to any interested parties. Participants were evenly divided between men and women, and were also of varying races and nationalities. Their majors ranged across a broad spectrum of disciplines at Virginia Tech: English, Psychology, Mathematics, Animal Science, Business Administration, Chemistry, Statistics, and Computer Science, to name a few. Participants were paid $15.00, with a bonus of $5.00 if all tasks were completed. All participants completed all tasks.

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 (lazy eye). We also accepted one participant who volunteered the information that he had a learning disability affecting use of symbols. We chose to accept these two participants because both are successful students and we believe their limitations exist in the Envision user community, so including them meets our commitment to "real world" conditions. We also observed that their performance during the experiment was within the bounds established by other participants.

4.2.1.2 TIME REQUIRED OF PARTICIPANTS

Pilot testing was conducted with three participants. One participant completed all trials for all 20 design points and related paperwork in under two hours. Two other participants each completed all trials for four design points and related paperwork, averaging 4.5 and 6.5 minutes per design point. Based on these results, we proceeded with an experimental plan that required only one session of under 2.5 hours per participant. This plan offered the advantage of increasing the probability of all participants completing all design points.

Table 4.1 — Graphical Encodings for Each Design Point, with Distribution of Tasks

		Graphical Encoding			Trials Involving:		
Design Point	**Symbol**	**Icon Color**	**Icon Shape**	**Icon Size**	**Type**	**Rel.**	**Both**
1	C_T	Type	Uncoded	Uncoded	12		
2	S_T	Uncoded	Type	Uncoded	12		
3	Z_T	Uncoded	Uncoded	Type	12		
4	C_TS_T	Type	Type	Uncoded	12		
5	C_TZ_T	Type	Uncoded	Type	12		
6	S_TZ_T	Uncoded	Type	Type	12		
7	$C_TS_TZ_T$	Type	Type	Type	12		
8	C_R	Relevance	Uncoded	Uncoded		12	
9	S_R	Uncoded	Relevance	Uncoded		12	
10	Z_R	Uncoded	Uncoded	Relevance		12	
11	C_RS_R	Relevance	Relevance	Uncoded		12	
12	C_RZ_R	Relevance	Uncoded	Relevance		12	
13	S_RZ_R	Uncoded	Relevance	Relevance		12	
14	$C_RS_RZ_R$	Relevance	Relevance	Relevance		12	
15	C_TS_R	Type	Relevance	Uncoded	4	4	4
16	C_RS_T	Relevance	Type	Uncoded	4	4	4
17	C_TZ_R	Type	Uncoded	Relevance	4	4	4
18	C_RZ_T	Relevance	Uncoded	Type	4	4	4
19	S_TZ_R	Uncoded	Type	Relevance	4	4	4
20	S_RZ_T	Uncoded	Relevance	Type	4	4	4

The second column of Table 4.1 shows symbols used to represent various encodings in figures and discussion that follow. Large letters represent the graphical device used (e.g., C for color, S for shape, and Z for size) and subscripted letters represent the type of data (e.g., $_T$ for document type, $_R$ for document relevance) encoded in the graphical device. Thus C_T signifies color representing document type. If a graphical device is not represented in the symbol for the design point, the graphical device was uncoded.

4.2.1.3 PLATFORM AND MATERIALS

All search results displays presented to participants were captured as a screen dump from Envision. These screen dumps and related tasks were presented to participants using a program written in SuperCard[2]. Using screen dumps displayed through SuperCard, instead of Envision itself, offered several advantages. First, using screen dumps prevented participants from inadvertently altering experimental conditions by changing the Graphic View layout during the experiment, as they might if using Envision itself. Second, using screen dumps eliminated reliance on the network connection required by Envision's distributed system. Use of the network and server would have introduced unpredictable load factors that could have affected system performance, making accurate timing of participant responses far more difficult. In addition, SuperCard has built-in functions that facilitate timing, accurate to the nearest sixtieth of a second.

Our SuperCard control program ran on an Apple Power Macintosh 9500 running System 7.5.2, with 32 megabytes of RAM and virtual memory enabled. The display was an Apple high-resolution color 20-inch monitor with automatic degaussing. Resolution was 1152 x 870 at 75 Hz, and pitch was 0.31 millimeters. The system was powered through an American Power Conversion Back-UPS 450 to ensure consistent conditions.

Individual monitors do not display colors consistently over time; they vary slightly from hour to hour and considerably from day to day. To minimize the impact of such monitor "drift," we recalibrated the monitor prior to beginning each session of the experiment, using the Colortron II colorimeter from Light Source (© 1995, Light Source Computer Images, Inc.). The Colortron II was itself recalibrated daily following the manufacturer's instructions. For monitor calibration, our target white point was 9300^{o} Kelvin, standard for Macintosh color displays, and the target gamma was 1.8. Across all calibrations, the mean measured white point was 7785.83^{o} Kelvin and the measured gamma was 1.70.

Various monitors display radically different colors in response to the same red-green-blue (RGB) input. Thus, to communicate the colors actually used in our experiments, immediately after each recalibration of the monitor, we recorded the CIELUV colorspace coordinates for each color used in Envision's color codes. The mean values for these coordinates are shown in Table 4.2. The CIELUV color space is a color description system developed by the Commission Internationale de l'Eclarage (CIE, or International Commission on Illumination), the international standards authority for color science, in 1976. The LUV color space in intended to be perceptually uniform, meaning that the distance between two colors in the coordinate space reflects human perception of the difference between two colors [Travis 1991] [Foley et al., 1990] [Light Source, Inc. 1995]. Given the CIELUV coordinates and an appropriate color calibration system, other researchers will be able to reproduce colors used these experiments.

[2] SuperCard is a registered trademark of Allegiant, Inc.

Table 4.2 CIELUV Coordinates for Colors Used

Color Name (& Use in Our Study)	L	U	V
Orange (Most Relevant)	74.20	78.33	98.81
Medium Green (Medium Relevance)	76.87	-24.54	23.35
Pale Blue (Least Relevance)	93.40	-16.40	-18.63
Pink (Books)	71.83	67.69	-32.98
Medium Blue (Journal Articles)	72.69	-21.62	-81.87
Lime Green (Proceedings Articles)	85.34	-15.60	97.84
Bisque (Control Space Background)	93.40	14.18	28.79

4.2.1.4 TASKS

In the first experiment, a *trial* consisted of a single information extraction (counting identification) task performed with a single display. For each design point, all 20 participants completed 15 counting identification tasks that required comprehension of the graphical encoding for completion. There were three training trials, followed by 12 and measured trials for each design point. A trial (task) began with presentation of one task description, presented in a small window at the bottom of the 20-inch monitor.

Each task required counting icons that represented documents matching specified criteria, where the necessary data were encoded only in the graphical devices being studied. For this experiment, the task description took the form of a question such as "How many icons represent documents that are journal articles?" where the journal article document type was encoded in icon shape.

We used counting identification tasks because the type of display scanning required for counting identification tasks, which require locating all icons that match given criteria, is very similar to the behavior of users of complex information visualization displays like Envision's Graphic View. Furthermore, counting identification tasks have greater potential to reveal minute differences in the time required for humans to process encoded information than locating a single icon, as might be required for a simple identification task.

Counting identification tasks also support study of interactions among graphical devices. In particular, filtering searches of multidimensional displays could be studied. As discussed in Chapter 2, these are searches in which more than one type of information (i.e., type and relevance) might be encoded in each single icon, but only part of that information (i.e., type or relevance) is sought. For filtering (or disjunction) tasks, the possibility exists that a visually dominant graphical

device such as color will cause only icons that include that device as part of the target condition to be selected. For example, if icon color represents document relevance and icon shape represents document type, participants who were asked how many icons representing documents of a specific type are shown might count only the most relevant documents of that type. With accurate responses requiring perception of all icons that match the task condition, we expected the likelihood of detecting such perceptual interference to be substantially increased.

Tasks for the experiment did not require use of either position encoding and the meaning of these codes remained constant, with the x-axis showing publication year (quantitative data) and the y-axis showing index terms (nominal data). 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, icon labels show document relevance rank during normal use. However, to avoid confounding because of redundancy with codes being tested, icon labels showed Envision document ID for all design points

For design points 15 through 20, a given design point presents multiple options for information extraction (e.g., drawing on a single graphical code out of several presented for filtering tasks, or some combination of the graphical codes for integration tasks). Accordingly, tasks were balanced among these options, enabling us to study interaction of graphical devices with one another. (See Table 4.1, where the rightmost three columns show distribution of tasks among the kinds of data represented.) For example, when both type and relevance were encoded in one display, three types of tasks were included in the study:

1) integration tasks, requiring use of both encodings, (e.g., "How many icons represent books of medium relevance?")
2) filtering tasks requiring use of document type only (e.g., "How many icons represent journal articles?"), and
3) filtering tasks requiring use of document relevance only (e.g., "How many icons represent highly relevant documents?").

For each non-redundant two-dimensional design point (design points 15-20 in Table 4.1), there were four integration tasks, and four filtering tasks each for document type and relevance. The total number of icons counted for each category of task (e.g., filtering for type, filtering for relevance, or integration) equaled 33±1. This balance was necessary for us to be able to compare results among the different categories of tasks, so that differences in time for task performance reflect perceptual differences, rather than differences in numbers of icons counted.

4.2.1.5 PROCEDURE

Participants were instructed to perform tasks as quickly as possible, without sacrificing accuracy. Participants were instructed to read the task description from the screen and then press the Return key. This initiated a trial, causing the relevant Graphic View screen dump to be displayed, the timer started, and the text insertion cursor moved to an answer field. The participant then performed the task by counting icons that matched the condition specified in the task description, typing a number in the answer field, and pressing the Return key, at which point the timer stopped, the screen dump window was closed, and the trial ended. The participant pressed the Return key again, activating a "Next Question" button, to signal readiness for the next trial. For each trial, time to task completion was recorded in a SuperCard database field, along with the response. On completion of all 15 tasks for a design point, we asked the participant to mark a form, rating that design point for ease of use and for the likelihood of selecting that visualization if they needed the information it encoded.

Dependent variables were accuracy and time for task completion. The order in which participants experienced design points was systematically varied and counterbalanced to reduce any learning transfer. However, within each design point, tasks were ordered the same way for all participants. Task order was controlled so that no two consecutive tasks required counting of the same target color, size, shape, or combination thereof.

Given correct answers, the total number of icons counted in all measured trials for each design point in the first experiment was 100±2. These were evenly distributed among document types and levels of relevance for unidimensional and redundant codes (design points 1-7 using type and design points 8-14 using relevance).

We were also careful that all three document types were represented in the displays and in tasks performed, so that the number of icons counted for each document type equaled 33±1 for the first 7 design points. Similarly, the number of icons counted for each level of relevance (i.e., most relevance, medium relevance, and least relevance) equaled 33±1 for design points 7-14. This practice parallels our use for design points 15-20 of 33±1 filtering tasks for type, 33±1 filtering tasks for relevance, and 33±1 integration tasks, as discussed above. Again, the balance ensures that difference in time to task completion are attributable to the code as a whole, instead of to differences in number of items counted or ease of counting some particular shape, color, or size.

4.2.2 Method: Experiment 2

4.2.2.1 PARTICIPANTS

For the discriminability experiment, participants were 10 graduate and undergraduate students at Virginia Tech, meeting the same conditions described above. Eight of them also completed the first experiment. No perceptual abnormalities were reported in this group. They were paid $6.00 for one half-hour session.

4.2.2.2 TASKS

For this discriminability experiment, a *trial* consisted of a single visual search task performed with a single display. For each trial, the participant was asked to count the icons in the display that matched a target icon, so that no information decoding was required—participants simply looked for icons that matched the target. The questions presented on the screen were basically of the same form used in the first experiment, but a target icon was shown rather than providing a text description of information to be decoded. For example, a trial might ask the participant to answer the question, "How many are _____ ?" where the underlined space in this example held a pink circle of medium size in the display on the screen.

For each design point, 10 participants completed 18 visual search tasks for completion. There were three training trials, followed by 15 measured trials for each design point. Measured trials were balanced so that the total number of icons counted was evenly distributed among the three elements of each code (e.g., the three colors or shapes or sizes), ensuring comparability among mean times to task completion. For each element, the number of icons counted was 33 ± 1.

4.2.2.3 PROCEDURE

Participants were instructed to perform tasks as quickly as possible without sacrificing accuracy. Dependent variables were accuracy and time for task completion. Again, the order in which participants experienced design points was systematically varied and counterbalanced to reduce any learning transfer. However, within each design point, tasks were ordered the same way for all participants. Task order was controlled so that no two consecutive tasks required counting of the same target color, size, or shape.

CHAPTER 5. RESULTS OF EMPIRICAL STUDY

As described in Chapter 4, dependent variables in both the experiments were error rate and time for task completion. Collection of objective data was automated using SuperCard. Timing was accurate to the nearest sixtieth of a second, beginning with presentation of the Graphic View screen dump and ending when participants entered their responses. Time to task completion and the participant response (numeric answer to the task question) were recorded in a SuperCard database for subsequent analysis. Subjective participant ratings for ease of use and likelihood of selecting the visualization for a given design point were collected on marked forms after completion of trials for each design point, constituting third and fourth dependent variables.

We analyzed data with three goals in mind:

1) to examine effectiveness in conveying both nominal and quantitative data by icon color, icon shape, and icon size; and if differences in effectiveness were statistically significant, to develop rankings of icon color, icon shape, and icon size based on their effectiveness in conveying both nominal and quantitative data;
2) to examine effectiveness of redundant graphical codes, with both double (two graphical devices) and triple (three graphical devices) redundancy, in comparison to unidimensional encodings (using a single graphical device) for nominal data and for quantitative data; and if differences in effectiveness were statistically significant, to develop rankings of redundant graphical codes for conveying both nominal and quantitative data; and
3) to examine interactions among icon color, icon shape, and icon size when used in non-redundant two-dimensional codes, both for integration and filtering tasks.

Rather than performing a specific test to address each goal, we used a series of analyses that addressed the three goals together, as described in the following sections.

5.1 ANALYSIS FOR MEAN TIME TO TASK COMPLETION

Time for task completion is a continuous variable. We plotted the data for time to task completion for each of several design points and observed that normality was tenable. For each participant at each design point, we constructed one value for mean time to task completion. For both unidimensional and redundantly encoded design points, this value was the mean time across the 12 measured trials. For non-redundant two-dimensional design points (15-20), we constructed four values for time and error rate — one overall value for all measured trials, one for the four filtering tasks dealing with type, one for the four filtering tasks dealing with relevance, and one for the four integration tasks. These values were constructed as above, except that for the filtering and integration groupings we averaged over four trials for time.

We performed separate analyses of variance for each of three groups of design points: points 1-7, encoding only document type in the icon design; points 8-14, encoding only relevance; and design points 15-20, encoding type in one graphic device and relevance in another, with the third uncoded. For each group of design points, 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 at the 1% level of significance. For each group of design points, 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.

5.1.1 Design Points 1-7, Conveying Document Type

Comparing results for mean time to task completion among the three unidimensional graphical codes (i.e., for nominal data: color representing type with shape and size both uncoded [C_T] versus shape representing type with color and size uncoded [S_T] versus size representing type with color and shape uncoded [Z_T], as shown in Table 4.1.), we expected to find significant differences among the means. Based on findings in the literature, we expected that color would yield the best mean time, followed by shape, with size last.

Evaluating the impact of redundant graphical codes (i.e., using two or more graphical devices to represent the same document attribute, with other graphical devices uncoded, in design points 4-7), we expected to find that redundant graphical codes yield faster mean times to task completion than those for unidimensional codes. Among two-dimensional redundant graphical codes (i.e., color and shape both representing type with size uncoded [C_TS_T] versus color and size both representing type with shape uncoded [C_TZ_T] versus shape and size both representing type with color uncoded [S_TZ_T]), we expected to find the most effective redundant code to be composed of the two graphical devices that yielded the best performance as unidimensional codes, yielding significantly lower mean time to task completion than the other two-dimensional redundant codes. Since we expected icon color and shape to form the two most effective unidimensional codes, we expected to find that the redundant code using both icon color and shape to convey type (C_TS_T, with size uncoded) would be significantly more effective than the other redundant codes. We also expected to find that the triply redundant code [$C_TS_TZ_T$], in which three graphical devices all represent the same document attribute, yielded the fastest time of all.

We rejected the null hypothesis that all 7 codes yield equal mean times to task completion ($F = 21.19$, $p = 0.0001$) and proceeded to the multiple range test, the results of which are shown in the figure below. In the figure, times and design points sharing a common underline are not significantly different. Times are in "ticks," or sixtieths of a second.

Design Point:	C_T	C_TZ_T	$C_TS_TZ_T$	C_TS_T	S_TZ_T	Z_T	S_T
Mean Time:	377	386	394	412	532	541	572

Figure 5.1 — Tukey's HSD Ordering:
Mean Times for Codes Conveying Document Type (Nominal Data)
Alpha = 0.01, Minimum Significant Difference = 92

With the risk of Type I error at 1% (alpha = 0.01), Tukey's HSD (Honestly Significant Difference) [Schulman 1992] 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 (C_T) produced the fastest mean time, followed by the redundant code with both color and size as type (C_TZ_T), then the triply redundant code for type ($C_TS_TZ_T$), and last the redundant code with both color and shape as type (C_TS_T), though the differences among these four are not statistically significant. The remaining three codes are ordered with the redundant code using both shape and size as type (S_TZ_T) first, followed by size as type (Z_T), with shape as type (S_T) last, though again the difference among these three is not significant.

The rankings below summarize results of the Tukey's HSD ordering. In the rankings, "<" represents statistically significant difference, while "≤" reflects non-significant difference.

Ranking of Unidimensional Codes Conveying Document Type (Nominal Data), Based on Time:

Color < Size ≤ Shape

Ranking of Redundant Codes Document Type (Nominal Data), Based on Time:

Color&Size ≤ Color&Shape&Size ≤ Color&Shape < Shape&Size

We expected to find significant performance benefit from redundant encoding, but that was not the case. Any code involving color produced a mean time approximately equal to that for color alone. To our surprise, the size code yielded slightly better mean time as a unidimensional code than did shape, an effect that was also apparent in results for redundant codes, for which the combination of color and size yielded faster performance than color and shape.

Our rankings for unidimensional codes parallel those of Christ [1975]. However, they differ from Mackinlay's (1986) rankings for nominal data which suggest the following order:

Color < Shape < Area

(Note that Mackinlay uses the term "area" to identify the same icon attribute that we call "size.") This suggests that the tasks our participants performed were closer to the counting identification tasks studied by Christ than to those studied by Mackinlay.

Previous studies, including those by Christ [1984], Jubis [1990], and Kopala [1979], found that redundant codes significantly reduced the time required for visual search tasks, especially when the redundant code uses both color and shape. Our finding that redundant codes involving color do not yield faster mean times than use of color alone is at odds with those findings. This may be due to differences in colors used in the studies.

5.1.2 Design Points 8-14, Conveying Document Relevance

Virtually identical analysis was performed with mean times to task completion for design points 8-14, conveying relevance, with analysis of variance again leading to the conclusion that the codes do not all require equal mean times to task completion ($F = 42.10$, $p = 0.0001$). Results of the Tukey's HSD test are shown below. As before, times and design points sharing a common underline are not significantly different. Times are in "ticks," or sixtieths of a second.

Design Point:	C_RZ_R	C_R	C_RS_R	$C_RS_RZ_R$	S_R	S_RZ_R	Z_R
Mean Time:	355	361	365	391	484	498	581

Figure 5.2 — Tukey's HSD Ordering:
Mean Times for Codes Conveying Document Relevance (Quantitative Data)
Alpha = 0.01, Minimum Significant Difference = 57

With the risk of Type I error at 1%, 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 contains only the unidimensional code using size, which produced faster results than shape among codes conveying document type, as shown in the previous Tukey table.
The ordering of these groups is slightly different from that for the type codes. In the first group, we find the fastest mean time from the redundant code with both color and size as relevance (C_RZ_R), followed by the code with color alone as relevance (C_R), then the redundant code with both color and shape as relevance (C_RS_R), with the triply redundant code fourth ($C_RS_RZ_R$), though differences among these mean times are not significant. In the second group, the code with shape (S_R) alone as relevance leads, followed by the redundant code with both shape and size as relevance (S_RZ_R), and the code with size alone as relevance (Z_R) is last.

The rankings below summarize results of the Tukey's HSD ordering. In the rankings, as before, "<" represents statistically significant difference, while "≤" reflects non-significant difference.

Ranking of Unidimensional Codes Conveying Document Relevance (Quantitative Data), Based on Time:

Color < Shape < Size

Ranking of Redundant Codes Document Relevance (Quantitative Data), Based on Time:

Color&Size ≤ Color&Shape ≤ Color&Shape&Size < Shape&Size

This ordering differs significantly from those suggested by literature. Mackinlay (1986) suggests that for quantitative data,

Area < Color

and that shape is unsuitable for encoding quantitative data. Mackinlay's rankings for these graphical devices parallel those of Cleveland and McGill (1985). Our results also differ from those predicted by Christ (1975), who ranks these devices as follows:

Color < Size < Shape

We believe our rankings differ from those of Mackinlay and from Cleveland and McGill because the perceptual tasks users performed in our experiments are closer to the counting identification tasks studied by Christ than to the graphical perception tasks studied by Cleveland and McGill, whose work influenced Mackinlay. We also believe that the shapes used in our code for relevance are more readily distinguished than those used in the earlier studies Christ reanalyzed, resulting in improved performance.

5.1.3 Design Points 15-20, Conveying Both Document Type and Document Relevance

Two-dimensional graphical codes may be either redundant (e.g., color and shape both representing document type) or non-redundant (e.g., color representing document type and size representing relevance). Redundant two-dimensional graphical codes have already been discussed in sections 5.1.1 and 5.1.2. Analysis of non-redundant graphical codes (all six possible combinations of one graphical device uncoded, with one graphical device representing document type and one representing document relevance) were again based on anovas and multiple range tests, supplemented in some instances by linear contrasts.

For the first analysis of these data, we used the mean time for completion of all 12 tasks involving each design point, combining the results of integration tasks with those for tasks filtering for relevance and those for tasks filtering for type. Once again, we rejected the null hypothesis that these six codes yield equal mean times to task completion ($F = 40.99$, $p = 0.0001$) and proceeded to the multiple range test.

Design Point:	C_TS_R	C_TZ_R	C_RS_T	C_RZ_T	S_TZ_R	S_RZ_T
Mean Time:	523	536	566	603	763	912

Figure 5.3 — Tukey's HSD Ordering: Mean Times for 2D Non-Redundant Codes for All 12 Trials
Alpha = 0.01, Minimum Significant Difference = 119

Tukey's HSD test showed once again that all four codes involving color yield significantly faster times than either of the achromatic codes. The order produced shows C_TS_R first, followed closely by C_TZ_R, with C_RS_T third, and C_RZ_T fourth. Differences among mean times for these four codes are not significant. Design point S_TZ_R stood alone behind the first group, requiring significantly more time for task completion, while design point S_RZ_T required significantly more time for task completion than any other design point among the two-dimensional nonredundant codes. The ordering of these last two points, S_TZ_R and S_RZ_T, is not what we might expect from earlier results, which indicated no significant difference between shape an size in conveying document type, but showed shape to be more effective than size in conveying document relevance.

5.1.3.1 FILTERING TASKS WITH NON-REDUNDANT 2D CODES

Non-redundant two-dimensional graphical codes were also compared with unidimensional codes to investigate possible interference in filtering tasks. For example, with design point C_TS_R, tasks may require only the type information encoded with color, perception of which may be affected by encoding document relevance in shape. Similarly, perception of the document relevance information encoded in shape may be affected by the use of color to convey document type. Our analysis required two anovas, one using mean times for all trials from design points 1-7 conveying only document type, along with mean times from the four trials filtering for document type with design points 15-20, and another anova using mean times for all trials from design points 8-14 conveying only document relevance, along with mean times from the four trials filtering for document relevance with design points 15-20. Details of these analyses follow.

For the analysis using mean times filtering for document type, we again rejected the null hypothesis ($F = 15.43$, $p = 0.0001$) that the 13 design points yield equal mean times and proceeded to multiple range tests. With alpha = 0.01, Tukey's HSD revealed a complex pattern of statistical significance.

Design Point with Mean Time below:

C_TZ_R	C_T	C_TZ_T	$C_TS_TZ_T$	C_TS_T	C_TS_R	S_TZ_T	Z_T	S_T	C_RS_T	C_RZ_T	S_TZ_R	S_RZ_T
347	377	386	394	413	455	532	541	572	594	648	698	968

Figure 5.4 — Tukey's HSD Ordering:
Mean Times for All Codes with Tasks Involving Document Type (Nominal Data)
Alpha = 0.01, Minimum Significant Difference = 162

All six codes using color to convey document type form a group yielding the fastest mean times, led by C_TZ_R and ending with C_TS_R, including C_T. Thus, we find no interference with perception

of color as type when either shape or size is used to encode relevance. The sixth grouping includes S_T, C_RS_T, and S_TZ_R, showing no interference in perception of shape as document type when either of color or size is used to encode document relevance. The sixth group also includes both Z_T and C_RZ_T, suggesting that perception of size as document type is not hindered by encoding document relevance in color. However, design point S_RZ_T stands alone in yielding the slowest mean time for tasks involving document type, suggesting that use of shape to encode document relevance interferes with perception of size as document type. With this single exception, we found no interference in perception of document type among color, size, and shape used in combination to convey both document type and document relevance.

The anova comparing the two-dimensional nonredundant codes, using only the tasks involving filtering for document relevance, with the seven unidimensional and multidimensional codes for document relevance, led us to reject the null hypothesis that all these design points yield equal mean times ($F = 15.18$, $p = 0.0001$). We then moved to the multiple range test. With alpha = 0.01, Tukey's HSD again revealed a complex pattern of significance, though with fewer groupings than for nominal data.

The first grouping includes C_R, C_RS_T, and C_RZ_T, showing no interference in conveying relevance with color when using either shape or size to encode type. The fourth grouping includes both S_R and C_TS_R, showing no hindrance in perceiving shape as document relevance when using color to encode document type. However, S_RZ_T is not in any grouping with S_R and yielded the slowest mean time, suggesting that use of size to encode document type interferes with perception of shape as document relevance. Z_R, C_TZ_R, and S_TZ_R are all in the fifth grouping, suggesting that encoding document type with color or shape does not hinder perception of size as document relevance.

Design Point with Mean Time Below:

C_RZ_R	C_R	C_RS_R	$C_RS_RZ_R$	C_RS_T	C_RZ_T	S_R	S_RZ_R	C_TS_R	Z_R	C_TZ_R	S_TZ_R	S_RZ_T
355	361	365	391	437	440	484	498	569	581	632	675	731

Figure 5.5 — Tukey HSD Ordering:
Mean Times for All Codes with Tasks Involving Document Relevance
Alpha = 0.01, Minimum Significant Difference = 123

To summarize, among the six non-redundant two-dimensional codes examined, each presenting two possible types of interference, we have found only two cases in which one code interferes

with perception of another encoding. Both cases involve the same code, S_RZ_T. Thus we see that use of size to convey document type interferes with perception of shape as document relevance, and, conversely, use of shape to encode document relevance interferes with perception of size as document type. We observed no such interference with the reverse encoding, S_TZ_R.

In our analysis of seven codes conveying type, described in section 5.1.1, we found that shape and size codes yield performance that is not significantly different from each other, though both are significantly slower than color. In conveying relevance, we found shape to be significantly faster than size. We see no pattern in the earlier results that suggests an explanation for the interference we found. However, participants voiced strong preference for codes involving color and also frequently commented on the difficulty of tasks for the last six design points, using non-redundant multidimensional codes. Participant frustration with these displays may explain why S_RZ_T and S_TZ_R produced the slowest performance, but does not explain the interference observed with the first of these encodings. The key may lie in results on user ratings for ease of use, discussed in Section 5.1.

Despite its apparent power to speed task completion, we did not find that color codes interfere with either shape or size encoding of either document type or document relevance. Our results differ from previous studies by Christ (1975, 1984), Luder and Barber (1984), and Umbers and Collier (1990), who found that color interferes with perception of all achromatic codes and with perception of shape and size in particular. Our different outcome may be due to the exact colors, shapes, and sizes used in our study, or to differences in tasks performed by participants in the studies.

5.2 Responses to Question 1: Ease of Use

As we did for mean time to task completion, we performed three separate one-way anovas using participant responses to each of the two subjective questions as the dependent variable. The first question was "How hard is it to use this visualization?" Responses ranged from 1 as Very Hard to 5 as Very Easy, so that higher values imply greater perceived ease of use. Subjective rating is already a single variable and cannot be distributed among types of tasks, so there is no further breakdown for non-redundant two-dimensional codes.

5.2.1 Responses to Question 1 for Codes Conveying Document Type

We again rejected the null hypothesis ($F = 26.84$, $p = 0.0001$) that participants found the seven codes equally easy to use and moved to the multiple range test.

Design Point:	$C_TS_TZ_T$	C_TS_T	C_T	C_TZ_T	S_TZ_T	S_T	Z_T
Mean Response:	4.70	4.65	4.55	4.55	3.65	3.35	2.30

Figure 5.6 — Tukey's HSD Ordering:
Responses to Question 1 for Codes Conveying Document Type (Nominal Data)
Alpha = 0.01, Minimum Significant Difference = 0.878

Once again the Tukey's HSD test shows the four design points involving color were easier to use than all achromatic codes. Design points S_TZ_T and S_T do not differ significantly from each other but are more difficult to use than the first group, while Z_T stands alone as the most difficult to use. This is in contrast to our earlier finding in section 5.1.1 that Z_T and S_T yield equal performance times, with the slight (but not significant) edge to Z_T. Thus, we obtain the following ranking, based on perceived difficulty of using the seven codes for document type:

Ranking of Unidimensional Codes Conveying Document Type,
Based on Ease of Use:

Color < Shape < Size

This ranking differs from our earlier finding in section 5.1.1, when time to task completion was the criterion:

Color < Size ≤ Shape

Participant opinion is that using a display in which size to encodes document type is significantly more difficult that using one in which shape encodes document type. This may partly explain why use of shape to encode document relevance interferes with the already-difficult use of size to encode document type, discussed in section 5.1.3.1 above.

We again found little significant benefit to redundant encoding, though the combination of shape and size did yield significantly greater ease of use than did size alone. Rankings for the redundant codes conveying document type, resulting from the Tukey HSD test, are shown below.

Ranking of Redundant Codes Conveying Document Type (Nominal Data)
Based on Ease of Use:

Color&Shape&Size ≤ Color&Shape ≤ Color&Size < Shape&Size

5.2.2 Responses to Question 1 for Codes Conveying Document Relevance (Quantitative Data)

As for the codes conveying document type, we rejected the null hypothesis ($F = 8.99$, $p = 0.0001$) that users found the seven codes conveying document relevance equally easy to use. We then again performed a multiple range test.

The four codes involving color were grouped as equally easy to use, as they were for type data. The unidimensional code with shape as document relevance is grouped with the redundant code using both shape and size, a pairing also seen in the codes conveying document type and paralleling the grouping by time, though the Question 1 relevance grouping also includes the triply redundant $C_RS_RZ_R$. Again, the unidimensional code with size as relevance is rated hardest to use, just as it held the last-place position by time to task completion. Rankings for ease of use, shown below, match those for relevance codes ranked by time to task completion.

Design Point:	C_R	C_RS_R	C_RZ_R	$C_RS_RZ_R$	S_R	S_RZ_R	Z_R
Mean Response:	4.70	4.70	4.65	4.50	3.95	3.90	2.95

Figure 5.7 — Tukey's HSD Ordering:
Responses to Question 1 for Codes Conveying Document Relevance (Quantitative Data)
Alpha = 0.01, Minimum Significant Difference = 0.6

Ranking of Unidimensional Codes Conveying Document Relevance (Quantitative Data) Based on Ease of Use:

Color < Shape < Size

Unfortunately, these results shed no light on the observed interference of perception of relevance encoded in shape by using size to encode type, discussed in sections 5.1.3.1 and 5.2.1.

Again we found little significant benefit to redundant encoding, though again the combination of shape and size to convey document relevance yielded significantly greater ease of use than did size alone. Rankings for the redundant codes conveying document type, resulting from the Tukey HSD test, are shown below.

Ranking of Redundant Codes Conveying Document Relevance (Quantitative Data),
Based on Ease of Use:

Color&Shape ≤ Color&Size ≤ Color&Shape&Size
with
Color&Shape&Size ≤ Shape&Size

5.2.3 Responses to Question 1 for Codes Conveying Both Document Type & Document Relevance

We once again rejected the null hypothesis ($F = 22.16$, $p = 0.0001$) that users found the codes equally easy to use. As in all previous analyses, results showed codes involving color grouped by the Tukey's HSD test on user responses as easier to use than the two achromatic codes.

Design Point:	C_TS_R	C_RS_T	C_TZ_R	C_RZ_T	S_TZ_R	S_RZ_T
Mean Response:	3.85	3.65	3.20	2.85	2.15	1.30

Figure 5.8 — Tukey's HSD Ordering:
Responses to Question 1 for 2D Non-Redundant Codes Conveying Document Type and Document Relevance
Alpha = 0.01, Minimum Significant Difference = 1.032

5.3 RESPONSES TO QUESTION 2: LIKELIHOOD OF USE

The second question we asked participants was, "If you needed the information visualized, how likely would you be to choose this visualization?" Responses again ranged from 1 to 5, with 1 as Very Unlikely and 5 as Very Likely.

5.3.1 Responses to Question 2 for Codes Conveying Document Type

We rejected the null hypothesis ($F = 30.46$, $p = 0.0001$) that users rate use the seven codes equally likely and moved to the multiple range test.

Design Point:	C_TZ_T	C_TS_T	$C_TS_TZ_T$	C_T	S_TZ_T	S_T	Z_T
Mean Response:	4.60	4.55	4.50	4.25	3.25	2.80	1.75

Figure 5.9 — Tukey's HSD Ordering:
Responses to Question 2 for Codes Conveying Document Type (Nominal Data)
Alpha = 0.01, Minimum Significant Difference = 0.878

As for other tests, we see here a clear benefit to codes using color. However, we see here for the first time some indication that redundant codes are more effective than unidimensional codes, observing that the three redundant codes involving color [C_TZ_T, C_TS_T, and $C_TS_TZ_T$] all have similar means that are higher, though not significantly so, than the mean for color alone [C_T]. We

also see that the unidimensional color code is considered only equally likely to be used with the redundant code using shape and size [S_TZ_T].

5.3.2 Responses to Question 2 for Codes Conveying Document Relevance (Quantitative Data)

Once again, we rejected the null hypothesis ($F = 27.70$, $p = 0.0001$) and moved to multiple range tests, the results of which are shown below.

Design Point:	C_RS_R	C_R	$C_RS_RZ_R$	C_RZ_R	S_R	S_RZ_R	Z_R
Mean Response:	4.65	4.45	4.35	4.35	3.70	3.50	2.35

Figure 5.10 — Tukey's HSD Ordering: Responses to Question 2 for Relevance Codes
Alpha = 0.01, Minimum Significant Difference = 0.765

With the risk of Type I error at 1%, Tukey's HSD showed that all four design points involving color are equally likely to be used, and more likely so than any achromatic code, while Z_R was less likely to be used than any other encoding. Unlike the results for question 2 regarding codes for document type, we found no significant increase in likelihood of using a visualization because of redundant encoding.

5.3.3 Responses to Question 2 for Codes Conveying Both Document Type and Document Relevance

We rejected the null hypothesis ($F = 25.88$, $p = 0.0001$) that the six non-redundant two-dimensional codes are equally likely to be used and moved to the multiple range test.

Design Point:	C_TS_R	C_RS_T	C_TZ_R	C_RZ_T	S_TZ_R	S_RZ_T
Mean Response:	3.70	3.65	2.95	2.55	1.95	1.05

Figure 5.11 — Tukey's HSD Ordering:
Responses to Question 2 for 2D Non-Redundant Codes
Alpha = 0.01, Minimum Significant Difference = 0.988

Tukey's HSD showed all design points using color more likely to be used than achromatic codes, though the combination of color as document relevance with size as document type [$C_R Z_T$] is significantly less likely to be used than either code involving both color and shape [$C_T S_R$ and $C_R S_T$] and the code with shape as document relevance and size as document type [$S_R Z_T$] is least likely to be used, though not significantly less than the other achromatic code [$S_T Z_R$].

5.4 TEST OF CORRELATION BETWEEN QUESTIONS 1 AND 2

Observing similarities between participant response to the two questions, we performed a rank correlation test. As expected, the correlation is strong (rho = 0.86917, p = 0.0001), suggesting that participants are more likely to choose visualizations they find easiest to use.

5.5 ANALYSIS FOR PARTICIPANT ERRORS

For both unidimensional and redundantly encoded design points, error rate was the proportion of trials in which the participant gave an incorrect answer in 12 trials for one design point (i.e., for the first experiment, the possible error rates are 0/12, 1/12,...,12/12). For non-redundant two-dimensional design points, we constructed four values for error rate, as we did for time, while error rate was proportional to only four trials for each filtering and integration grouping. Error rate was not continuous nor was it normally distributed, because many participants made no errors on a majority of design points, so that analysis of variance was not possible.

We then constructed Table 5.1, summarizing the number of trials for each design point (out of 240) in which an error was made. As we did for measurements of mean time to task completion and participant responses to our questions, we analyzed these measurements of error frequency in three groups: one group for all seven design points representing only document type, one group for all seven design points representing only document relevance, and one group for the six design points using non-redundant two-dimensional codes to represent both document type and 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.

Table 5.1 Trials per Design Point in Which an Error was Made

Design Point		Trials w/Errors
1	C_T	12
2	S_T	14
3	Z_T	27
4	$C_T S_T$	5
5	$C_T Z_T$	6
6	$S_T Z_T$	8
7	$C_T S_T Z_T$	11
8	C_R	7
9	S_R	23
10	Z_R	24
11	$C_R S_R$	3
12	$C_R Z_R$	3
13	$S_R Z_R$	15
14	$C_R S_R Z_R$	4
15	$C_T S_R$	18
16	$C_R S_T$	24
17	$C_T Z_R$	57
18	$C_R S_T$	27
19	$S_T Z_R$	46
20	$S_R Z_T$	68

Design Point:	C_TS_T	C_TZ_T	S_TZ_T	$C_TS_TZ_T$	C_T	S_T	Z_T
Trials with Errors:	5	6	8	11	12	14	27

Figure 5.12 — Fisher's Exact Test of Proportions of Trials with Errors for Codes Conveying Document Type (Nominal Data)
Alpha (per comparison) = 0.01

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

Color ≤ Shape ≤ Size

The ranking by error frequency of two- and three-dimensional redundant graphical codes conveying type is

Color&Shape ≤ Color&Size ≤ Shape&Size ≤ Color&Shape&Size

Note that all four redundant codes yield fewer errors than the three unidimensional codes. While the difference between $C_TS_TZ_T$ and Z_T is not quite significant, it is nearly so($p = 0.011$). A further test comparing the proportion of errors for all redundant codes conveying document type to those for all non-redundant codes conveying document type revealed a significant difference favoring redundant codes, with the two-tailed FET = 0.0000929.

As for the codes conveying document type, for the seven codes conveying only document relevance, the Chi Square test again showed a significant difference in error frequency among the codes ($p = 0.001$). We then used Fisher's Exact Test to establish the groupings shown below.

Design Point:	C_RS_R	C_RZ_R	$C_RS_RZ_R$	C_R	S_RZ_R	S_R	Z_R
Trials with Errors:	3	3	4	7	15	23	24

Figure 5.13 — Fisher's Exact Test of Proportions of Trials with Errors for Codes Conveying Document Relevance (Quantitative Data)
Alpha (per comparison) = 0.01

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 again 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 ($p = 0.00000369$).

Examining the six non-redundant two-dimensional codes conveying both type and relevance, we found significant differences among the proportions of errors ($p = 0.001$). We then used Fisher's Exact Test to establish the groupings shown in Figure 5.14. While the difference between C_TZ_R and C_RZ_T is not significant, it is nearly so ($p = 0.018$).

Design Point:	C_TS_R	C_RS_T	C_TZ_R	C_RZ_T	S_TZ_R	S_RZ_T
Trials with Errors:	18	24	27	46	57	68

Figure 5.14 — Fisher's Exact Test of Proportions of Trials with Errors for Codes Conveying Both Document Type and Document Relevance
Alpha (per comparison) = 0.01

5.6 ANALYSIS FOR EXPERIMENT 2 — DISCRIMINABILITY TEST

The second experiment was a test of the discriminability of icons used in the first experiment. The dependent variable was mean time for task completion. There were two goals in data analysis:

- to document the relative discriminability of the five sets of graphic devices (i.e., two sets of three colors, one set for document type and one set for document relevance; two sets of three shapes, one set for document type and one set for document relevance; and one set of sizes, used for both document type and document relevance) used in the first experiment.
- to document the relative discriminability of the elements (i.e., individual colors, shapes, and sizes) within each of the five sets.

As for the first experiment, our analysis was based on analysis of variance, supplemented with multiple range tests. Also as we did for the first experiment, we used the more conservative Tukey's HSD test with risk of Type I error at 1%, rather than making Huynh-Feldt corrections to degrees of freedom.

To support analysis for the first goal, we constructed a mean from the times required for completion of all 15 tasks for each subject for each set of the five sets of graphic devices. Analysis of variance showed that the five sets of graphic devices were not equally discriminable ($F = 20.65$, $p = 0.0001$). The results of Tukey's HSD test are shown below. Mean times are in ticks, or sixtieths of a second.

Type colors and relevance colors are not significantly different, though type colors were more quickly distinguished than any other set of devices. Relevance colors and relevance shapes were also equally distinguished. However, relevance colors were more quickly distinguished than either type shapes or sizes. The mean times for relevance shapes, type shapes, and sizes are not significantly different.

Type Colors C_T	Relevance Colors C_R	Relevance Shapes S_R	Type Shapes S_T	Sizes $Z_{R\|T}$
297.7	362.6	443.3	514.3	519.0

Figure 5.15 — Tukey's HSD Ordering: Mean Time for Five Sets of Graphical Devices
Alpha = 0.01, Minimum Significant Difference = 107.17

These results support our earlier finding that the two color codes used in these experiments were easier to use than any other codes. The first experiment produced different rankings for the relative effectiveness of shape and size: for encoding document type, size produced slightly (though not significantly) faster mean times than did shape, but for encoding document relevance, size was significantly slower. In particular, we note that the two sets of colors were equally discriminable, as were the two sets of shapes. The same sizes were used for both type and relevance and were as discriminable as the shapes, though less so than either set of colors.

Next, we used a multiple range test to examine the discriminability of, or time required to distinguish among, elements within each set of graphic devices. For all tests, the risk of Type I error (alpha) was set at 1%.

For type colors, Tukey's HSD test showed the 3 colors to require equal times to distinguish:

Blue	Lime	Pink
292.8	299.9	300.5

Figure 5.16 — Tukey's HSD Ordering: Mean Time for Type Colors
Alpha = 0.01, Minimum Significant Difference = 131.7

For relevance colors, the results of Tukey's HSD test also showed no difference in discriminability:

Orange	Pale Blue	Green
322.9	333.7	431.3

Figure 5.17 — Tukey's HSD Ordering: Mean Time for Relevance Colors
Alpha = 0.01, Minimum Significant Difference = 131.7

Some difference in discriminability of relevance shapes was revealed by Tukey's HSD test. In particular, finding diamonds required more time than stars.

Star	Triangle	Diamond
369.5	442.0	518.4

Figure 5.18 — Tukey's HSD Ordering: Mean Time for Relevance Shapes
Alpha = 0.01, Minimum Significant Difference = 131.7

Results for type shapes, however, showed that these shapes required equal time to distinguish:

Book	Proceedings Article	Journal Article
476	517	550

Figure 5.19 — Tukey's HSD Ordering: Mean Time for Type Shapes
Alpha = 0.01, Minimum Significant Difference = 131.7

Finally, we found that the three sizes did not require equal time to distinguish, with the large circles recognized significantly faster than the medium circles. However, the small circles required time equal to both the largest and the medium size.

Large	Small	Medium
424.5	442.0	518.42

Figure 5.20 — Tukey's HSD Ordering: Mean Time for Sizes
Alpha = 0.01, Minimum Significant Difference = 131.7

CHAPTER 6. DISCUSSION OF EMPIRICAL RESULTS

6.1 COMPARISON OF RANKINGS

6.1.1 Rankings for Codes Conveying Nominal Data

We have produced five rankings for codes conveying document type (nominal data) in the course of this analysis, ranking by:

- mean time to task completion,
- frequency of errors during trials,
- responses to the question on ease of use (Q1),
- responses to the question on likelihood of use (Q2), and
- discriminability of the sets of graphical devices used in the codes.

These rankings, along with those of Mackinlay [1986] and Christ [1984], are shown in Table 6.1. Columns to the left of the vertical bar show rankings obtained from these studies, while 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. In the columns for Table 6.1, Q1 pertained to ease of use and Q2 to likelihood of use. The rightmost column pertains to accuracy on identification tasks.

Table 6.1 — Rankings of Codes Conveying Nominal Data

	Time	Errors	Ease of use Q1	Likelihood of use Q2	Discriminability	Mackinlay	Christ Search Time	Christ Ident. Accuracy
Color	1	1	1	1	1	1	1	1
Shape	3	2	2	2	2	2	3	3
Size	2	3	3	3	3	3	2	2

We observe that our rankings by error frequency, discriminability, and responses to questions about ease of use and likelihood of use are the same and correspond exactly to those of Mackinlay. Kendall's Tau, measuring correlation among these rankings, equals 1. For rankings of three items (n = 3), the only possible values for Kendall's Tau are -1, -1/3, 1/3, and 1. Tau = 1 signifies perfect correlation. We note that with n = 3 (that is, rankings of only 3 items), testing Kendall's Tau is not reliable because this test of correlation was designed for comparing rankings of larger sets of items, but Kendall's Tau gives us some objective measure of similarities and differences among these small rankings.

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, yielding Kendall's Tau = 1 for these rankings, but Kendall's Tau = 1/3 when comparing our ranking by time (and Christ's) with our

rankings by error frequency, discriminability, and responses to Q1 and Q2 (and Mackinlay's), indicating only partial agreement. 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 ease of use (Q1), likelihood of use (Q2), error frequency, and discriminability, 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.

The difference among our own rankings, as well as between those of Mackinlay and Christ, is not surprising, given the variety of measures of effectiveness used. We note, for example, the finding that redundant codes conveying document type do not consistently yield faster performance (see Section 5.1.1), but they do yield significantly fewer errors (see Section 5.5). Thus, designers should determine the relative importance of these measures of effectiveness in choosing among rankings to guide design decisions.

6.1.2 Rankings for Codes Conveying Quantitative Data

We have produced the same five rankings for codes conveying quantitative data that we produced for nominal data. These rankings, along with those of Mackinlay [1986], Christ [1984], and Cleveland and McGill [1984] [1985] are shown in Table 6.2. As for Table 6.1, Q1 pertains to user ratings for ease of use and Q2 to user ratings for likelihood of using each design point. The vertical bar separates our results from those of other studies. 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 6.2 — Rankings of Codes Conveying Quantitative Data

	Time	Errors	Ease of Use Q1	Likelihood of Use Q2	Discriminability	Christ Search Time	Mackinlay	Cleveland & McGill C&M
Color	1	1	1	1	1	1	2	2
Shape	2	2	2	2	2	3	3	3
Size	3	3	3	3	3	2	1	1

We observe that our rankings by time to task completion, error frequency, responses to Q1 and Q2, and discriminability are the same. Comparing our rankings, we find Kendall's Tau = 1, signifying perfect correlation. 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.

Kendall's Tau measurement of agreement between our rankings and those of Mackinlay and Cleveland and McGill is -1/3, indicating more disagreement than agreement.

All of 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. Comparing our rankings with Christ's, Kendall's Tau equals 1/3, indicating incomplete agreement between the rankings.

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, the disagreement reflects a fundamental difference in the nature of the tasks on which the rankings are based. Recall our comments in Section 2.4, regarding our reservations about the assumptions and tasks on which Cleveland and McGill based their rankings, which influenced Mackinlay. The *graphical perception* task used in their experiments required making a determination about a specific quantitative value or comparing exactly two graphical items in a display which did not contain extraneous data. Our experiments, on the other hand, required users to search a display and count only those items which 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, since the tasks used in our experiments are more similar to the visual search and counting identification tasks he described (see Section 2.1.1.1). 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 codes and those of earlier studies, including those re-analyzed by Christ. We have found only one such study, that of Smith and Thomas [1964], which reports discriminability data.

6.2 RECOMMENDATIONS TO DESIGNERS

Mackinlay's [1986] rankings suggest that designers' choices of graphical devices should vary, depending on the data represented (see Section 2.3). Cleveland and McGill [1984] [1985], on the other hand, are concerned only with graphical perception, which is, by definition, limited to quantitative data. Christ's [1984] rankings ignore the type of data represented, focusing instead of the kinds of tasks performed.

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*,

since variations in these choices lead to different conclusions about which rankings are best. 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 [1984], whose research is from the vantage point of psychophysics, while our rankings differ more with those of Mackinlay [1986] and Cleveland and McGill [1984] [1985].

The difference between our rankings and those of both Mackinlay and Cleveland and McGill do not necessarily imply that either is wrong. Where *counting identification tasks* are 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, as we have seen with icon shape and size when conveying nominal data (see Table 6.1). 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 because of concern about color-impaired users. However, the clear power of color codes to improve both accuracy and speed of performance lead 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 gray-scale printer, or color printouts are subject to photocopying. We note that redundant encoding offers the additional advantage of improving accuracy (see Section 5.5).

Because decisions about every detail of graphical code design impact user performance, from the exact size of icons to tiny 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.

CHAPTER 7. SUMMARY AND FUTURE WORK

Providing access to graphically encoded information requires attention to a range of human cognitive and perceptual 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 codes which might otherwise be reduced to opinion and personal taste. However, literature (see Chapter 2, Related Research) offers inconclusive and often conflicting viewpoints, suggesting that further research is needed.

These studies provide 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 by all measures except time for task completion, which places shape behind size. For quantitative data, we found, by all measures, that encodings with icon shape are more effective than with icon size. 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.

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.

Further studies of this type are needed to empirically compare effectiveness of other graphical devices suggested by authors such as Christ [1984], Cleveland and McGill [1984] [1985], and Mackinlay [1986]. Other graphical devices which might be studied include, among others: texture, flash rate, letters and digits, and orientation or angle. 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. Our 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. However, results of this study for the six non-redundant two-dimensional design points suggest that users find use of these codes difficult and undesirable, and that they pay a significant price in both time to task completion and increased errors when using these codes.

Our study did not address issues pertaining to partially redundant three-dimensional graphical codes, those in which two graphical device both represent one data type while a third graphical device represents the other data type. As for non-redundant two-dimensional graphical codes, tasks for partially redundant three-dimensional codes may be integration or filtering tasks. These design points are left for future study. Results of integration tasks might be compared with those of tasks

involving the same graphical devices with the same semantics where only one graphical encoding is present (e.g., results from displaying shape as type and size as relevance, with color uncoded, would be compared to results from displaying shape as type, with both color and size uncoded, and to those from displaying size as relevance, with both shape and color uncoded). Results of filtering tasks where both data types are encoded might be compared with use of the same graphical devices to represent the same data types where only a single graphical encoding is used. That is, the same comparisons would be made as for the conjunction search, with different results expected. While worthwhile, this study would not be expected to yield rankings by effectiveness. Even if such rankings were produced, their value is unclear since design decisions are likely to be driven by less complex user tasks. The primary insights gained would be closer to the realm of psychophysics than to user interface design.

In summary, 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.

REFERENCES

Ahlberg, Christopher and Shneiderman, Ben. (1994) Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, in *Proceedings of CHI '94 Human Factors in Computing Systems,* Boston, MA, April 24-28, 1994, Ed. Adelson, Beth; Dumais, Susan, and Olson, Judith. ACM Press/Addison Wesley. pp. 313-317, 479-480.

Ashcraft, Mark A. (1989) *Human Memory and Cognition*. HarperCollins Publishers.

Bates, M. (1984) The fallacy of the perfect thirty-item on-line search. *RQ*, 42(1), 43-50.

Beale, Russell; McNab, Rodger J.; and Witten, Ian H. (1997) Visualizing sequences of queries: A tool for information retrieval. In *Proceedings 1997 IEEE Conference on Information Visualization: An International Conference on Computer Visualization and Graphics*, August 27029, 1997, London, England, 57-62.

Bishop, A. P. and Crook, M. N. (1960) Absolute identification of colour for targets presented against white and coloured backgrounds. Wright Air Development Division Report No WADD TR 60-611.

Brown, Judith R.; Earnshaw, Rae; Jern, Mikael; and Vince, John. (1995) *Visualization: Using Computer Graphics to Explore Data and Present Information*. New York: John Wiley and Sons, Inc.

Bundesen, C. and Pedersen, L. F. (1983) Color segregation and visual search. *Perception and Psychophysics*, 33(5), 487-493.

Cahill, Mary-Carol, and Carter, Robert. C., Jr. (1976) Color code size for searching displays of different density. *Human Factors*, 18(3), 273-280.

Carswell, C. Melody. (1992a) Choosing specifiers: an evaluation of the basic tasks model of graphical perception. *Human Factors*, 34(5), 535-554.

Carswell, C. Melody. (1992b) Reading graphs: Interactions of processing requirements and stimulus structure. In Burns, Barbara (ed.), *Percepts, Concepts and Categories: The Representation and Processing of Information*. New York: North-Holland. 605-645.

Carswell, C. Melody and Wickens, Christopher D. (1987) Information integration and the object display: An interaction of task demands and display superiority. *Ergonomics*, 30(3), 511-527.

Carswell, C. Melody and Wickens, Christopher D. (1990) The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object integration. *Perception and Psychophysics*, 47, 157-168.

Carswell, C. Melody and Wickens, Christopher D. (1996) Mixing and matching lower-level codes for object displays: evidence for two sources of proximity compatibility. *Human Factors* , 38(1), 1-22.

Carter, R. C. (1982) Visual search with color. *Journal of Experimental Psychology: Human Perception and Performance*, 8(1), 127-136.

Cavanagh, J. P. (1972) Relationship between the immediate memory span and the memory search rate. *Psychological Review,* 79, 525-530.

Cavanagh, Patrick; Arguin, Martin; and Treisman, Anne. (1990) Effect of surface medium on visual search for orientation and size features. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3), 479-491.

Chalmers, Matthew and Chitson, Paul. (1992) Bead: Explorations in Information Visualization, in (ed.) Belkin, Nicholas; Ingwersen, Peter; and Pejtersen, Annelise Mark, *Proceedings of SIGIR '92, The Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 330-337.

Chapanis, A. and Halsey, R. (1956) Absolute judgments of spectral colors. *Journal of Psychology*, 42, 99-103.

Christ, Richard E. (1975) Review and analysis of color coding research for visual displays. *Human Factors*, 17(6), 542-570.

Christ, Richard E. (1984) Research for evaluating visual display codes: an emphasis on colour coding. In R. Easterby & H. Zwaga (Eds.), *Information Design: The design and evaluation of signs and printed material* (pp. 209-228). New York, NY: John Wiley and Sons Ltd.

Christ, Richard E. and Corso, Gregory M. (1983) The effects of extended practice on the evaluation of visual display codes. *Human Factors*, 25(1), 71-84.

Chuah, Mei C. and Roth, Steven, F. (1996) On the semantics of interactive visualization. In *Proceedings IEEE Symposium on Information Visualization '96,* October 28-29, 1996, San Francisco, California. Ed. Card, Stuart; Eick, Stephen G.; and Gershon, Nahum. 29-36.

Cleveland, W. S. and McGill, R. (1984) Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 531-554.

Cleveland, W. S. and McGill, R. (1985) Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(August), 828-833.

DeSchepper, Bruce and Treisman, Anne. (1996) Visual memory for novel shapes: Implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(1), 27-47.*

Dubin, David. (1995a) Comparing and Assessing VIRIs. Working document included in the Program for the VIRI Workshop, SIGIR 95.

Dubin, David. (1995b) Document analysis for visualization. In Fox, Edward A.; Ingwersen, Peter; and Fidel, Raya (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval — SIGIR '95*, Seattle, Washington, July 9-13, 1995, 199-204.

D'Zmura, Michael. (1991) Color in visual search. *Vision Research*, 31(6), 951-966.

D'Zmura, Michael; Lennie, Peter; and Tiana, Carlo. (1997) Color search and visual field segregation. *Perception and Psychophysics*, 59(3), 381-388.

Fairchild, Kim M.; Poltrock, Steven E.; and Furnas, George W. (1988) SemNet: Three-Dimensional Graphic Representations of Large Knowledge Bases. In (ed.) Guindon, Raymonde,*Cognitive Science and its Application for Human-Computer Interaction.* Hillsdale, New Jersey: Lawrence Erlbaum and Associates. 201-233.

Foley, James D.; van Dam, Andries; Feiner, Steven K.; and Hughes, John F. (1990) *Computer Graphics: Principles and Practice, Second Edition*. New York: Addison-Wesley.

Fox, E.A.; France, R. K.; Sahle, E.; Daoud, A.; and Cline, B. E. (1993) Development of a modern OPAC: from REVTOLC to MARIAN. In Korfhage, R.; Rasmussen, E.; and Willet, P. (Eds.) *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, PA, USA, June 27-July 1, 1993, 248-259.

Fox, Edward A.; Hix, Deborah; Nowell, Lucy T.; Brueni, Dennis J.; Wake, William C.; Heath, Lenwood S.; and Rao, Durgesh. (1993) Users, user interfaces, and objects: Envision, a digital library, in *Journal of the American Society for Information Science* 44(5), September 1993, 44(5), 480-491.

Frakes, William B. and Baeza-Yates, Ricardo. (Ed.) (1992) *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, New Jersey: Prentice Hall.

Gershon, Nahum and Eick, Stephen G. (1997) Information visualization. *IEEE Computer Graphics and Applications*, July/August, 29-31.

Goettle, Barry P.; Wickens, Christopher D.; and Kramer, Arthur F. (1991) Integrated displays and the perception of graphical data. *Ergonomics*, 34(8), 1047-1063.

Harmon, Donna. (1992) Relevance feedback revisited. In Belkin, N.; Ingerwersen, P.; and Pejtersen, A.M. (Eds.), *Proceedings of the SIGIR '92, the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* June 21-24, 1992, Copenhagen, 1-10.

Healey, Christopher G. (1996) Choosing Effective Colours for Data Visualization. In *Proceedings Visualization '96*. Ed. Yagel, Roni and Nielson, Gregory M. October 27-November 1, 1996. San Francisco, California. 263-270.

Heath, L.; Hix, D.; Nowell, L.; Wake, W.; Averboch, G.; Guyer, S.; and Fox, E. (1995) "Envision: A User-Centered Database of Computer Science Literature." In *Communications of the ACM*, 38(4), 52-53.

Horton, William. (1994) *The Icon Book: Visual Symbols for Computer Systems and Documentation*. New York: John Wiley and Sons.

Humphreys, Glyn W. and Boucart, Muriel. (1997) Selection by color and form in vision. *Journal of Experimental Psychology: Human Perception and Performance,* 23(1), 136-153.

Jones, Patricia M.; Wickens, Christopher D.; and Deutsch, Stuart J. (1990) The display of multivariate information: an experimental study of an information integration task. *Human Performance*, 3(1), 1-17.

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.

Keppel, Geoffrey. (1991) *Design and Analysis: A Researcher's Handbook, Third Edition*. Engelwood Cliffs, New Jersey: Prentice Hall.

Kim, Min-Shik and Cave, Kyle R. Spatial attention in visual search for features and feature conjunctions. *Psychological Science*, 6(6), 376-380/

Kopala, C. J. (1979) The use of color-coded symbols in a highly dense situation display. In *Proceedings of the Human Factors Society - 23rd Annual Meeting*, 397-401.

Kosslyn, S. M. (1985) Graphics and human information processing: A review of five books. *Journal of the American Statistical Association*, 80(391), 499-512.

Lavie, Nilli. (1997) Visual feature integration and focused attention: Response competition from multiple distractor features. *Perception and Psychophysics*, 59(4), 543-556.

Light Source, Inc. (1995) *Colortron Color System: Color Control from Original...to Monitor... to Print*.

Lohse, Gerald L.; Biolsi, Kevin; Walker, Neff; and Rueter, Henry H. (1994) A classification of visual representations. *Communications of the ACM*, 37(12), 36-49.

Lohse, Gerald; Walker, Neff; Biolsi, Kevin; and Rueter, Henry. (1991) Classifying graphical information. *Behaviour and Information Technology*, 10(5) 419-436.

Luder, C. B. and Barber, P. J. (1984) Redundant color coding on airborne CRT displays. *Human Factors*, 26(1), 19-32.

Lubow, R. E. and Kaplan, Oren. (1997) Visual search as a function of type of prior experience with target and distractor. *Journal of Experimental Psychology: Human Perception and Performance*, 23(1), 14-24.

Mackinlay, Jock D. (1986a) *Automatic Design of Graphical Presentation*. Doctoral dissertation, Computer Science Department, Stanford University, Stanford, California, December 1986. (Figure from p. 68) Also Tech. Report Stan-CS 86 1038.

Mackinlay, Jock. (1986b) Automating the design of graphical presentations of relational information. *Transactions on Graphics*, 5(2), 110-141.

Macdonald, W. A. and Cole, B. L. (1988) Evaluating the role of colour in a flight information cockpit display. *Ergonomics*, 31(1), 13-37.

Merwin, D. H. and Wickens, C. D. (1993) Comparison of eight color and gray scales for displaying continuous data. In *Proceedings of the Human Factors Society,* 1993, v. 2, 1330-1334.

Miller, G. A. (1956) The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 63, 81-97.

Moraglia, G., Maloney, K. P., Fekete, E. M., & Al-Basi, K. (1989). Visual search along the colour dimension. *Canadian Journal of Psychology*, 43(1), 1-12.

Mordkoff, J. Toby; Yantis, Steven; and Egeth, Howard E.(1990) Detecting conjunctions of color and form in parallel. *Perception and Psychophysics*, 48(2), 157-168.

Munsell, Albert H. (1969) *A Grammar of Color*. Edited by Faber Birren. New York: Van Nostrand Reinhold.

Nagy, A. L. and Sanchez, R. R. (1990) Critical color differences determined with a visual search task. *Journal of the Optical Society of America*, 7(7), 1209-1217.

Nowell, Lucy T.; France, Robert K.; Hix, Deborah; Heath, Lenwood S.; and Fox, Edward A. (1996) Visualizing search results: some alternatives to query-document similarity. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Inforamtion Retrieval*, Zurich, Switzerland, August 1996. 67-75.

Nowell, Lucy T. and Hix, Deborah. (1993) Visualizing Search Results: User Interface Development for the Project Envision Database of Computer Science Literature. In *Human-Computer Interaction: Software and Hardware Interfaces, vol. 19B of Advances in Human Factors/ Ergonomics, Proceedings of HCI International '93, 5th International Conference on Human Computer Interaction jointly with the 9th Symposium on Human Interface (Japan)*, Orlando, Florida, August 8-13, 1993. Pages 56-61. Elsevier, New York, 1993.

Nowell, Lucy Terry and Hix, Deborah. (1993) Query composition: Why does it have to be so hard? In *Proceedings of the East-West International Conference on Human-Computer Interaction, 1993, Vol. I*. Moscow, Russia. August 1993. 226-241. Also available as tech report TR93-19 from Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061

Olsen, K. A.; Korfhage, R. R.; Sochats, K. M.; Spring, M. B.; and Williams, J. (1993) Visualization of a document collection: The VIBE system. *Information Processing and Management 29*, 1(1993), 69-81.

Ott, Lyman. (1988) *An Introduction to Statistical Analysis and Data Analysis, Third Edition*. Boston: PWS-Kent Publishing Company.

Payne, David G.; Lang, Virginia A.; and Blackwell, Jason M. (1995) Mixed versus pure display format in integration and nonintegration visual display monitoring tasks. *Human Factors*, 37(3), 507-527.

Perlman, Gary and Swan, J. Edward II. (1993) Color coding versus texture coding to improve visual search performance. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting -- 1993*, vol. 1, October 11-15, Seattle,WA, 343-347.

Perlman, Gary and Swan, J. Edward II. (1994) Relative effects of color, texture, and density coding on visual search performance and subjective preference. *In Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting,* October 24-28, 1994, Nashville, Tennessee, 235-239.

Quinlan, Philip T. and Humphreys, Glyn W. (1987) Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception and Psychophysics*, 41(5), 455-472.

Rice, John F. (1991) Display color coding: 10 rules of thumb. *IEEE Software*, January 1991, 86-88.

Risch, J.S.; Rex, D.B.; Dowson, S.T.; Walters, T.B.; May, R.A.; and Moon, B.D. (1997) 42-49.

Rogers, Wendy A.; Lee, Mark, D.; and Fisk, Arthur D. (1995) Contextual effects on general learning, feature learning, and attention strengthening in visual search. *Human Factors*, 37(1), 158-172.

Salton, Gerard. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. New York: Addison-Wesley.

Salton, G.; Wong, A.; & Yang, C. (1975) A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613- 620.

Sarkar, Manojit and Brown, Marc H. (1994) Graphical fisheye views. *Communications of the ACM*, 37(12), 73-84.

Schneider, W. and Shiffren, R. M. (1977) Controlled and automatic human information processing I: Detection, search, and attention. *Psychological Review*, 84, 1-66.

Schulman, Robert S. (1992) *Statistics in Plain English with Computer Applications*. New York: Van Nostrand Reinhold.

Shneiderman, Ben. (1987) *Designing the User Interface*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Smallman, H. S. and Boynton, R. M. (1990) Segregation of basic colors in an information display. *Journal of the Optical Society of America,* 7(102), 1985-1994.

Smith, A. R. (1979) Color gamut transform pairs. *Proceedings of the ACM SIGGRAPH 79 International Conference on Computer Graphics*, 276-283.

Smith, L. and Thomas, D. (1964) Color versus shape coding in information displays. *Journal of Applied Psychology*, 48(3), 137- 146.

Smith, S. L. (1962) Color coding and visual search. *Journal of Experimental Psychology*, 64(5), 434-440.

Sternberg, S. (1966) High-speed scanning in human memory. *Science, 153*, 652-654.

Swierenga, Sarah J.; Boff, Kenneth R.; and Donovan, Rebecca S. (1991) Effectiveness of coding schemes in rapid communication displays. *Proceedings of the Human Factors Society 35th Annual Meeting — 1991, Volume 2*, September 2-6, 1991, San Francisco, California, 1522-1526.

Tan, Kay C. (1990) *Effects of Stimulus Class on Short -Term Memory Workload in Complex Information Displays*. Doctoral thesis for the Department of Industrial Engineering and Operations Research, Virginia Tech, Blacksburg, Virginia, May, 1990.

Theeuwes, J. (1990) Perceptual selectivity is task dependent: evidence from selective search. *Acta Psychologica*, 74, 81-99.

Travis, David. (1991) *Effective Color Displays: Theory and Practice*. New York: Academic Press.

Treisman, Anne. (1991) Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and PerformanceI,* 17(3), *652-676*.

Treisman, A. and Gormican, S. (1988) Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 95(1), 15-48.

Treisman, Anne and Sato, Sharon. (1990) Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3), 459-478.

Tufte, Edward R. (1983) *The Visual Display of Quantitative InformationI*. Cheshire, CT: Graphics Press.

Tufte, Edward R. (1990) *Envisioning Information*. Cheshire, CT: Graphics Press.

Umbers, I. G. and Collier, G. D. (1990) Coding techniques for process plant VDU formats. *Applied Ergonomic*s, 21(3), 187-198.

Van Orden, Karl F.; Divita, Joseph; and Shim, Matthew J. (1993) Redundant use of luminance and flashing with shape and color as highlighting codes in symbolic displays. Human Factors, 35(2), 195-204.

Venturino, Michael. (1991) Automatic processing, code dissimilarity, and the efficiency of successive memory searches. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3) 677-695.

Walker, P. (ed.) (1988) Chambers Science and Technology Dictionary. New York: Chambers/Cambridge, 1988.

Wickens, Christopher D. (1989) Attention and skilled performance. In D. Holding (Ed.), *Human Skills*, Chichester, England: Wiley.

Wickens, Christopher D. (1992) Engineering Psychology and Human Performance, 2nd. Ed. New York, NY: HarperCollins.

Wickens, C. D. and Andre, A. D. (1990) Proximity compatibility and information display: effects of color, space, and objectness on information integration. *Human Factors*, 32(1), 61-77.

Wickens, Christopher D. and Carswell, C. Melody. (1995) The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37(3), 473-494.

Wickens, Christopher D.; Merwin, David H.; and Lin, Emilie L. (1994) Implications of graphics enhancements for visualization of scientific data: dimensional integrality, stereopsis, motion, and mesh, in *Human Factors*, vol. 36, no. 1, 44-61.

Winer, B. J.; Brown, Donald R.; and Michels, Kenneth M. (1991) *Statistical Principles in Experimental Design, Third Edition.* New York: McGraw-Hill.

VITA

Lucille (Lucy) Terry Nowell

EDUCATION:

PhD Computer Science, Virginia Tech, Blacksburg, VA. 1997
Dissertation Title: *Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size To Convey Nominal and Quantitative Data*

MS Computer Science, Virginia Tech, Blacksburg, VA. 1993,

MFA Drama (Design), University of New Orleans, New Orleans, LA. 1982
MA Theatre (Design), University of Alabama, Tuscaloosa, AL), 1974
BA Speech, University of Alabama, Tuscaloosa, AL, 1972

RESEARCH INTERESTS:

Digital electronic libraries, information visualization, user interface design for information retrieval systems and digital electronic libraries, methods of formative usability evaluation, design description methods, dynamic user modeling and intelligent interfaces.

PROFESSIONAL EXPERIENCE:

- Associate Professor, Computer Science Department, Lynchburg College, Lynchburg, VA. 1994- present. Department chair, 1997-present.

  Courses taught: Computer Concepts, PC WordProcessing, PC Spreadsheet Use, PC Database Management, Structured Programming (CS1), Data Structures, Programming Langauges, Database Management Systems, Operating Systems, Artificial Intelligence, Special Problems in Human-Computer Interaction

- Summer Intern, IBM Thomas J. Watson Research Center, Information Design and Access Department. Hawthorne, New York. June-July, 1996. Developed system integration and user interface requirements document for digital library system.

- Graduate Research Assistant, Project Envision, Department of Computer Science, Virginia Tech, Blacksburg, VA. Oct. 1991-Dec. 1994. Designed user interface for multimedia database of computer science literature: Interviewed users. Performed task analysis. Developed user interface specifications. Supervised programmers. Conducted formative usability evaluation.

- Graduate Teaching Assistant, Department of Computer Science, Virginia Tech, Blacksburg, VA. Aug. 1990-Sept. 1991.

- Associate Professor, Dramatic Arts and Speech Communication Department — assigned to Computer Science Department, Lynchburg College, Lynchburg, VA. 1988-1993 (on leave for graduate study at Virginia Tech 1990-1993).

PROFESSIONAL EXPERIENCE (CONT.):

- Associate Professor, Dramatic Arts and Speech Communication Department, Lynchburg College, Lynchburg, VA. 1987-1988. Director of theatre program.

- Assistant Professor, Dramatic Arts Department, Lynchburg College, Lynchburg, VA. 1979-1987. Director of theatre program during 1986-87. See above for courses taught.

- Instructor, Dramatic Arts Department, Lynchburg College, Lynchburg, VA. 1976-1979.

- Graduate Teaching Assistant, Speech and Theatre Department, University of Alabama, Tuscaloosa, AL. 1973-1974.

PUBLICATIONS:

Nowell, Lucy Terry; France, Robert K.; and Hix, Deborah. (1997) Exploring Search Results with Envision. Two-page abstract for formal and participatory demonstrations, *CHI '97 Conference on Human Factors in Computing Systems — Extended Abstracts,* Atlanta, Georgia, March 1997, 14-15.

Nowell, Lucy Terry. (1997) Graphical Issues in Information Visualization.Two-page abstract for CHI doctoral consortium and interactive poster, to appear in *CHI '97 Conference on Human Factors in Computing Systems — Extended Abstracts*, Atlanta, Georgia, USA, March 1997, 65-66.

Nowell, Lucy Terry; France, Robert K.; Hix, Deborah; Heath, Lenwood S.; and Fox, Edward A. (1996) Visualizing Search Results: Some Alternatives to Query-Document Similarity. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Zurich, Switzerland, August 1996, 67-75.

Fox, Edward A.; Barnette, Dwight N.; Shaffer, Clifford A; Heath, Lenwood S.; Wake, William; Nowell, Lucy T.; Hix, Deborah; and Hartson, Rex. (1995) Progress in Interactive Learning with a Digital Library in Computer Science. Invited paper for *Proceedings of ED-MEDIA 95, World Conference on Educational Media and Hypermedia, Graz, Austria*, June 17-21, 1995, 7-12.

Heath, L.; Hix, D.; Nowell, L.; Wake, W.; Averboch, G.; Guyer, S. and Fox, E. (1995) Envision: A User-Centered Database of Computer Science Literature. In *Communications of the ACM*, April 1995, 52-53.

Fox, Edward A.; Hix, Deborah; Nowell, Lucy T.; Brueni, Dennis J.; Wake, William C.; Heath, Lenwood S.; and Rao, Durgesh. (1993) Users, User Interfaces, and Objects: Envision, a Digital Library. In *Journal of the American Society for Information Science* , 44(5) Sept. 1993, 480-491.

Nowell, Lucy T. and Hix, Deborah. (1993) Visualizing Search Results: User Interface Development for the Project Envision Database of Computer Science Literature. In *Proceedings of HCI International '93, Orlando*, August 8-13, 1993, vol. 2, 56-61. Elsevier, New York, 1993.

Nowell, Lucy Terry and Hix, Deborah. (1993) Query Composition: Why Does It Have to be So Hard? In *Proceedings of East-West International Conference on Human-Computer Interaction, 1993,* Moscow. August 1993, 226-241.

PUBLICATIONS (CONT.):

Brueni, Dennis J.; Fox, Edward A.; Heath, Lenwood S; Hix, Deborah; Nowell, Lucy T.; and Wake, William C. (1993) What If There Was Desktop Access to the Computer Science Literature? In S. C. Kwasny (Ed.), *Proceedings of ACM Computer Science Conference: CSC '93*. Indianapolis, February 1993, 15-22. New York: ACM Press, 1993.

Nowell, Lucy T. and Hix, Deborah. (1992)User Interface Design for the Project Envision Database of Computer Science Literature. In *Proceedings of Virginia Computer Users Conference*. Blacksburg, VA. October, 1992,29-33.

SELECTED AWARDS AND HONORS:

CHI'97 Doctoral Consortium participant, ACM Conference on Human-Computer Interaction, Atlanta, GA, March 1997

ACM SIGIR Student Travel Grant for participation in ACM SIGIR '96, Zurich, Switzerland

Delegate, Computing Research Association/NSF Windows of Opportunity Symposium for Women in Computer Science, representing Virginia Tech. Arlington, VA. 1993

Upsilon Pi Epsilon National Computer Science Honor Society, Virginia Tech Chapter, since 1991

Graduate Council Fellow, University of Alabama, 1972-1973

Alumni Honors Scholar, University of Alabama, 1969-1972