Bennett, K. B., Nagy, A. L., & Flach, J. M. (2006). Visual displays. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (Third ed., pp. 1191-1221). New York, NY: Wiley.

Visual Displays

Kevin B. Bennett, Allen L. Nagy, and John M. Flach

20.1.0 Introduction.

Advances in computer science and artificial intelligence currently provide new forms of computational power with the potential to support human problem solving. One use of this computational power is to provide an expert system or an automatic assistant which provides "advice" to the human operator at the appropriate times. An alternative use is to integrate the information graphically (or more generally "perceptibly"). Here computational power is used to create and manipulate representations of the target world, rather than to create autonomous machine problem solvers. Perhaps the most general term that has been applied to this endeavor is "representation aiding" (Zachary, 1986; Woods and Roth, 1988; Woods, 1991). The emphasis in this chapter will be on representation aiding. However, we see these as complementary tools in the designer's tool chest and we expect that for very complex systems both approaches will be necessary.

Representation aiding offers a unique opportunity to improve overall performance of human-machine systems. Other approaches to decision support are hampered, to a certain degree, by the maturity of the technologies on which they are based. For example, although there has been a great deal of progress in the use of production systems and neural networks as the drivers for decision support, it is clear that additional strides need to be made before these types of systems will be practical in applied settings. On the other hand, the technologies needed to produce computer graphics are relatively more mature. The challenge in representation aiding centers around how best to use these technological capabilities to support human decision making and problem solving.

In this chapter we review issues that are relevant to meeting this challenge. In contrast to most other treatments of display design, we do not provide a "cook book" of detailed guidelines and recommendations (primarily because they tend to be conflicting and difficult to apply). Instead, we chose to describe a set of general heuristics for display design. Because these heuristics are necessarily abstract, we have made the discussion more concrete by illustrating them within the context of a simple domain. We describe how the heuristics apply to that domain and annotate our written descriptions with concrete graphic examples. Our goal is to transfer functional knowledge of display design to practitioners.

We begin our discussion with a description of basic physiological, perceptual, and technological considerations in display design. These considerations are the foundation for display design, and represent the baseline conditions that must be met for a display to be effective. We next consider four alternative approaches to display design. Each approach emphasizes a different conceptual aspect of the display design puzzle, and each approach has both strengths and weaknesses. A fifth approach is outlined; this approach draws from the strengths of the previous approaches and incorporates new considerations that are particularly relevant to the design of displays for complex, dynamic domains. We end the chapter by considering the limitations of our discussion and examples and additional challenges for display design.

20.2.0 Physiological, Perceptual, And Technological Considerations

This section considers fundamental aspects of the visual system and visual perception that are relevant for display design. A visual display is most often represented by a difference in perceived brightness or a difference in perceived color between the graphical elements that comprise the display and the background of the display field. This section is concerned primarily with the detection

and perceived appearance of these differences. Although this chapter is focused primarily on emissive displays, it is useful to begin by discussing some of the differences between reflective and emissive displays and the implications of these differences for visual perception. Emissive displays, such as the CRT, generate the light that is used to produce text, symbols, or pictures that carry information. Reflective displays such as road signs, pages in a text book, and the speedometer in an automobile do not produce any light, but reflect some portion of the light that falls on them. Though emissive displays are much more versatile and flexible in some respects, it is probably safe to say that the use of reflective displays to present information was, and still is, far more common. With regard to the visual system and visual perception there are some fundamental differences between reflective and emissive displays. We will begin by examining properties of achromatic, or colorless, displays that illustrate these differences and later in this section take up chromatic displays.

20.2.1 Reflective Displays

The surface of a reflective display reflects some portion of the light energy that falls on it in many different directions. The percentage of light reflected (the "reflectance") and the dependence of this percentage on the wavelength of the light (the "spectral reflectance function") are determined by the physical properties of the surface (Nassau, 1983). We will begin by discussing surfaces with flat spectral reflectance functions that reflect approximately the same percentage of light for all wavelengths. Images are placed on the surface by changing the properties of the surface in local regions. For example, suppose a printer for a personal computer deposits black ink on a grey page so as to form text. The grey page reflects a percentage, perhaps 50%, of the light energy at each wavelength falling on it. The ink deposited on the page appears very dark because it reflects only a small percentage, for example 5%, of the light energy falling on it. Suppose an observer

views this page tacked to a wall painted uniformly white so that the surface of the wall has a reflectance of 90%.

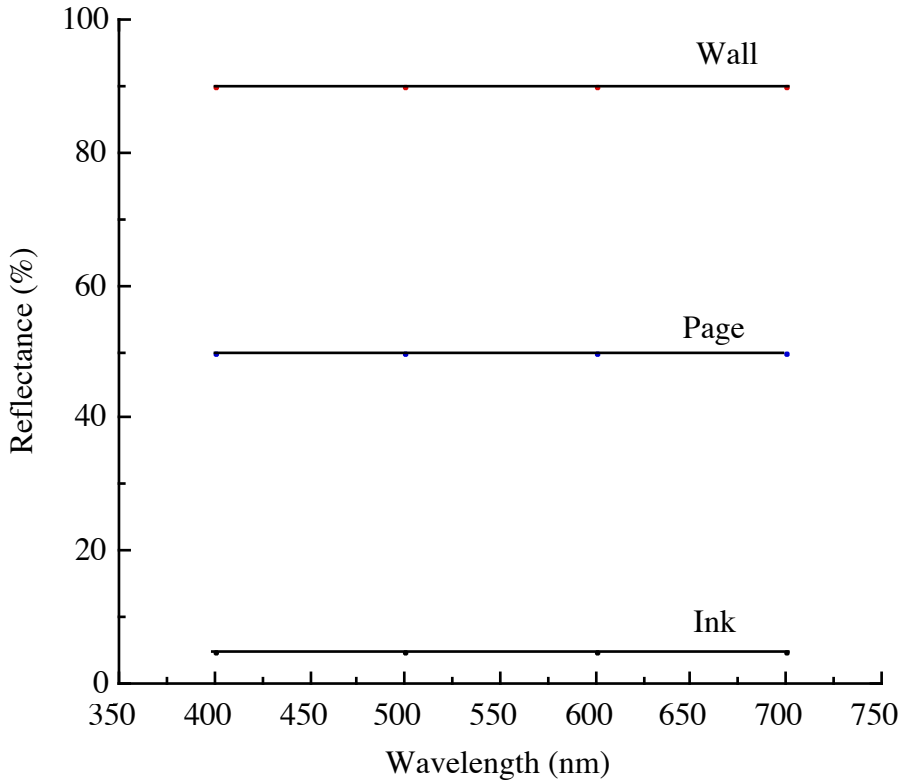
The reflectance of surfaces varies with the angle of incidence of the illumination and the angle at which the reflectance is measured. Reflectances of surfaces can be described with two components, a specular component and a diffuse component (Shafer, 1985; Hunter and Herold, 1987). The specular component is "mirror-like" in that a large proportion of the light is reflected off at an angle equal to the angle of incidence. The diffuse component is characterized by light reflected off in all directions. Shiny surfaces, like mirrors, have a large specular component and a small diffuse component while matte surfaces, like a velvet cloth) have a large diffuse component and a small specular component. For simplicity we will ignore these complexities here. The graph shown in Figure 1a illustrates idealized spectral reflectance curves for the page, the ink, and the wall. Real spectral reflectance curves would only approximate flat curves. Surfaces with flat curves are neutral in the sense that they do not change the spectral quality of the light that falls on them.

==========

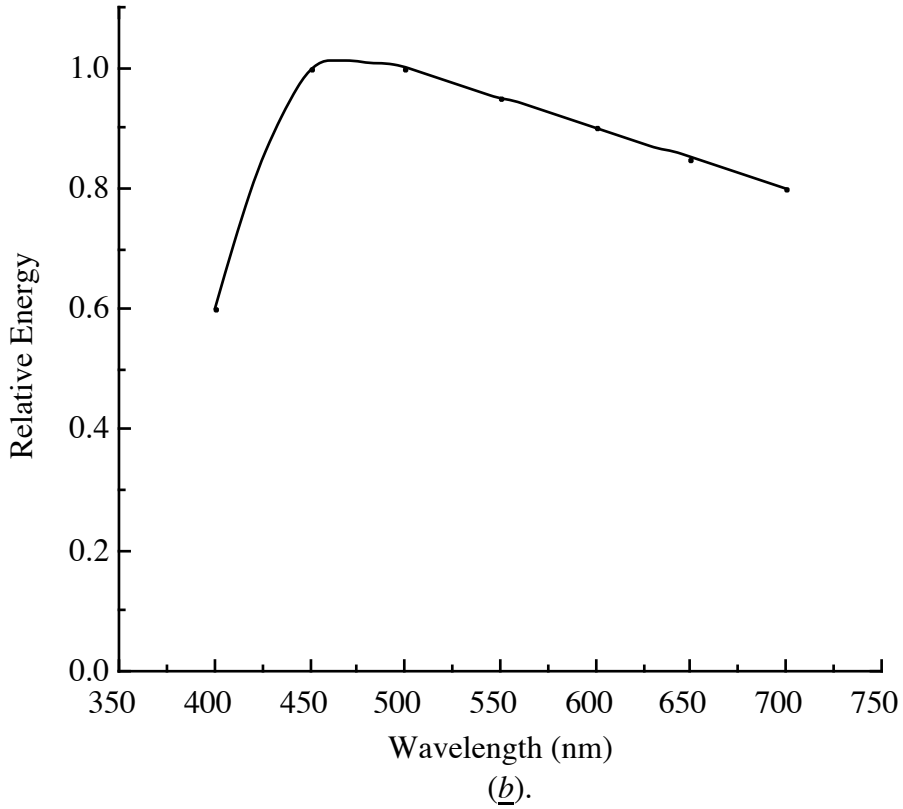
Insert Figure 1 about here

==========

In order to characterize the light reflected back from the surface, we need to know something about the light falling on the surface. A typical spectrum for sunlight is shown in Figure 1b, where the relative energy is plotted as a function of wavelength. This spectrum is referred to as typical, because the spectrum for sunlight varies with time of day, time of year, latitude, and atmospheric conditions. Not all of the energy in sunlight is effective in generating a visual response. Some wavelengths of light are more likely to be absorbed by the receptors in the eye, the rods and cones,



(*a*).



(*b*).

Figure 1 Figure 1a shows idealized spectral reflectance curves for the ink, the page, and the wall in the example described in the text. Figure 1b shows the relative energy at each wavelength in sunlight.

than others. A function describing the relative effectiveness of different wavelengths for photopic or cone vision (See Figure 2) was standardized by the CIE in 1924 (See Wyszecki and Stiles, 1982). This function, which is known as the photopic luminosity function, has served as a standard in science and industry ever since.

==========

Insert Figure 2 about here

==========

A similar function for scotopic or rod vision was standardized in 1951 (See Wyszecki and Stiles, 1982). Since most displays are viewed under photopic conditions, we will concentrate on cone vision here. In order to get a measure of the visual effectiveness of the light energy from the sun we multiply the energies at each wavelength in Figure 1b by the value of the photopic luminosity function at that wavelength. The sum or integral of these weighted energies, multiplied by a constant to convert the energy units to a convenient unit of visual effectiveness, is known as the luminance of the source. A commonly used unit for luminance is the Candela/sq. meter.

For our purposes, the more important measure is the amount of light that actually falls on the wall, the page, and the ink. This quantity is known as illuminance, the amount of visually effective light that actually falls on a surface in space. We will assume that the wall is evenly illuminated so that this measure is the same across the wall, the text, and the page. A common unit of illuminance is the Lux. The measurement of luminance, and the related quantity illuminance, is itself a complex topic, and many different types of units are used in measuring light (for discussions of light measurement see Wyszecki and Stiles, 1982; and Grum and Bartleson, 1980). In order to find the amount of visually effective light reflected from the surface, we multiply the reflectance at each wavelength times the illuminance provided by the sunlight at each wavelength. Alternately, we

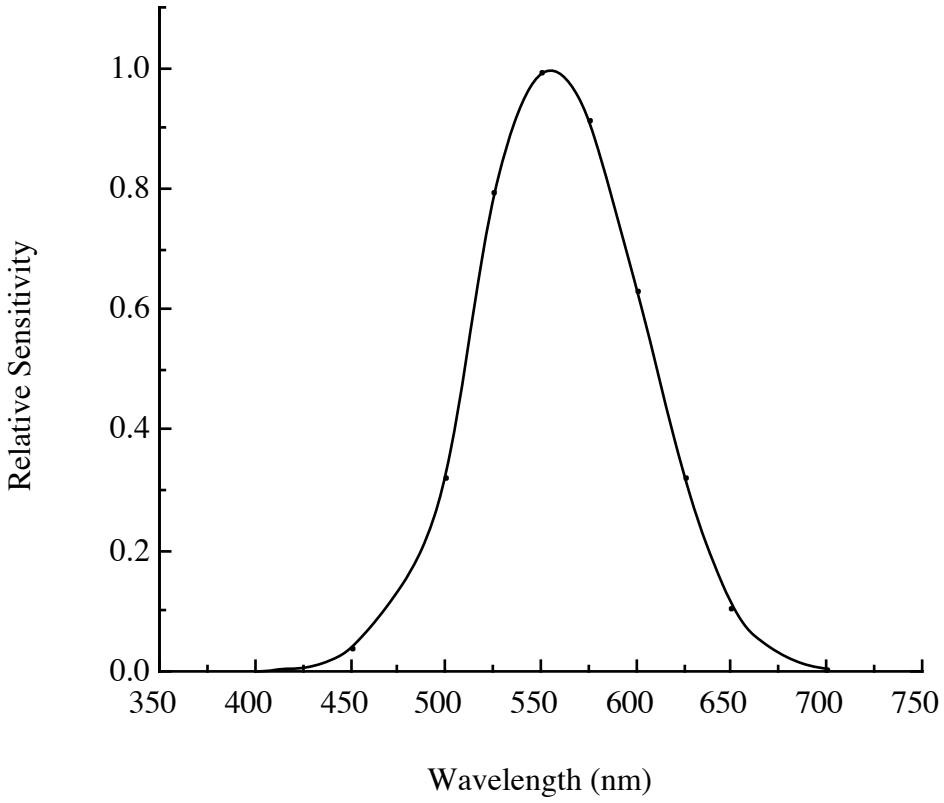


Figure 2 The CIE 1924 photopic luminosity curve.

could measure the amount of visually effect light reflected in a particular direction directly using a device called a photometer (for a discussion of devices for measuring light see Post, 1992).

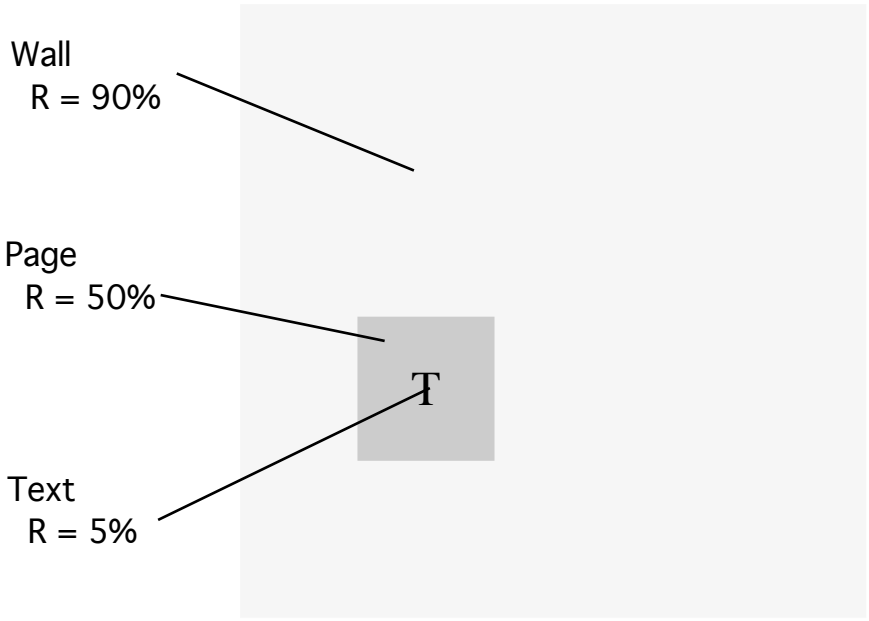
An important property of reflective displays, such as our page of printed text mounted on the wall, is that the physical contrast between the text and the page, or the page and the wall, does not vary with the amount of light falling on them as long as all of the surfaces are illuminated at the same level. The term "physical contrast" is used to refer to the difference in the light reflected from two regions of a scene. The physical contrast of a stimulus on a background is often defined as the "contrast ratio", $\Delta L/L$, the difference between the light reflected from the stimulus and the background divided by the background level. In our example the physical contrast between the text and the page could be specified as the difference in the amounts of light reflected by the ink and by the page divided by the amount of light reflected by the page. Note that as the amount of light falling on the wall is changed, the physical contrast ratios calculated for the text and the page, the text and the wall, and the page and the wall will remain constant (see Figure 3). The reader can verify this by calculating the contrast ratios in the present example. The light level, which appears in both the numerator and the denominator of the contrast ratio, cancels out and the contrast ratios are determined by the reflectances alone.

==========

Insert Figure 3 about here

==========

The human visual system appears to have evolved to take advantage of the reflective properties of surfaces. One of the earliest relationships established in the study of visual perception is that the intensity difference between a stimulus and a background necessary for detection is a constant proportion of the intensity of the background field. This rule, known as Weber's Law, is often



Contrast ratios:

$$\text{text/page} = \frac{0.5 \bullet I - 0.05 \bullet I}{0.5 \bullet I} = \frac{0.45}{0.50} = 0.90$$

$$\text{page/wall} = \frac{0.9 \bullet I - 0.5 \bullet I}{0.9 \bullet I} = \frac{0.40}{0.90} = 0.44$$

$$\text{text/wall} = \frac{0.9 \bullet I - 0.05 \bullet I}{0.9 \bullet I} = \frac{0.85}{0.90} = 0.94$$

Figure 3 The diagram illustrates the calculation of contrast ratios for the page, the text, and the wall. The values of "r" indicate the reflectances of the three surfaces in the figure. The symbol "I" in the equations represents the illumination level which is identical for all three surfaces in the figure and therefore cancels out of the equations.

written in equation form as $\Delta I = k * I$. Here ΔI refers to the difference between the intensity of the stimulus and the intensity of the background, k is the proportionality constant or the Weber Fraction, and I is the intensity of the background field. Weber's Law indicates that the visual system becomes less sensitive to differences between the stimulus and the background as the intensity of the background field increases. That is, in order to keep the stimulus detectable, the difference in intensity between the stimulus and the background must be increased as the intensity of the background is increased. Notice, however, that if we rearrange Weber's Law by dividing both sides of the equation by I, we get $(\Delta I/I) = k$.

At threshold levels of stimulation the difference between the intensities of the stimulus and the background (ΔI) divided by the background intensity (I) is constant. This is exactly the situation for the reflective displays described above. Thus, if the text on a page is detectable at any light level, then it will remain detectable as the light level is changed. A somewhat different form of Weber's Law also applies to the discrimination of two stimuli presented on a background. In this case, at threshold the difference in the contrasts between the two stimuli relative to the background must be a constant proportion of one of the contrasts (Whittle, 1986, 1992; Nagy and Kamholz, 1995). Thus for reflective displays, if two stimuli at different contrast levels on a background are discriminable from each other they will remain discriminable as the illumination level is changed. It is well known that Weber's Law is only approximately true and that it breaks down under many conditions, perhaps most importantly when the light levels involved are low and approach absolute threshold. However, the change in the sensitivity implied by Weber's Law is an important property of the visual system. It is a component of another property of the visual system known as "lightness constancy." Lightness constancy refers to the fact that the visual system operates in such a manner as to keep the perceived appearance of reflective objects approximately constant under changing

illumination levels. That is the wall, the page, and the text in our example appear white, grey and black, respectively, whether they are viewed outdoors under intense sunlight or indoors under dim illumination. Lightness constancy depends on many other factors in addition to the change in sensitivity indicated by Weber's Law, and has been a topic of intense interest in the last couple of decades (Gilchrist, Delman, and Jacobson, 1983; Adelson, 1993).

20.2.2 Emissive Displays

We will use a CRT as an example of an emissive display. CRT's generate light by shooting beams of electrons at substances called phosphors which are painted on the screen of the CRT. When the electrons hit a point on the screen, light energy is given off by the phosphor at that point. The intensity of the light given off can be changed by varying the strength of the beam of electrons directed at the point. Images are created on the screen by varying the intensity of the electron beam hitting different points on the screen. The physical contrast between different regions of the screen can be defined in the same manner as for reflective displays.

Suppose that we mount the CRT on the white wall and use it to generate a page of dark text on a grey page. Suppose also that we adjust the CRT so that the page gives off 50 units of light and the text gives off 5 units of light. The physical contrast ratio between the text and the page would be 0.90 as it was for the reflective display (see Figure 3). Suppose that the white wall is illuminated initially so that 90 units of light are reflected from it. Also suppose, for the moment, that the surface of the CRT reflects none of this light. In this case the contrast ratios between the three surfaces would be the same as in our first example with the reflective page of text, and we might expect that the CRT display would look very similar to the reflective display.

Note what happens as the illumination falling on the wall is increased, however. The intensity

of the light reflected from it increases, but the intensities of the lights from the text and the page on the CRT do not change. The contrast ratio between the text and the page on the CRT remains constant, but the contrast ratios between the page and the wall, and the text and the wall increase. Thus we might expect that the appearances of the text and the page to change considerably as the light level falling on the wall is changed. If we regard the text and the page as individual incremental stimuli against the large background provided by the wall, then Weber's Law suggests that their discriminability will decrease as the light reflected from the wall increases. The decrease in discriminability occurs because the difference in contrast ratios decreases with increasing light level. In this case the decrease in the sensitivity of the visual system with increasing background light level reduces the ability to detect the difference between the text and the page which remains constant.

Any light which is reflected from the glass face of the CRT will reduce the discriminability of the text on the page even further, because it will be reflected from both the region containing the dark text and the region containing the page. The reflected light actually reduces the physical contrast between the text and the page and makes them even less discriminable. Thus emissive displays behave quite differently than reflective displays in natural environments. These differences do not present much of a problem when emissive displays are placed in a constant environment such as an office illuminated by a fixed light source. However, when emissive displays are placed in natural environments in which the illumination level may vary by a factor of a million or more, the problems caused by the varying contrast ratios are evident. For example, this problem occurs when emissive displays are used in aircraft. The detectability and the appearance of elements within the display may vary dramatically. In order to keep the appearance of the text and the page constant, the light levels given off by the CRT must be adjusted in accord with the change in the

illumination of the wall.

20.2.3 Factors Affecting Perceived Contrast

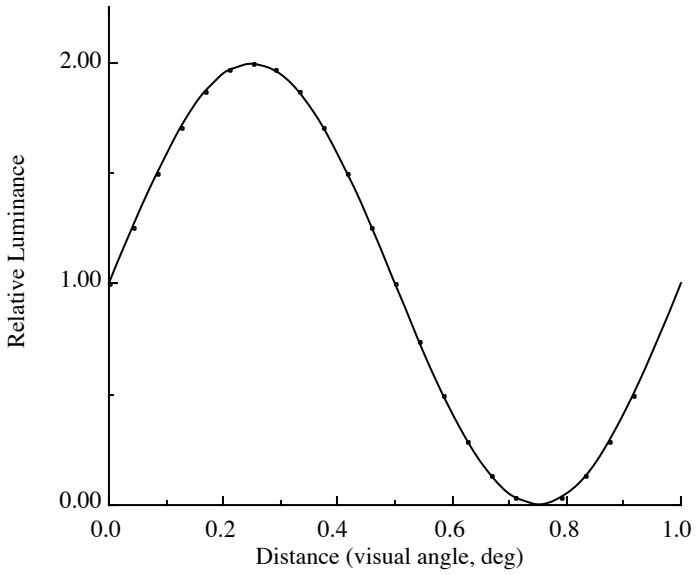
Besides the physical contrast there are many other factors such as adaptive state, location in the visual field, eye movements, and the interpretation of the perceived illuminant which affect the perceived contrast of a stimulus against a background. One of the most important of these factors is stimulus size. In the last few decades this problem has been investigated very successfully with an approach based on Fourier analysis (for extensive reviews see Ginsberg, 1986; Olzack and Thomas, 1986; Graham, 1989; and DeValois and DeValois, 1990). Fourier analysis suggests that any pattern of light and dark on the retina can be described as a sum of sinusoidal components of different frequency and amplitude. The application of this idea to visual perception involves measuring an observer's sensitivity to a number of sinusoidal patterns of different spatial frequency (see Figure 4). These repetitive spatial patterns of light and dark are known as gratings.

==========

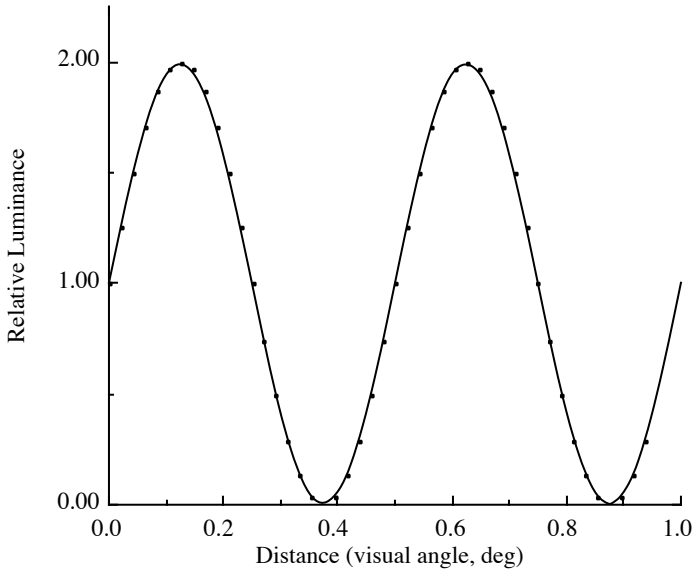
Insert Figure 4 about here

==========

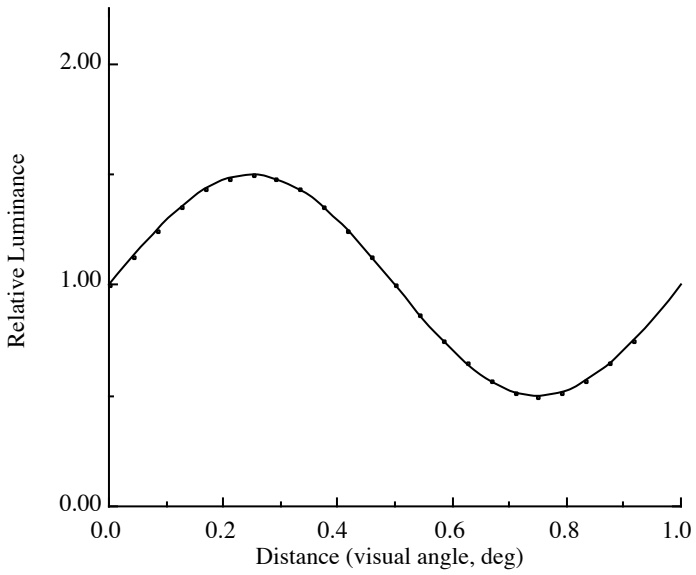
Spatial frequency is essentially a measure of the size of the bars in the pattern. The greater the spatial frequency the more cycles occur per degree of visual angle and the smaller the bars. Visual angle is used as the unit of size because it gives a measure of the size of the image on the retina. A book 12 inches long makes a larger image on the retina when it is held up close to the eye than when it is held far away. In order to get a measure of the size of an image on the retina the distance between the book and the observer's eye must be taken into account. Thus the visual angle subtended by an object is defined as twice the arcTan of the heigth/2 divided by the distance (See Figure 5).



(*a*).

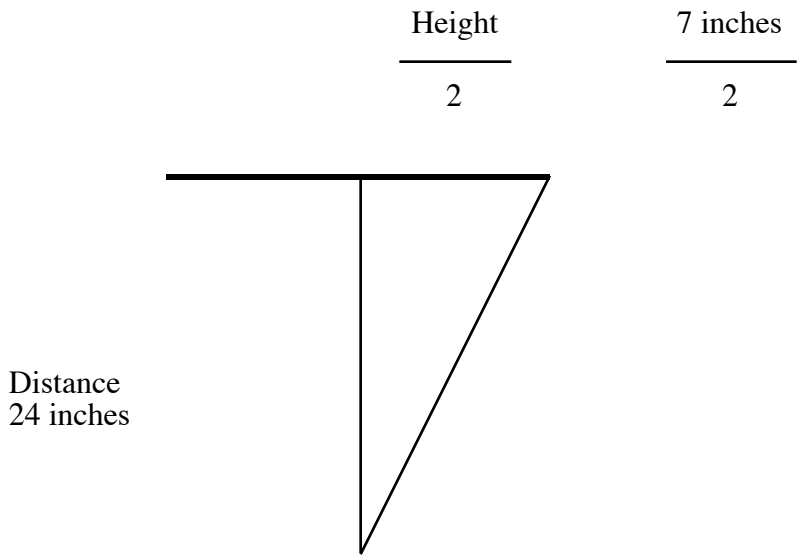


(*b*).



(*c*).

Figure 4 Plots showing the variation in luminance for sinusoidal patterns.
Figure 4a illustrates a spatial frequency of 1 cycle/degree at a contrast of 100%.
Figure 4b illustrates a spatial frequency of 2 cycles/degree at a contrast of 100%.
Figure 4c illustrates a spatial frequency of 1 cycle/degree at a contrast of 50%.



$$\text{visual angle} = 2 \bullet [\arctan(\text{height}/2/\text{distance})]$$

$$= 2 \bullet [\arctan (3.5/24)]$$

$$= 2 \bullet [\arctan 0.1458]$$

$$= 2 \bullet 8.3 \text{ degrees}$$

$$= 16.6 \text{ degrees}$$

Figure 5 The diagram illustrates the calculation of visual angle as described in the text.

==========

Insert Figure 5 about here

==========

Sensitivity is measured by finding the physical contrast level at which a given pattern of light and dark is just visible. In order to give a measure of sensitivity, the reciprocal of the threshold is calculated by dividing one by the threshold contrast. The measure of physical contrast typically used in this approach is slightly different than the contrast ratio described above, and is called the "Michelson Contrast." It is defined as Lmax - Lmin divided by Lmax + Lmin, where Lmax is defined as the maximum luminance level in the pattern and Lmin is defined as the minimum luminance in the pattern. The curve described by plotting contrast sensitivity against the spatial frequency of the grating pattern is called the "contrast sensitivity function." The spatial frequency of the pattern is defined as the number of cycles that occur in one degree of visual angle.

A typical contrast sensitivity function for photopic or cone vision obtained from a human observer is shown in the Figure 6. The curve shows that when spatial frequency is low (i.e. the bars are large), the sensitivity to contrast is low. As spatial frequency is increased the sensitivity increases up to spatial frequencies of about 5 to 10 cycles per degree. With further increases in spatial frequency (i.e. smaller and smaller bars), sensitivity falls off rapidly until at a spatial frequency of approximately 50 cycles/deg a grating of 100% contrast (the highest physical contrast obtainable) is not visible. Spatial patterns of even greater frequency also are not visible. Thus, very fine patterns are visible only if the spatial frequency is below 50 cycles/degree and if they have very high contrast.

==========

Insert Figure 6 about here

==========

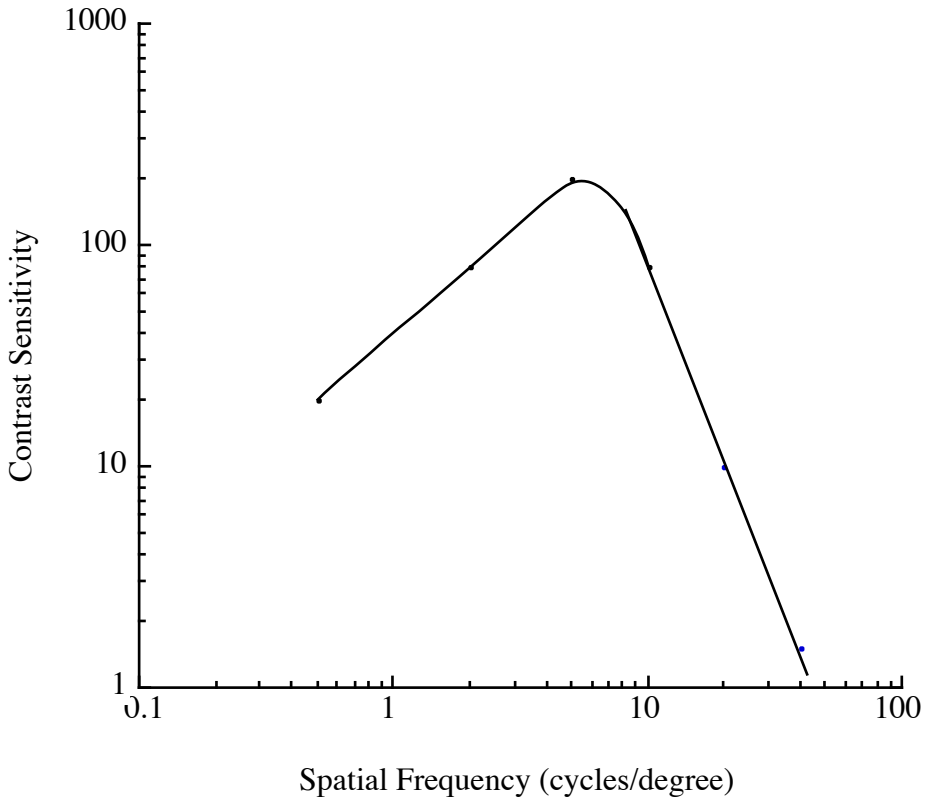


Figure 6 A typical plot of a contrast sensitivity function for a human observer. Based on data given in DeValois and DeValois (1990).

Many physical factors have been shown to affect the contrast sensitivity function, including overall light level, number of cycles present in the pattern, and the location of the pattern in the visual field. The shape of the curve as well as the overall sensitivity can vary considerably. The shape and height of the curve are affected by several components within the visual system that play a role in determining the contrast sensitivity function. For example the optics of the eye, the lens and cornea, which form an image of the pattern on the retina, influence the contrast sensitivity function, because they do not form a perfect image of the external pattern on the retina. A good introductory treatment of the optics of the eye is given by Millidot (1982). The distribution of rods and cones on the retina also plays a role in determining the contrast sensitivity function. The rods and cones absorb light and initiate neural signals in the visual system. Thus their size and the distances between them have some affect on the contrast sensitivity function. A good introduction to the sampling properties of rods and cones is given by Wandell (1995). The way the rods and cones are connected to the neurons that carry signals out of the eye also plays a role in determining the contrast sensitivity function, because many receptors are connected to each neuron. Psychophysical evidence suggests that the visual system may be organized into approximately 5 to 7 neural channels, each sensitive to a different band of spatial frequencies (Olzack and Thomas, 1986). Thus the contrast sensitivity function is the result of many factors which have been studied intensely over the last few decades. Nevertheless, it is a very useful and fundamental description of the ability of a human observer to detect contrast in patterns of different size. For example, recent studies suggest that the recognition of text may be mediated by the same mechanisms that mediate the contrast sensitivity function (Solomon and Pelli, 1994; Alexander, Xie, and Derlacki, 1994).

The perceived contrast of patterns that are well above threshold is not simply related to the contrast sensitivity function (see Cannon and Fullenkamp, 1991). That is, if we measure the threshold

contrast for sinusoidal patterns at a number of different spatial frequencies and then increase the physical contrast of all of these patterns so that the contrast for each one is 5 times the threshold contrast, the patterns will not appear to have equal contrasts. This is similar to the situation in audition where equal loudness curves for tones of different frequencies do not have the same shape as the audibility curve, a plot of threshold as a function of frequency, and change shape as the loudness level is raised. Thus the contrast sensitivity function can be used to predict whether a pattern of a given spatial frequency is visible, but it cannot be used to predict accurately the perceived contrast of patterns that are well above threshold. For example, if a display designer wants to equate the perceived contrast of patterns of different size that are well above threshold, the contrast sensitivity function cannot be used to do this accurately.

The notions of visual angle, spatial frequency, and contrast sensitivity that were briefly introduced above are very useful in thinking about both reflective and emissive displays. Here we will concentrate on emissive displays. Assume we have a standard CRT display that is 9.5 inches wide and 7 inches high with 640 columns of pixels each containing 480 rows (standard 640 X 480 resolution). If the observer views this display from a distance of 2 feet, then the screen subtends about 22.4 degrees horizontally and 16.6 degrees vertically (see Figure 5), and each pixel subtends about 0.035 degrees. If we want to make patterns of light and dark bars on the screen, we might want to know the highest spatial frequency that can be represented. If we make alternate pixels black and white we need two pixels to make one cycle, which will subtend 0.07 degrees. Thus the highest spatial frequency that can be represented accurately will be 1/.07 or slightly over14 cycles per degree.

Looking back at our representative contrast sensitivity function, we see that this frequency is

well below the upper limit of approximately 50 cycles per degree. Looking at the vertical axis we find that the sensitivity at 14 cycles per degree is approximately 30. In order for an observer to detect this pattern on the screen, we can determine that the Michelson contrast will have to be approximately 1/30 or 3.3%. These calculations also tell us something else. Patterns with spatial frequencies higher than 14 cycles per degree cannot be represented accurately on the monitor. Thus if we want to view an image with a lot of fine details at high spatial frequencies, such as a digitized photograph which subtends 9.5 by 7 inches, spatial frequencies greater than 14 cycles/degree that were visible when the original photograph was viewed from a distance of two feet, will not be accurately represented on the monitor if they are composed of spatial frequencies above 14 cycles/deg.

One solution to this problem is to use a monitor with higher resolution or smaller pixels. For example if we could pack 1280 x 960 pixels into the same 9.5 x 7 inch screen, patterns with spatial frequencies up to nearly 29 cycles per degree could be represented. In order to make a display that matches the upper limit on the resolution of the visual system we would need to pack about 2240 x 1660 pixels into the display. A 9.5 X 7 CRT with this resolution would permit the presentation of patterns with spatial frequencies up to 50 cycles/degree at a viewing distance of two feet. This would be very difficult to accomplish with present technology, making the display and the computer hardware that drives it very expensive.

It is also possible to portray patterns with spatial frequencies greater than 14 cycles/deg, on the original CRT by moving the observer farther away so that each pixel subtends a smaller visual angle. The drawback to this approach is that the entire display field now subtends a smaller portion of the field of view. For example if we move the observer back to a distance of about four feet,

patterns with spatial frequencies up to nearly 29 cycles/degree could be portrayed on the screen. This example illustrates a fundamental trade-off in emissive displays, the trade-off between field-of-view and resolution. With a fixed number of pixels, this trade-off is always present in an emissive display. If the pixels are spread over a larger viewing area the resolution will be poor. If they are packed into a smaller viewing area the resolution will improve, but the field of view will decrease.

The resolution of an emissive display may be limited either by the display itself, or by the hardware that drives it (i.e., the video card in a computer or the signals generated on a television cable). The detail in an image, or the spatial frequencies that can be portrayed, and the field of view that is visible, will be limited by this resolution and the size of the screen.

20.2.4 COLOR

Though black and white pictures carry much of the information in the real world they do not carry information about color. Color in images is certainly important for aesthetic reasons, but in addition to the aesthetic qualities it brings to an image it serves two important basic functions (Boynton, 1993). First, chromatic contrast between two regions in image can add to the luminance contrast between these regions to make the difference between the regions much more noticeable, especially when the luminance contrast is small. Second, since color is perceived to be a property of an object (though in fact it also depends on illumination, as we will see), it is useful in identifying objects, searching for them, or grouping them. Boynton (1993) regards the second function of color, which he describes as related to categorical perception, as the more important one.

It is probably because of these categorical properties, that color is often used as a coding device and as a means of segregating information in visual displays (see Widdel and Post, 1993).

Several excellent treatments of the basics of human color vision and the science of specifying colors for applications are available (Boynton, 1992; Pokorny and Smith, 1986; Post, 1995; Wyszecki and Stiles, 1982). So a very brief review will be given here. Normal human color vision depends on the presence three types of cone receptors in the retina. These cones differ in the type of light absorbing pigment contained in them. One of these pigments absorbs the greatest percentage of the light falling on it, in the short-wavelength region of the spectrum; hence the cone containing it is referred to as the "S" cone. The second pigment absorbs best in the middle of the spectrum, and the cone containing it is referred to as the "M" cone. The third pigment absorbs best at slightly longer wavelengths than the "M" pigment and the cone containing it is referred to as the "L" cone.

The differences in the signals generated in these cones by a given light, provide some information about the spectral content of the light. For example a light source which gives off more energy in the long- wavelength portion of the spectrum than in the middle or short wavelength regions would tend to stimulate the L cones more than the other two cone types. On the other hand, a light source that gives off more energy in the short-wavelength region would tend to stimulate the S cones more than the other two types. The differences in the stimulation of the cone types serve as a means for discriminating between the lights, and result in the perception of color.

Since there are only three types of cones, normal human color vision is said to be three dimensional or trichromatic. Furthermore, since there are only three signals from different types of cones in the visual system, it follows that only three numbers are needed to specify the perceptual quality of a color. Much effort has gone into developing systems of specifying colors with three numbers such that they represent the perceptual qualities of the stimulus in useful ways. The fact that only

three numbers are needed to specify the chromatic quality of a stimulus also means that there are many physically different stimuli that stimulate the three cones in the same way and thus appear to be the same color. Stimuli which are physically different but appear to be the same are called "metamers."

Consider once again our example of the reflective display. Suppose that we print the text on the gray page using red ink rather than black ink. The ink appears red because it tends to absorb short and middle wavelength light that falls on it while reflecting long wavelength light. A spectral reflectance curve showing the percentage of light reflected as a function of wavelength for red ink might look like the curve shown in Figure 7. To get the light reflected back from the ink we multiply the reflectance at each wavelength times the energy at each wavelength. In order to calculate the luminance of this light we would weight the reflected energy at each wavelength by the photopic luminosity function and integrate or sum over the entire curve as we did for achromatic stimuli previously. However, the text appears to differ from the grey page and the white wall in color as well as in lightness. In order to characterize this difference we would like some means of measuring the colors of the text, the page, and the wall. The most widely used system for doing this is based on the CIE 1931 chromaticity diagram. This diagram is based on color matches of normal human observers. A good introduction to the color matching experiment and the development of chromaticity diagrams can be found in Boynton (1992). In the color matching task, observers were asked to adjust the intensities of three primary lights that were mixed together in a single stimulus field so as to match the colors of a wide variety of other lights presented in another stimulus field. The CIE chromaticity diagram uses three numbers (related to the intensities of the primaries needed to make a match in the color matching experiment) to represent the color or, more specifically, the "chromaticity" of a stimulus. These numbers are called the chromaticity coordinates of the

color and are referred to as x, y, and z. The color matching data were normalized so that the values of these three chromaticity coordinates add up to 1 for any real color. As a result only two of the chromaticity coordinates need to be given to specify a color, because the third can always be obtained by subtracting the sum of the other two from 1. Therefore, all colors can be represented in a two dimensional diagram like the CIE 1931 diagram shown in Figure 8, where only x and y are plotted. Many measuring instruments have been developed and are commercially available for measuring the CIE coordinates of a color (see Post, 1992 for some discussion of these).

==========

Insert Figure 7 about here

==========

==========

Insert Figure 8 about here

==========

The chromaticity coordinates specify the chromatic properties of a color but do not specify its appearance, because the appearance of the color can change with many viewing conditions that do not change its chromaticity coordinates. For example, the size of the stimulus, in terms of visual angle, can affect the color appearance even though the chromaticity coordinates of the ink used to make it do not change (Poirson and Wandell, 1993). This is a severe limitation on the meaning and usefulness of the CIE chromaticity diagram. One would like to have a system in which the appearance of the color is specified, but this is a very difficult problem that has not yet been solved. Nevertheless, the specification of colors in the chromaticity diagram is still very useful, because any two stimuli with the same chromaticity coordinates will appear to be identical in color when viewed under the same conditions. What the chromaticity coordinates specify is how to make a color that will appear the same as a given sample under the same viewing conditions.

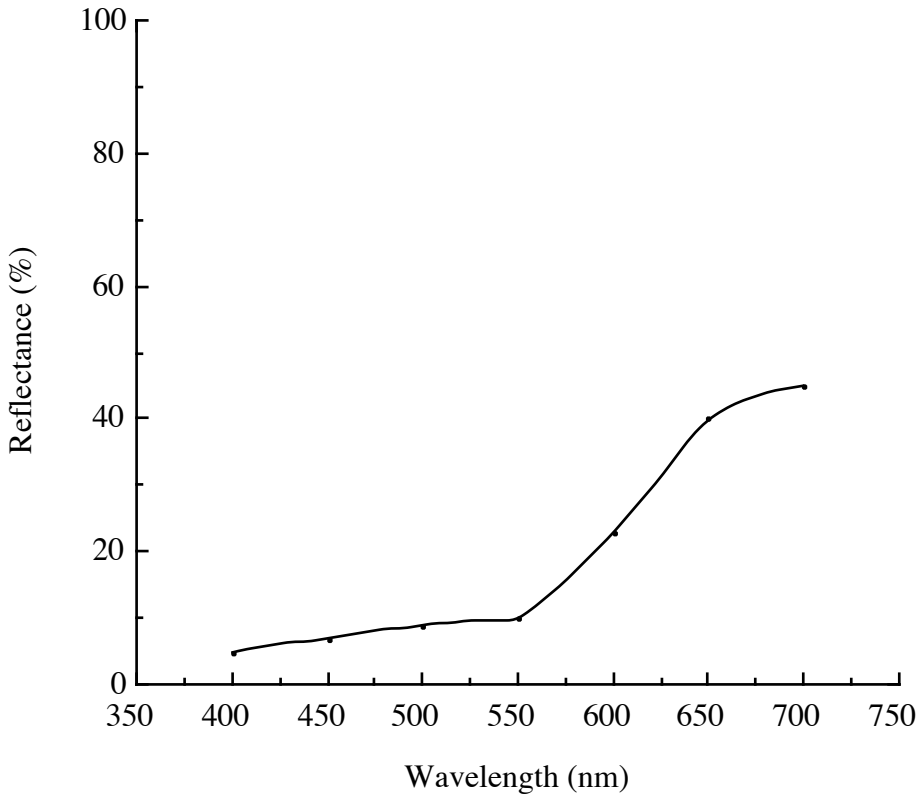


Figure 7 A spectral reflectance curve for red ink.

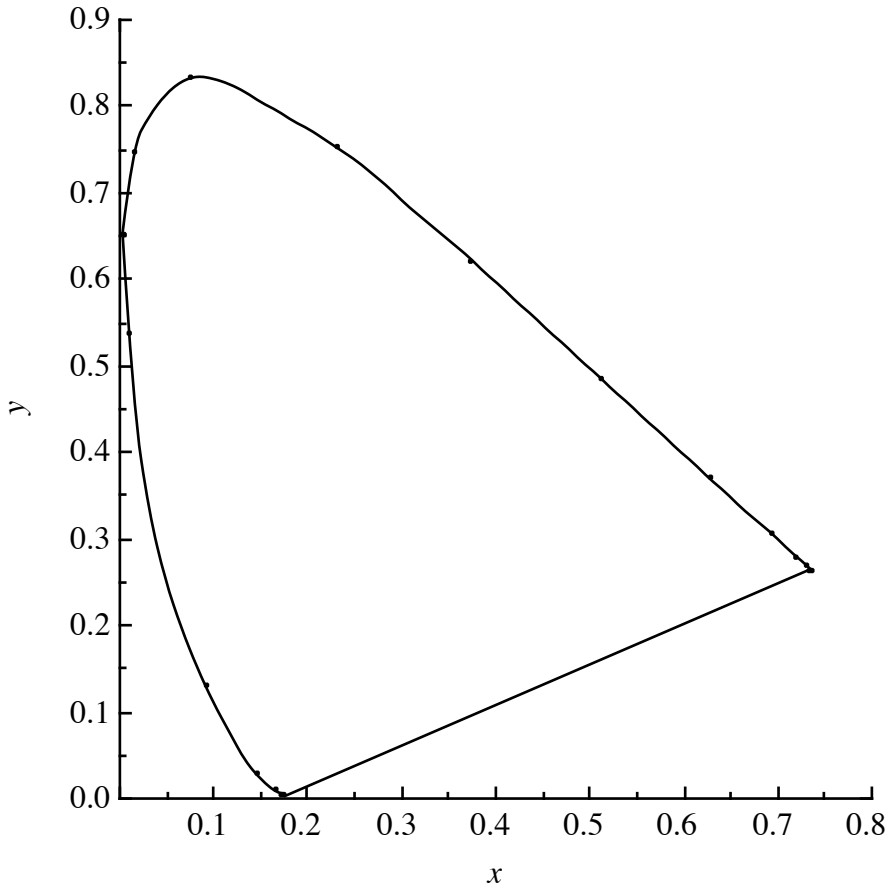


Figure 8 The CIE 1931 chromaticity diagram. Plotted from data given in Wyszecki and Stiles (1982).

The chromaticity coordinates of a reflective display change with the chromaticity of the light used to illuminate it. The change occurs because the amount of light reflected back from an object at each wavelength depends in part on the amount of light falling on it. Therefore, when the chromaticities of objects, or dyes, or paints are specified they are usually given with reference to a standard light source (for a discussion of standardized light sources see Wyszecki and Stiles, 1982). One might expect that the change in the chromaticity coordinates accompanying a change in the light source would result in change in the color appearance of a reflective display. Such changes in light source are actually quite common. As noted above the spectral quality of daylight changes with time of day, atmospheric conditions, season, and location on earth. A large variety of artificial light sources are commercially available, and these can differ considerably in the spectral quality of the light given off. However, these changes do not generally result in large changes in the appearances of objects, because mechanisms within the visual system act to maintain a constant color appearance despite these changes in illumination. Color constancy has generally been shown to be less than perfect (Brainard and Wandell, 1992; Arend and Reeves, 1986). However, it appears to work well enough to prevent confusing changes in the appearance of reflective objects. The visual mechanisms mediating color constancy have been investigated intensely over the past few decades (Maloney and Wandell, 1986 D'Zmura and Lennie, 1986). Selective adaptation within the three cone mechanisms is thought to be one of the major mechanisms mediating color constancy (Worthy and Brill, 1986) much as the change in sensitivity described by Weber's Law plays a role in lightness constancy.

While mechanisms of color constancy work to maintain a constant appearance in reflective displays they actually work against the maintenance of a constant appearance in emissive displays, much as mechanisms of lightness constancy worked against the constant appearance of black and

white emissive displays. Color CRT's take advantage of the fact that human color vision is trichromatic by using only three different phosphors. Each phosphor emits light of a different color when it is stimulated. The light from the three phosphors is mixed together in different proportions to give all other colors including white.

The chromaticity of a color produced on an emissive display does not change with changes in the illumination of the surroundings. Thus the mechanisms of color constancy, activated by changes in the illumination of the surroundings, introduce changes in the appearance of these chromaticities which may be quite noticeable to the observer. Under some conditions these changes in appearance may be large enough to cause some confusion in identifying objects on the basis of color.

20.2.4.1 Factors Affecting Perceived Color Contrast

Much as the perception of achromatic contrast is affected by many factors, color contrast is affected by many factors such as light level, adaptive state, location in the visual field and stimulus size. The spatial frequency approach has also been applied to the detection of color contrast. It is possible to produce grating patterns which vary sinusoidally in color with little or no variation in luminance. The color contrast between the bars of the grating required for detection of the pattern can be measured as a function of the spatial frequency (Kelly, 1974; Noorlander and Koenderink, 1983; Mullen, 1985; Sekiguchi et al., 1993). Typical results for red/green and yellow/blue gratings are shown in Figure 9. Comparison of the results for chromatic patterns with those shown for luminance patterns reveals clear differences. Sensitivity to color contrast is high at low spatial frequencies, but begins to fall off dramatically at rather low spatial frequencies as compared to luminance contrast. Above spatial frequencies of approximately 12 cycles/deg color contrast is not

detectable even at the highest color contrasts producible. Thus chromatic contrast information is limited to fairly low spatial frequencies, or large patterns, as compared to luminance contrast information. Within this range of spatial frequencies the color appearance of the bars of a pattern that is well above threshold is also affected by spatial frequency (Poirson and Wandell, 1993). As the spatial frequency of the pattern is increased the apparent color contrast between the bars is reduced. Thus the detectability of color contrast and the color appearance of stimuli is dramatically affected by stimulus size.

==========

Insert Figure 9 about here

==========

[Insert] a summary paragraph which pulls various aspects of previous discussion with a concrete example.

20.3.0 Four Alternative Approaches to the Problem of Display Design

The previous sections have discussed physiological, perceptual, and technological considerations in designing visual displays. This has been the traditional focus for human factors research: to design displays that are legible. For example, the knowledge that a user will be seated a particular distance from a particular type of display under a particular set of ambient lighting conditions can be used to determine the appropriate size and luminance contrast that will be necessary for the characters to be seen. Thus, the previous considerations provide us with an understanding of the baseline conditions of display design that must be met (are necessary) for an individual to use a display.

Although these considerations are necessary for the design of effective displays, they are not

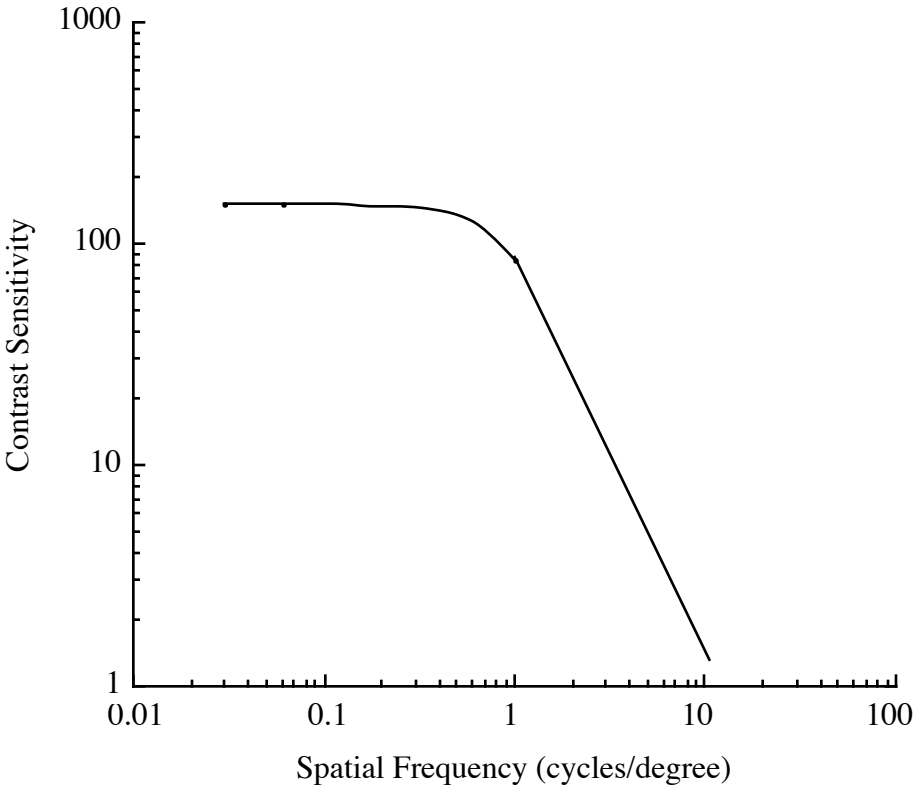


Figure 9 A typical plot of contrast sensitivity for isoluminant chromatic gratings. Based on data given in Mullen (1985).

sufficient. Compliance with these considerations will make the data required to complete domain tasks available, but may not provide the information necessary to support an observer in decision making and action. Woods (1991) makes an important distinction between design for "data availability" and design for "information extraction." Designs that consider only data availability often impose unnecessary burdens on the user: to collect relevant data, to maintain these data in memory, and to integrate these data mentally to arrive at a decision. These mental activities require extensive knowledge, tax limited cognitive resources (attention, short-term memory) and therefore increase the probability of poor decision-making and errors.

Our discussion of design for information extraction will begin with a consideration of four broadly defined approaches to display design. Each of these approaches are complementary in the sense that they approach the display design problem from different conceptual perspectives (i.e., graphical-arts, psychophysical, attention-based, and problem solving / decision making).

20.3.1 Aesthetic Approach to Statistical Graphics

Tufte (1983; 1990) reviews the design of displays from an aesthetic, graphic arts perspective. Tufte (1983) describes principles of design for "data graphics" or "statistical graphics" which are expressly designed to present quantitative data. One principle is the "data-ink ratio": a measurement of the relative salience of data vs. non-data elements in a graph. It is computed by determining the amount of ink that is used to convey the data and dividing this number by the total amount of ink that is used in the graphic. A higher data-ink ratio (a maximum of 1.0) represents the more effective presentation of information. A second measure of graphical efficiency is "data density." Data density is computed by determining the number of data points represented in the graphic and dividing this number by the total area of the graphic. The higher the data density the more effective

the graphic. Other principles include eliminating graphical elements that interact (e.g., moire vibration), eliminating irrelevant graphical structures (e.g., containers and decorations), and aesthetics (e.g., effective labels, proportion and scale).

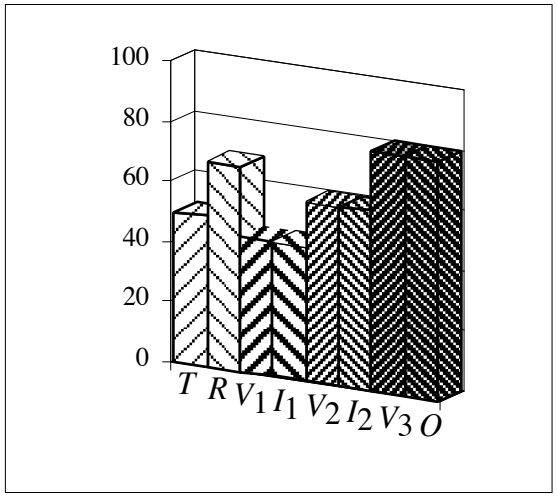
The two versions of a statistical graphic that are shown in Figure 10a and 10b illustrate several of Tufte's principles. The version in Figure 10a is poorly designed, while the version in Figure 10b is more effectively designed. In Figure 10b the irrelevant data container (the box) that surrounds the graph in Figure 10a has been eliminated. In addition, several other non-data graphical structures have been removed (grid lines). In fact, these grid lines are made conspicuous by their absence in Figure 10b. Together, these manipulations produce both a higher data-ink ratio and a higher data density for the version in Figure 10b. In Figure 10a the "striped" patterns on the bar graphs produce an unsettling Moire vibration and have been replaced in Figure 10b with gray-scale patterns. In addition, the bar graphs in Figure 10b have been visually segregated by spatial separation. Finally, the three dimensional perspective in Figure 10a complicates visual comparisons and has been removed in Figure 10b.

==========

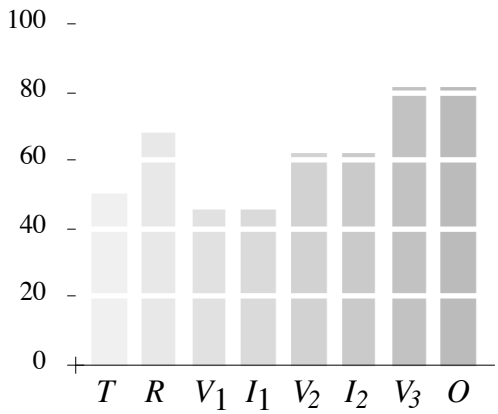
Insert Figure 10 about here

==========

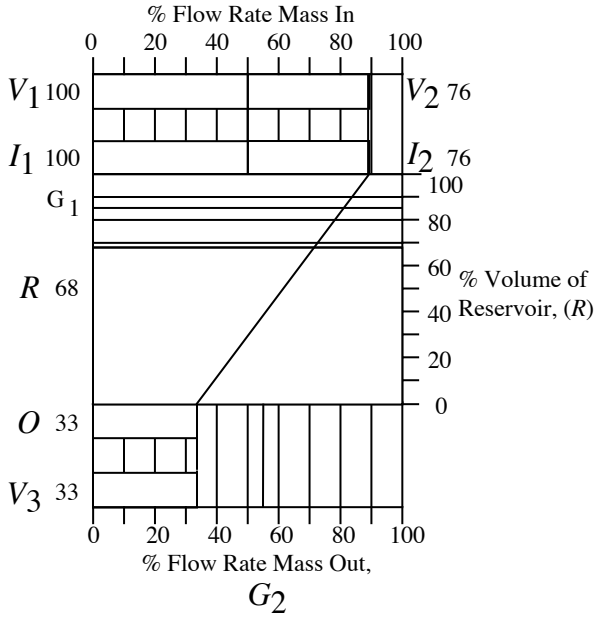
Tufte (1990) broadens the scope of these principles and techniques by considering non-quantitative displays as well. Topics that are discussed include micro/macro designs (the integration of global and local visual information), layering and separation (the visual stratification of different categories of information), small multiples (repetitive graphs that show the relationship between variables across time, or across a series of variables), color (appropriate and inappropriate use of), and narratives of space and time (graphics that preserve or illustrate spatial relations or relation-



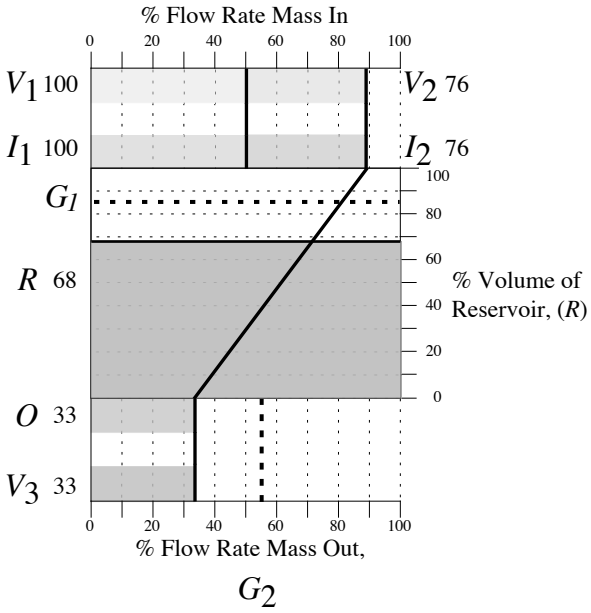
(*a*).



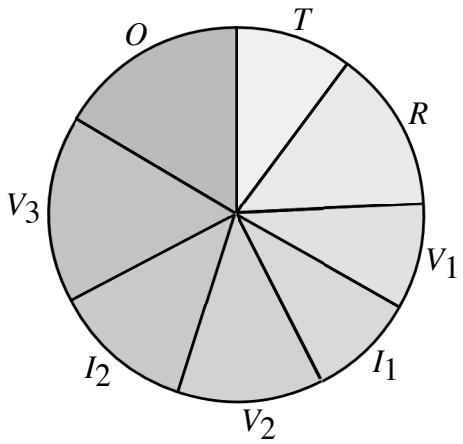
(*b*).



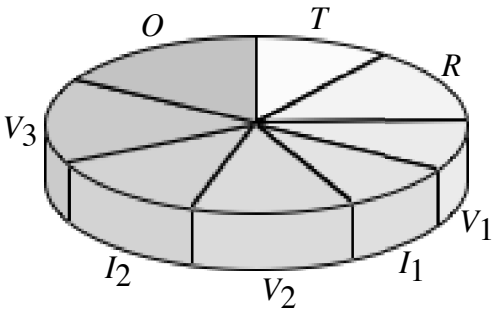
(*c*).



(*d*).



(*e*).



(*f*).

Figure 10. Six alternative mappings. Figure 10a and10 b represents alternative versions of a separable (bar graph) format that are less effective (10a) and more effective (10b) mappings. Similarly, Figure 10c and 10d represents alternative versions of a configural display format that are less (10c) and more effective (10d), primarily due to layering and separation. Figure 10e and 10f represents the least effective mappings.

ships over time). The following quotations summarize many of the key principles.

- It is not how much information there is, but rather, how effectively it is arranged (p. 50).
- Clutter and confusion are failures of design, not attributes of information (p. 51).
- Detail cumulates into larger coherent structures... Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, add detail (p. 37).
- Micro / macro designs enforce both local and global comparisons and, at the same time, avoid the disruption of context switching. All told, exactly what is needed for reasoning about information (p. 50).
- Among the most powerful devices for reducing noise and enriching the content of displays is the technique of layering and separation, visually stratifying various aspects of the data...What matters -- inevitably, unrelentingly -- is the proper relationship among information layers. These visual relationships must be in relevant proportion and in harmony to the substance of the ideas, evidence, and data displayed (pp. 53-54).

This final principal, layering and separation, is graphically illustrated in Figures 10c and 10d. These two versions of the same display vary widely in terms of the visual stratification of the information that they contain. In Figure 10c all of the graphical elements are at the same level of visual prominence; in Figure 10d there are at least three levels of visual prominence. The lowest layer of visual prominence is associated with the non-data elements of the display. The various display grids have thinner, dotted lines and their labels have been reduced in size and made thinner. The medium layer of perceptual salience is associated with the individual variables. The graphical forms that represent each variable have been gray-scale coded, which contributes to separating these data from the non-data elements. Similarly, the lines representing the system goals (G1 and G2) have been made bolder and dashed. In addition, the labels and digital values that correspond to the individual variables are larger and bolder than their non-data counterparts. Finally, the high-

est level of visual prominence has been reserved for those graphical elements which represent higher-level system properties (e.g., the bold lines that connect the bar graphs). The visual stratification could have been further enhanced through the use of color. The techniques of layering and separation will facilitate an observer's ability to locate and extract information.

To summarize, Tufte (1983; 1990) addresses the problem of presenting three-dimensional, multivariate data on flat, two-dimensional surfaces (primarily focusing on static, printed material) very admirably. He attacks the problem from a largely aesthetic perspective and provides numerous examples of both good and bad display design that clearly illustrate the associated design principles. Although there are aspects of dynamic display design for complex domains that are not considered, the principles can be generalized.

20.3.2 Psychophysical approach to statistical graphics

Cleveland and his colleagues have also developed principles for the design of statistical graphics. However, in contrast to the aesthetic conceptual perspective of Tufte, Cleveland has used a psychophysical approach. As an introduction consider the following quotation (Cleveland, 1985, p. 229):

> When a graph is constructed, quantitative and categorical information is encoded by symbols, geometry, and color. Graphical perception is the visual decoding of this encoded information. Graphical perception is the vital link, the raison d'etre, of the graph. No matter how intelligent the choice of information, no matter how ingenious the encoding of the information, and no matter how technologically impressive the production, a graph is a failure if the visual decoding fails. To have a scientific basis for graphing data, graphical perception must be understood. Informed decisions about how to encode data must be based on knowledge of the visual decoding process.

In their efforts to understand graphical perception Cleveland and his colleagues have consid-

ered how psychophysical laws (e.g., Weber's law, Stevens' law) are relevant to the design of graphic displays. For example, psychophysical studies using magnitude estimation have found that judgments of length are less biased than judgments of area or volume. Therefore, visual decoding should be more effective if data has been encoded into a format that requires length discriminations, as opposed to area or volume discriminations. Cleveland and his colleagues have tested this, and similar intuitions, empirically. Their experimental approach was to take the same quantitative information, to provide alternative encodings of this quantitative information (graphs which required different "elementary graphical-perception tasks"), and to test observers' ability to extract the information.

The results of these experiments provided a rank-ordering of performance on basic graphical perception tasks: position along a common scale, position along identical, nonaligned scales, length, angle / slope, area, volume, and color hue / color saturation / density (ordered from best to worst performance, Cleveland, 1985, p. 254). Guidelines for display design were developed based on these rankings. Specifically, graphical encodings should be chosen that require the highest ranking graphical perceptual task of the observer during the visual decoding process. For example, consider the three graphs illustrated in Figure 10b, 10e, and 10f. For decoding information contained in the Figure 10b an observer is required to judge position along a common scale (in this case, the vertical extent of the various bar graphs). For Figure 10e the observer is required to judge angles and/or area. Finally, to decode the information in Figure 10f the observer is required to judge volume (note that because of the three dimensional representation angles and area are no longer valid cues). According to the rankings, Cleveland and his colleagues would therefore predict that performance would be best with the bar chart, intermediate with the pie chart, and worst with the three-dimensional pie chart.

20.3.3 Attention-based Approach to display design

A third perspective on display design is to consider the problem in terms of visual attention and object perception. From this conceptual perspective designers have a number of interface resources at their disposal for encoding information in graphical displays (e.g., chromatic contrast, luminance contrast, the integration of individual variables into geometrical objects, and animation). A great deal of basic research has attempted to identify the factors that control the distribution of attention to visual stimuli. The results have important theoretical and practical implications for display design.

Understanding these implications requires a brief consideration of the continuum of attention demands that operators might face in complex, dynamic domains. At one end of the attention continuum are tasks that require selective responses to specific elements in the display ("focused" tasks). This might refer to a response contingent on the height of a single bar in a bar graph or on the position of a pointer on a radial display. At the opposite end of this continuum are tasks that require the distribution of attention across many features that must be considered together in order to choose an appropriate response ("integration" tasks). For example, the response might be contingent on the relative position of numerous bars within a bar graph. Thus, tasks can be characterized in terms of the relative demands for selective attention to respond to specific features with specific actions and distributed or divided attention in which multiple display elements must be considered together in order to choose the appropriate actions.

Attention-based approaches to display design have examined how the design of visual representations can help to meet the cognitive load posed by this continuum of attention demands. Garner (Garner, 1970, 1974; Garner and Felfoldy, 1970) and Pomerantz (Pomerantz, 1986; Pomerantz

and Pristach, 1989; Pomerantz, Sager, and Stoever, 1977) have used the speeded classification task to examine the dimensional structure of stimuli. Carswell and Wickens (1987) have generalized these results by investigating perceptual dimensions that are representative of those found in visual displays. Three qualitatively different relationships between stimulus dimensions have been proposed: "separable," "integral," and "configural" (Pomerantz, 1986).

Separable dimensions. A separable relationship is defined by a lack of interaction among stimulus dimensions. Each dimension retains its unique perceptual identity within the context of the other dimension. Observers can selectively attend to an individual dimension and ignore variations in the irrelevant dimension. On the other hand, no new properties emerge as a result of the interaction among dimensions. Thus, performance suffers when both dimensions must be considered to make a discrimination. This pattern of results suggests that separable dimensions are processed independently. An example of separable dimensions are color and shape: the perception of color does not influence the perception of shape, and vice versa.

Integral dimensions. An integral relationship is defined by a strong interaction among dimensions such that the unique perceptual identities of individual dimensions are lost. Integral stimulus dimensions are processed in a highly inter-dependent fashion: a change in one dimension necessarily produces changes in the second dimension. In their discussion of two integral stimulus dimensions Garner and Felfoldy (1970, p. 237) state that "in order for one dimension to exist, a level on the other must be specified." As a result of this highly interdependent processing a redundancy gain occurs. However, focusing attention on the individual stimulus dimensions becomes very difficult, and performance suffers when attention to one (selective attention) or both (divided attention) dimensions are required. An example of an integral stimulus is perceived color: it is a

function of both hue and brightness.

Configural dimensions. A configural relationship refers to an intermediate level of interaction between perceptual dimensions. Each dimension maintains its unique perceptual identity, but new properties are also created as a consequence of the interaction between them. These properties have been referred to as "emergent features." Using parentheses as our graphic elements will allow us to demonstrate several examples of emergent features. Depending upon the orientation a pair of parentheses can have the emergent features of vertical symmetry, () and)(, or parallelism,)) and ((. Pomerantz and Pristach (1989, p. 636) state that "Emergent features may be global (i.e., not localized to any particular position within the figure), such as symmetry or closure, or they may be local, such as vertices that result from intersections of line segments." There are two significant aspects of performance with configural dimensions. First, relative to integral and separable stimulus dimensions there is a smaller divided attention cost, suggesting that performance can be enhanced when both dimensions must be considered to make a discrimination. The second noteworthy aspect of this pattern of results is that there is an apparent failure of selective attention (see Bennett and Flach, 1992, for a further discussion of why this failure may be apparent, and not inherent).

20.3.3.1 Proximity Compatibility Principle

Wickens and his colleagues (e.g., Wickens and Carswell, in press) have applied the results of the visual attention research to the problem of display design. Their principle of proximity compatibility emphasizes the relationship between task demands and the graphical form of a display. Perceptual proximity (display proximity) refers to the perceptual similarity between information sources in a display. Perceptual proximity can be defined along several dimensions including: 1)

spatial proximity (e.g., physical distance -- near or far), 2) chromatic proximity (e.g., the same or different colors), 3) physical dimensions (e.g., information is encoded using the same or different physical dimensions -- length vs. volume), 4) perceptual code (e.g., digital vs. analog), and 5) geometric form (e.g., object vs. separate displays). For example, when individual variables are mapped into a closed geometric form the display is high in display proximity; when each variable has its own unique representation (e.g., a bar graph) the display is low in proximity.

Processing proximity (mental proximity) refers to the continuum of attentional demands. That is, to the extent to which information from the various sources in a display must be (or need not be) considered together to accomplish a particular task. There are three major categories of processing proximity: integrative processing, non-integrative processing, and independent processing. Information from different sources must be explicitly combined in integrative processing, and this represents a high level of processing proximity. Integrative processing includes both computational processing (involving numerical operations) and boolean processing (involving logical operations). Non-integrative processing represents an intermediate level of processing proximity and involves "some other features of similarity instead of (or in addition to) their need for combination" (Wickens and Carswell, p. ?). Examples include 1) metric similarity (similarity of units), 2) statistical similarity (extent of covariation), 3) functional similarity (semantic relatedness), 4) processing similarity (similarity of computational procedures), and 5) temporal similarity (temporal proximity). Finally, independent processing refers to the situation where different information sources need not be considered together (in fact, one information source is independent from another).

Briefly stated, the principle of proximity compatibility maintains that the <u>display proximity</u>

should match the task proximity. Performance on integrated tasks (high mental proximity) is predicted to be facilitated by displays that have high perceptual proximity (e.g., object display). Similarly, performance on focused tasks (low mental proximity) is predicted to be facilitated by displays that have low perceptual proximity (e.g., bar graph displays).

20.3.3.2 Implications for display design

Researchers continue to investigate the potential trade-offs between display type (object vs. separate) and task type (integrated vs. focused). Initially, a straight-forward trade-off was predicted: object displays would produce superior performance for integration tasks, while separable displays would produce superior performance for focused tasks. In general, laboratory research comparing performance differences between object and separate displays when integration tasks must be completed has revealed significant advantages for object displays (Bennett and Flach, 1992). However, there is a general consensus that these performance advantages are not attributable to objectness, per se (Bennett and Flach, 1992; Bennett, Toms, and Woods, 1993; Buttigieg and Sanderson, 1991; Sanderson et al., 1989; Wickens and Carswell, in press). Instead, the quality of performance at integration tasks is dependent upon the quality of the mapping between the emergent features produced by a display and the inherent data relationships that exist in the domain (this point will be discussed at length in subsequent sections).

There is much less consensus on the second major prediction regarding the potential costs for configural displays (relative to separable displays) when individual variables must be considered. We believe that a single display may support performance at both integration and focused attention tasks (Bennett and Flach, 1992). The attention and object perception literature (in particular, the principle of configurality) leaves open the possibility that a single geometric display may be de-

signed to support performance for both distributed and focused attention tasks. One way to consider objects is as a set of hierarchical features (including elemental features, configural features, and global features) that vary in their relative salience. For example, Treisman (1986) observed that "if an object is complex, the perceptual description we form may be hierarchically structured, with global entities defined by subordinate elements and subordinate elements related to each other by the global description" (p. 35.54). Observers may focus attention at various levels in the hierarchy at their discretion, and in particular, there may be no inherent cost associated with focussing attention on elemental features. From a practical standpoint, any potential costs associated with low-level data can be eliminated outright by annotating the graphical representation with digital information.

20.3.4 Problem Solving And Decision Making Approach to Display Design

The fourth perspective on display design that will be discussed is problem solving and decision making. Recently, there has been an increased appreciation for the creativity and insight that experts bring to human-machine systems. Under normal operating conditions an individual is perhaps best characterized as a decision maker. Depending on the perceived outcomes associated with different courses of action, the amount of evidence that a decision maker requires to choose a particular option will vary. In models of decision making, this is called a decision criterion. Under abnormal or unanticipated operating conditions an individual is most appropriately characterized as a creative problem solver. The cause of the abnormality must be diagnosed, and steps must be taken to correct the abnormality (i.e., an appropriate course of action must be determined). This involves monitoring and controlling system resources, selecting between alternatives, revising diagnoses and goals, determining the validity of data, overriding automatic processes, and coordinating the activities of other individuals. Thus, the literature on reasoning, problem solving, and

decision making has important insights for display design.

There is a vast literature on problem solving, ranging from the seminal work of the Gestaltists (e.g., Wertheimer, 1959), the paradigmatic contributions of Newell and Simon (1972), to contemporary approaches. For the Gestalt psychologists perception and cognition (more specifically, problem solving) were intimately intertwined. The key to successful problem solving was viewed as the formation of an appropriate gestalt, or representation, that revealed the "structural truths" of a problem. For example, Wertheimer (1959, p. 235) states that "Thinking consists in envisaging, realizing structural features and structural requirements..." The importance of a representation is still a key consideration today; it is probably not an overstatement to conclude that the primary lesson to be learned from the problem solving literature is that the representation of a problem has a profound influence on the ease or difficulty of its solution.

Historically, decision research has focused on developing models that describe the generation of multiple alternatives (potentially all alternatives), the evaluation (ranking) of these alternatives, and the selection of the most appropriate alternative. By and large, perception was ignored. In contrast, recent developments in decision research, stimulated by research on naturalistic decision making (e.g., Klein, Orasanu, and Zsambok, 1993) has begun to give more consideration to the generation of alternatives in the context of dynamic demands for action. Experts are viewed as generating and evaluating a few "good" alternatives. The emphasis is on recognition (e.g., how is this problem similar, or dissimilar, to problems that I have encountered before?). As a result, perception plays a dominant role. This change in emphasis has increased awareness of perceptual processes and dynamic action constraints in decision making.

These trends have, either directly or indirectly, led researchers in interface design to focus on

the representation problem. Perhaps the first explicit realization of the power of graphic displays to facilitate understanding was the STEAMER project (Hollan et al., 1987), an interactive inspectable, training system. STEAMER provided alternative conceptual perspectives - "conceptual fidelity" of a propulsion engineering system through the use of analogical representations. In addition, the current approach to the design of human computer interfaces (direct manipulation -- Shneiderman, 1986, 1993; Hutchins, Hollan, & Norman, 1986) can be viewed as an outgrowth of this general approach. More recently scientific visualization (the role of diagrams and representation in discovery and invention) is being vigorously investigated (reference). Thus, the challenge for display design from this perspective is to provide appropriate representations that support humans in their problem solving endeavors.

20.4.0 Display Design: Representation Aiding

It should be noted that in the aesthetic, psychophysical, and attention-based approaches there is little consideration to a domain or problem behind the display. It was not necessary for us to describe the "problem" behind the displays shown in Figure 10. However, the correspondence between the visual structure in a representation and the constraints in a problem is fundamental to the problem solving and decision making approaches. Recently, a number of research groups have recognized that effective interfaces depend on both the mapping from human to display (the coherence problem) and the mapping from display to a work domain or problem space (the correspondence problem). Terms used to articulate this recognition include direct perception (Moray, Lee, Vicente, Jones, and Rasmussen 1994) ecological interface design (Rasmussen and Vicente, 1989) representational design (Woods, 1991), or semantic mapping (Bennett and Flach, 1992). Woods and Roth (1988) have illustrated the two components of the interface problem in terms of

a "cognitive triad" that we illustrate in Figure 11. The three components of the cognitive system triad are 1) the cognitive demands produced by the domain of interest, 2) the resources of the cognitive agent(s) that meet those demands, and 3) the representation of the domain through which the agent experiences and interacts with the domain (the interface). Thus, the focus of our approach is not on information processing characteristics, graphical forms, events, trajectories, tasks, or procedures per se. Instead, the focus is on the interactive and mutually constraining relationships between the individual, the interface, and the domain that are labeled coherence and correspondence. The overall level of human machine system performance is determined by the quality of these relationships.

==========

Insert Figure 11 about here

==========

20.4.1 The Correspondence Problem: The Semantics Of Work

Correspondence refers to the issue of content --- what information should be present in the interface in order to meet the cognitive demands of the work domain? Correspondence is defined neither by the domain itself, nor the interface itself: it is a property that arises from the interaction of the two. Thus, in Figure 11 correspondence is represented by the labelled arrow that connects the domain and the interface. One convenient way to conceptualize correspondence is as the quality of the mapping between the interface and the work space, where these mappings can vary in terms of the degree of specificity (consistency, invariance, or correspondence). As we will demonstrate, within this mapping there can be a one-to-one correspondence, a many-to-one, a one-to-many, or a many-to-many mapping between the information that exists in the interface and the structure within the work space.

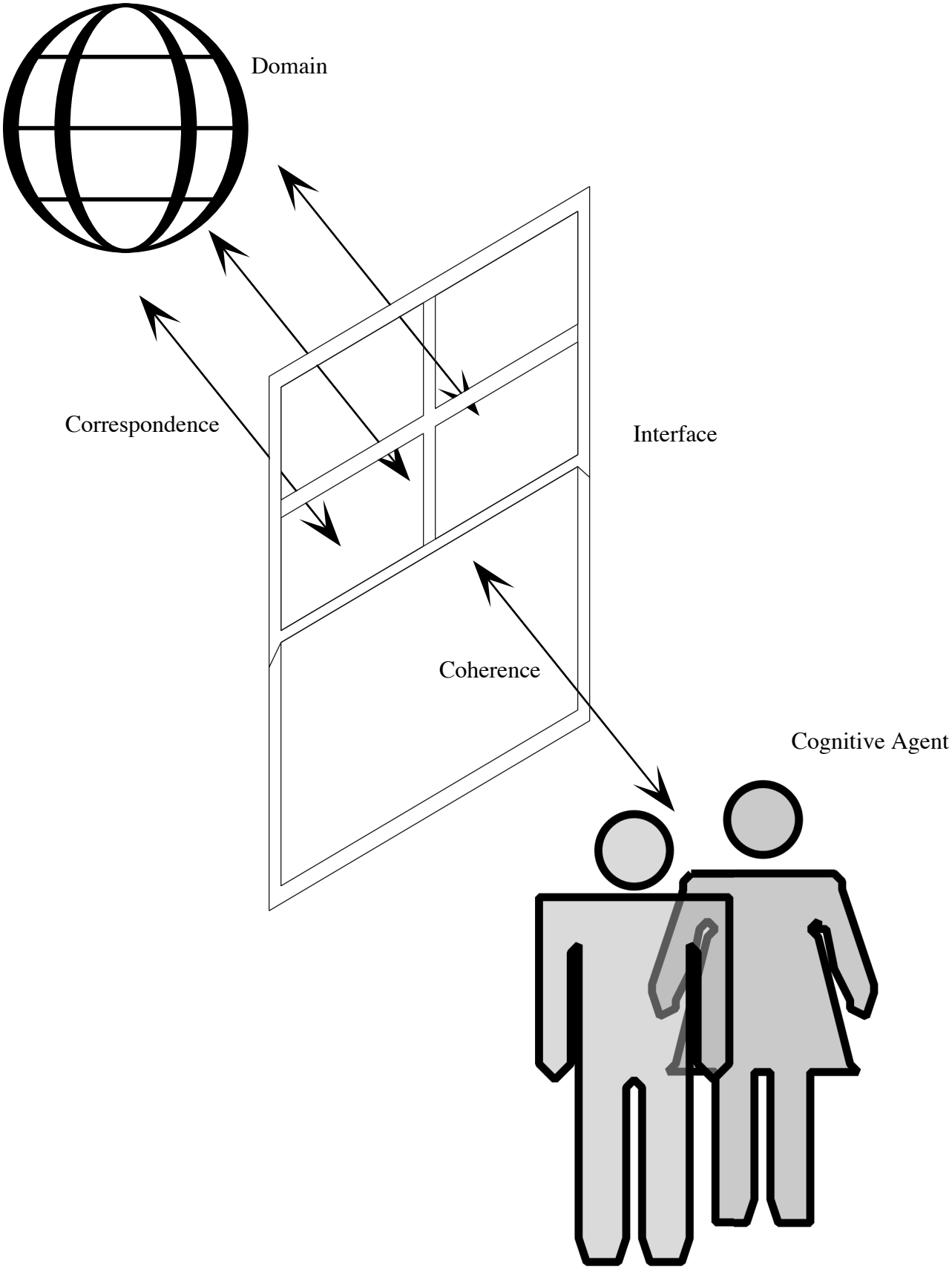


Figure 11. The "cognitive triad": A cognitive systems engineering perspective. Any domain produces cognitive demands that must be met by the cognitive agents interacting with (or controlling) the domain. The cognitive agent possesses cognitive resources that must be used to meet these demands. The interface is the medium (or representation) through which the cognitive agent views and controls the domain. The effectiveness of an interface is determined by both correspondence and coherence.

20.4.1.1 Rasmussen's Abstraction Hierarchy

Addressing the issue of correspondence requires a deep understanding and explicit description of the "semantics" of a work domain. Rasmussen's abstraction hierarchy (1986) is a theoretical framework for describing domain semantics in terms of a nested hierarchy of functional constraints (including goals, physical laws, regulations, organizational/structural constraints, equipment constraints, and temporal/spatial constraints). One way to think about the abstraction hierarchy is that it provides structured categories of information (i.e., the alternative conceptual perspectives) that an individual must consider in the course of accomplishing system goals. Consider the following passage from Rasmussen (1986, p. 21):

> During emergency and major disturbances, an important control decision is to set up priorities by selecting the level of abstraction at which the task should be initially considered. In general, the highest priority will be related to the highest level of abstraction. First, judge overall consequences of the disturbances for the system function and safety in order to see whether the mode of operation should be switched to a safer state (e.g., standby or emergency shutdown). Next, consider whether the situation can be counteracted by reconfiguration to use alternative functions and resources. This is a judgment at a lower level of function and equipment. Finally, the root cause of the disturbance is sought to determine how it can be corrected. This involves a search at the level of physical functioning of parts and components. Generally, this search for the physical disturbance is of lowest priority (in aviation, keep flying-- don’t look for the lost light bulb!).

Thus, in complex domains, situation awareness requires the operator to understand the process at different levels of abstraction. Further, the operator must be able to understand constraints at one level of abstraction in terms of constraints at other levels. The correspondence question asks whether the hierarchy of constraints that define a work domain are reflected in the interface.

20.4.2 The Coherence Problem: The Syntax Of Form

Coherence refers to the mapping from the representation to the human perceiver. Here the fo-

cus is on the visual properties of the representation. What distinctions within the representation are discriminable to the human operator? How do the graphical elements fit together or coalesce within the representation? Is each element distinct or separable? Are the elements absorbed within an integral whole, thus losing their individual distinctness? Or do the elements combine to produce configural or global properties? Are some elements or properties of the representation more or less salient than other elements or properties?

In general, coherence addresses the question of how the various elements within a representation compete for attentional and cognitive resources. Just as work domains can be characterized in terms of a nested hierarchy of constraints, so to, can complex visual representations be perceived as a hierarchy of nested structures, with local elements combining to produce more global patterns or symmetries.

20.4.3 The Mapping Problem

In human-machine systems a display is a representation of an underlying domain, and the user's tasks are defined by that domain, rather than by the visual characteristics of the display itself. Thus, whether a display will be effective or not is be determined by both correspondence and coherence. More specifically, the effectiveness of the display is determined by the quality of the mapping between the constraints that exist in the domain and the constraints that exist in the display. The display constraints are defined by the spatio-temporal structure (the visual appearance of the display over time) that results from the particular representation chosen.

Three rather fuzzy distinctions might be useful when thinking about the types of representations that might be used to accomplish the mapping of domain constraints to constraints within the interface --- analogical, configural, and metaphorical. Analogical representations might be consid-

ered when the constraints of the work domain are fundamentally spatial. For example, STEAMER used an analogical representation of the spatial layout of the feedwater system to show the connections among component processes. Also, the standard flight display for representing pitch and roll (attitude) is an analog to the spatial relations between the aircraft and the horizon. In general, where the domain constraints themselves are naturally spatial, designers should consider whether the interface might provide a direct analog of these constraints.

Configural representations use geometric relations to represent constraints that are not spatial. A simple example is using an axis in a graph to represent time. In configural representations the geometrical display constraints will generally take the form of symmetries --- equality (e.g., length, angle, area), parallel lines, colinearity, or reflection. In addition, Gestalt properties of closure and good form are useful. These display constraints will produce the emergent features that were discussed in Section ?. The core problem in implementing effective configural displays is to provide visual representations that are perceived as accurate reflections of the abstract domain constraints: Are the critical domain constraints appropriately reflected in the geometrical constraints in the display? Are breaks in the domain constraints (e.g., abnormal or emergency conditions) reflected by breaks in the geometrical constraints (e.g., emergent features such as non-equality, non-parallelism, non-closure, bad form)? Only when this occurs will the cognitive agent be able to obtain meaning about the underlying domain in an effective fashion. One source of ideas for configural displays is the graphical representations that engineers use to make design decisions. For example, Beltracchi (1987; 1989; See also Lindsay & Staffon, 1988; Moray, Lee, Vicente, Jones, and Rasmussen 1994; Rasmussen, Pejtersen, & Goodstein, 1994) has designed a configural display for controlling the process of steam generation in nuclear power plants based on the Temperature/Entropy graphic used to evaluate thermodynamic engines (Rankine Cycle Display).

Metaphorical representations use spatial or symbolic relations from other, more familiar, work domains as metaphors; the goal is to enhance the transfer of skills from one domain to another. Perhaps the most obvious example is the "desk top" metaphor that is used in personal computer systems. Another example is the BookHouse metaphor, developed by Goodstein & Pejtersen (Goodstein & Pejtersen, 1989; Pejtersen, 1992) to facilitate library information retrieval. Rasmussen, Pejtersen, and Goodstein (1994, pp. 289-291) describe the metaphor and its justification:

> The use of the BookHouse metaphor serves to give an invariant structure to the knowledge base ... Since no overall goals or priorities can be embedded in the system, but depend on the particular user, a global structure of the knowledge base reflects subsets relevant to the categories of users having different needs and represented by different rooms in the house ... This gives a structure for the navigation that is easily learned and remembered by the user ... The user "walks" through rooms with different arrangements of books and people ... It gives a familiar context for the identification of tools to use for the operational actions to be taken. It exploits the flexible display capabilities of computers to relate both information in and about the data base, as well as the various means for communicating with the data base to a location in a virtual space ... This approach supports the user's memory of where in the BookHouse the various options and information items are located. It facilitates the navigation of the user so that items can be remembered in given physical locations that one can then retraverse in order to retrieve a given item and/or freely browse in order to gain an overview.

Whether analogical, configural, metaphorical or some combined representation are used the most important key to successful design in these domains is the mapping. Salience in the display must reflect importance in terms of the work domain. For analogical displays the spatial analogs must scale appropriately to the real task constraints. For configural displays the geometric symmetries must correspond to higher-order constraints on the process. For metaphorical displays, the intuitions and skills elicited by the representational domain must map appropriately to the target domain.

20.4.4 A simple example

In this section, we will attempt to illustrate the rationale for constructing interfaces that address both the correspondence and coherence problem with a simple example. The example will be based on a simple process control task. Various types of representations will be considered. The representations are chosen to represent the continuum of visual forms from separable, through configural, to integral geometries.

The process is a generic one that might be found in process control, and it is represented graphically in the lower portion of Figure 12. There is a reservoir (or tank, represented by the large rectangle in the middle of the figure) that is filled with a fluid (for example, coolant). The volume, or level, of the reservoir (R) is represented by the filled portion of the rectangle. Fluid can enter the reservoir through the two pipes and valves located above the reservoir; fluid can leave the reservoir through the pipe and valve located below the reservoir. We will categorize the information in this simple process using a simple distinction in which the term "low-level" data refers to local constraints or elemental state variables that might be measured by a specific sensor. The term "higher-level properties" will be used to refer to more global constraints.that reflect relations or interactions among multiple variables.

==========

Insert Figure 12 about here

==========

Low-level data (process variables). There are two goals associated with this simple process. First, there is a goal (G_1) associated with R, the level of the reservoir. The reservoir should be maintained at a relatively high level to ensure that increases in demand (the required output flow rate, (O) can be met. The second goal (G_2) refers to the output flow rate that must be maintained in order to meet an external demand. These goals are achieved and maintained by adjusting three

Low-Level Data (process variables)			High-Level Properties (process constraints)		
T	=	time	K_1 = I_1 - V_1		Relation between commanded flow (V) and actual flow (I or O)
V_1	=	setting for valve 1	K_2 = I_2 - V_2		
V_2	=	setting for valve 2	K_3 = O - V_3		
V_3	=	setting for valve 3			
I_1	=	flow rate through valve 1			
I_2	=	flow rate through valve 2			
O	=	flow rate through valve 3			
R	=	volume of reservoir	K_4 =	$\Delta R = (I_1 + I_2) - O$	Relation between reservoir volume (R), mass in ($I_1 + I_2$), and mass out (O)
G_1	=	volume goal	K_5 = R - G_1		Relation between actual states (R, O) and goal states (G_1, G_2)
G_2	=	output goal (demand)	K_6 = O - G_2		

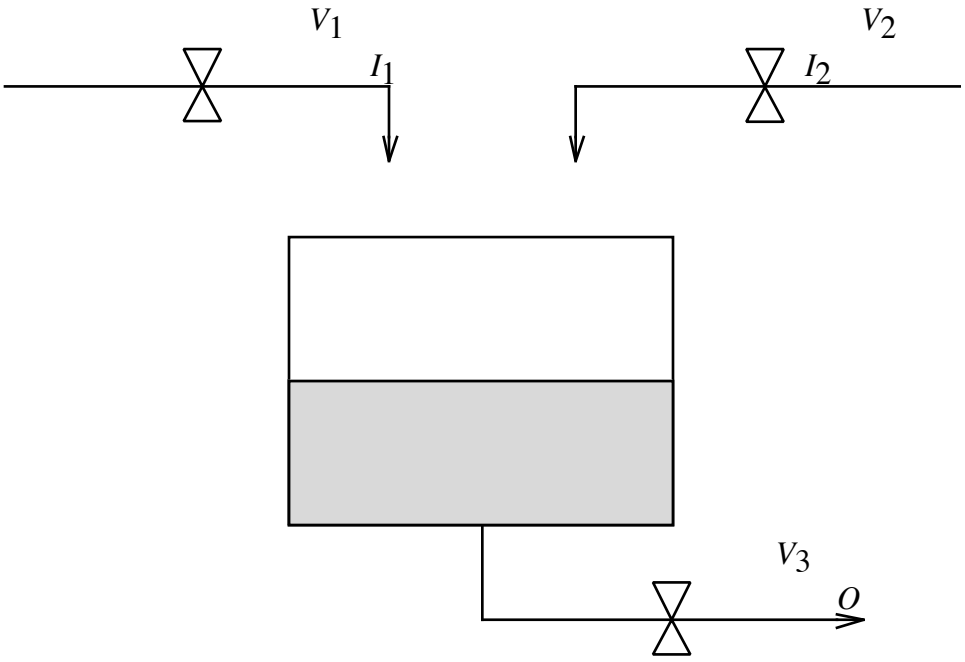


Figure 12. A simple domain from process control that has a reservoir for storing mass, two input streams that increase the volume of mass in the reservoir, and a single output stream that decreases the volume. The low level data (the measured domain variables), the high-level properties (constraints that arise from the interaction of these variables and the physical design) and the domain goals (requirements that must be met for the system to be functioning properly) are listed.

valves (V_1, V_2, & V_3) that regulate flow through the system (I_1, I_2, & O). Thus, this simple process is associated with a number of process variables that can be measured directly: these low-level data are listed in the upper, left-hand portion of Figure 12 (V_1, V_2, V_3, I_1, I_2, O, G_1, G_2, and R).

High-level properties (process constraints). In addition, there are relationships between these process variables that must be considered when controlling the process. The most important high-level properties are goal-related: Does the actual reservoir volume level (R) match the goal of the system (G_1)? Does the actual system flow rate (O) match the flow rate that is required (G_2)? Even for this simple process some of the constraints or (high-level properties) are fairly complex. For example, an important property of the system is mass balance. The mass balance is determined by comparing the mass leaving the reservoir (O, the output flow rate) to mass entering the reservoir (the input flow rates of I_1 & I_2). This relationship determines the direction and the rate of change for the volume inside the reservoir (ΔR). For example, if mass in and mass out are equal then mass is balanced, ΔR will equal 0.00, and R will remain constant.

Controlling even this simple process will depend upon a consideration of both high-level properties and low-level data. As the previous example indicates, decisions about process goals (e.g., maintaining a sufficient level of reservoir volume) generally require consideration of relationships between variables (is there a net inflow, net outflow, or is mass balanced?), as well as the values of the individual variables themselves (what is the current reservoir volume?).

20.4.5 An abstraction hierarchy analysis applied to the simple process

The constraints of the simple process in Figure 12 will be characterized in terms of the abstraction hierarchy. Typically the hierarchy has five separate levels of description, ranging from the

physical form of a domain to the higher-level purposes it serves. The highest level of constraints refers to the "functional purpose" or design goals for the system. For our simple process these are constraints K_5 and K_6. For example, consider the relationship between R and G1. When the actual reservoir volume (R) equals the goal reservoir volume (G_1) the difference between these two values will assume a constant value (0.00). This process constraint is represented by the equation associated with the higher-level property K_5 in Figure 12. For an actual work domain, the associated values (costs & benefits) underlying these particular goals might be considered. The "abstract functions" or physical laws that govern system behavior are another important source of constraints. In our example, K_4 reflects the law of conservation of mass. Change of mass in the reservoir (ΔR) should be determined by the difference between the residual mass in (I1 + I2) and the mass out (O). K_1, K_2, and K_3 represent similar constraints associated with the mass flow. Flow is proportional to valve setting (this assumes a constant pressure head). Further constraints arise as a result of the generalized function (sources, storage, sink). In our example, there are two sources, a single store, and a single sink. Also, the physical processes behind each "general function" represents another source of constraint (in this case, two feedwater streams, a single output stream, a reservoir for storage). Finally, the level of "physical form" provides information concerning the physical configuration of the system, including information related to causal connections, length of pipes, position of valves on pipes, size of the reservoir, etc. Also, the moment-to-moment values of each of the variables (T, V_1, V_2, V_3, I_1, I_2, O, & R) could be considered at the level of physical form. All of these constraints will be satisfied if the process is being controlled in a proper fashion.

To summarize, an abstraction hierarchy analysis provides information about the hierarchically nested constraints that constitute a domain’s semantics, and therefore defines the information that

must be present in the interface for an individual to perform successfully. The product of this analysis (interrelated categories of information) provides a structured framework for display development, as we will demonstrate shortly. It should be emphasized that this analysis and description is independent of the interface, and therefore differs from traditional task analysis. Although space limitations do not permit a complete discussion, we view abstraction hierarchy analysis and task analysis (traditional or "cognitive") as complementary processes that are necessary for the development of effective displays.

20.4.5.1 Coherence and correspondence: Alternative mappings

In order for the principles of correspondence, coherence, process constraints, display constraints (and the mappings between them) to be useful for design they must be illustrated clearly. In this section we provide six example displays that provide alternative mappings for our simple process (see Figure 13). The discussion is organized in terms of the distinction between integral, configural, and separable dimensions that was outlined in Section ?. One goal is to illustrate what these terms, derived from the attention literature, mean in the context of display design for complex systems.

==========

Insert Figure 13 about here

==========

The second goal is to focus on the quality of the mapping that each display provides, especially with respect to the ability of each display to convey information at the various levels of abstraction (see Section ?). To illustrate the quality of the mapping explicitly we have provided a summary listing (at the right of each display in Figure 13) that sorts the associated process constraints into two categories ("P" & "D"). Process constraints that are directly represented in the display (that is,

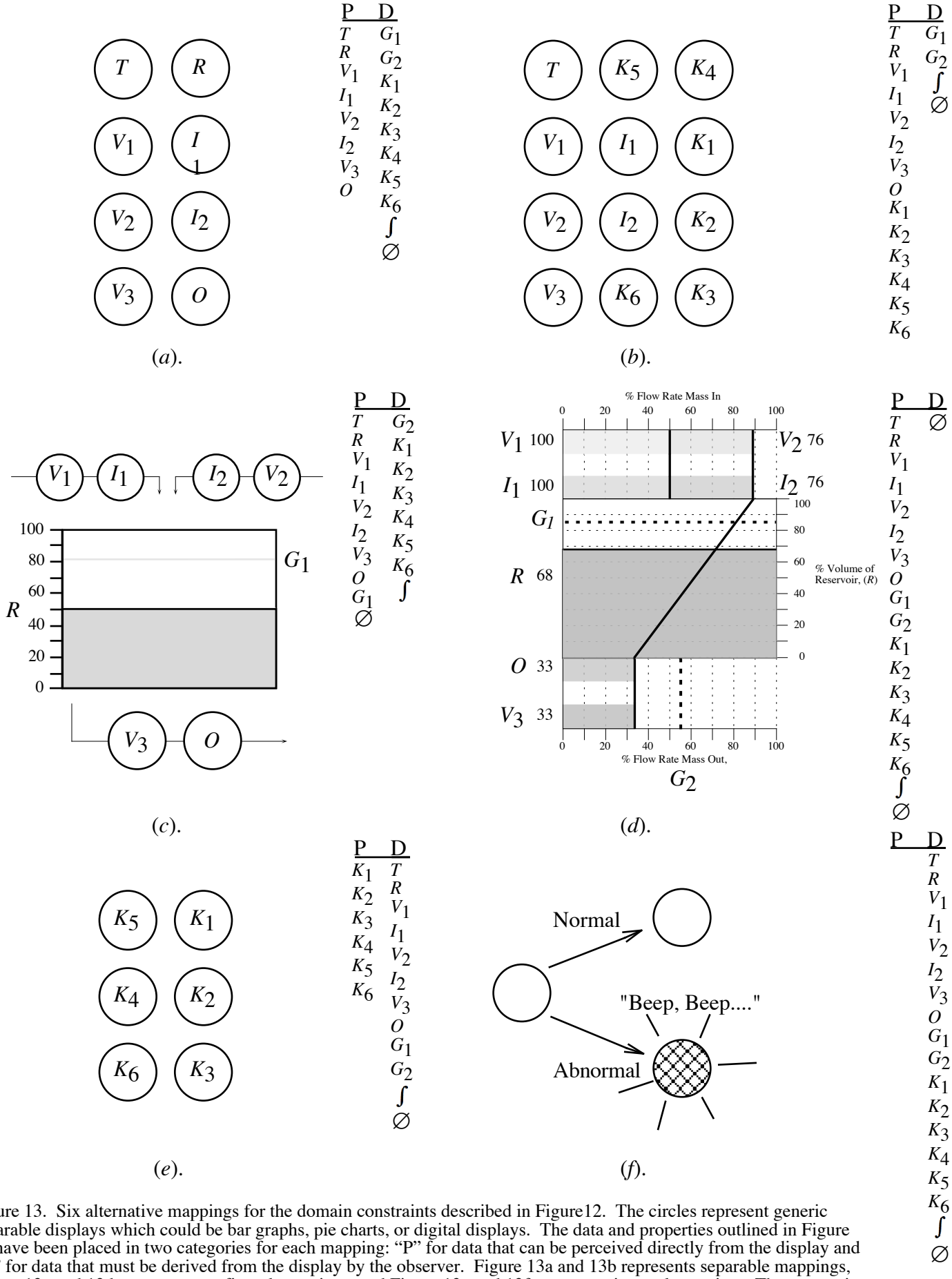


Figure 13. Six alternative mappings for the domain constraints described in Figure12. The circles represent generic separable displays which could be bar graphs, pie charts, or digital displays. The data and properties outlined in Figure 12 have been placed in two categories for each mapping: "P" for data that can be perceived directly from the display and "D" for data that must be derived from the display by the observer. Figure 13a and 13b represents separable mappings, Figure 13c and 13d represents configural mappings, and Figure 13e and 13f represents integral mappings. These mappings illustrate how the terms separable, configural, and integral have a different meaning when applied to display design (as opposed to attention).

which can be "seen") have been placed in the P category (Perceived). Process constraints that are not directly represented, and must be computed or inferred, are placed in the D category (Derived). Process constraints that are related to physical structure are represented by the theta symbol (ø); process constraints related to the functional structure are represented by the symbol (∫)

Separable displays. Figure 13a represents a separable display which contains a single display for each individual process variable present. Each display is represented in the figure by a circle, but no special significance should be attached to the symbology: the circles could represent digital displays, bar graphs, etc. For example, four instantiations of this display are shown in Figure 10a 10b, 10e, and 10f. For Figure 10a and 10b the display constraints are the relative heights of the bars in response to changes in the underlying variables.

In terms of the abstraction hierarchy the class of displays represented by Figure 13a provides information only at the level of physical function: individual variables are directly represented. Thus, there is not likely to be a selective attention cost for low-level data. However, there is likely to be a divided attention cost, because the observer must derive the high-level properties. To do so, the observer must have an internalized model of the functional purpose, the abstract functions, the general functional organization, and the physical process. For example, to determine the direction (and cause) of ΔR would require detailed internal knowledge about the process, since no information about physical relationships (ø) or functional properties (∫) is present in the display.

Simply adding information about high-level properties does not change the separable nature of the display. In Figure 13b a second separable display has been illustrated. In this display the high level properties (constraints) have been calculated and are displayed directly, including information related to functional purpose (K_5 & K_6) and abstract function (K_1, K_2, K_3 & K_4). This does

off-load some of the calculational requirements (e.g., ΔR). However, there is still a divided attention cost. Even though the high-level properties have been calculated and incorporated into the display, the relationships among and between levels of information in the abstraction hierarchy are still not apparent. The underlying cause of a particular system state still must be derived from the separate information that is displayed. Thus, while some low level integration is accomplished in the display, the burden for understanding the causal structure still rests in the observer's stored knowledge.

Configural displays. The first configural display, illustrated in Figure 13c, provides a direct representation of much of the low-level data that is present in the display in Figure 13a. However, it also provides additional information that is critical to completing domain tasks: information about the physical structure of the system (ø). This "mimic" display format was first introduced in STEAMER (Hollan, Hutchins, & Weitzman, 1984), and issues in the animation of these formats have been investigated more recently (Bennett, 1993; Bennett and Madigan, 1994; Bennett and Nagy, in press).

The mimic display is an excellent format for representing the generalized functions in the process. It has many of the properties of a functional flow diagram or flow chart. The elements can represent physical processes (e.g., feedwater streams) and, by appropriately scaling the diagram, relations at the level of physical form can be represented (e.g., relative positions of valves). Also, the moment-to-moment values of the process variables can easily be integrated within this representation. This display not only includes information with respect to generalized function, physical function, and physical form, but the organization provides a visible model illustrating the relations across these levels of abstraction. This visual model allows the observer to "see" some of the log-

ical constraints that link the low-level data. Thus, the current value of I_2 can be seen in the context of its physical function (feedwater stream 2) and its generalized function (source of mass) and in fact, its relation to the functional purpose in terms of G_1 is also readily apparent from the representation.

Just as in the displays listed in Figure 13a and 13b, there is not likely to be a cost in selective attention with respect to the lower-level data. However, although information about physical structure illustrates the causal factors that determine higher-level system constraints, the burden of computing these constraints (e.g., determining mass balance) rests with the observer. Thus, what is missing in the mimic display is information about abstract function (information about the physical laws that govern normal operation).

The second configural display, illustrated in Figure 13d, is slightly more complex (the logic is similar to Vicente, 1991) and will be described in detail before discussing the quality of the mapping that it provides. The valve settings V_1 and V_2 are represented as back-to-back horizontal bar graphs that increase or decrease in horizontal extent with changes in settings. The measured flow rates (I_1 & I_2)have the same configuration of graphical elements and are located below the valve settings in the display. The horizontal bar graphs depicting valve settings and flow rates for a particular pipe (e.g., V_1 & I_1) are connected with a line (in this case both of the lines are perpendicular because the settings and flow rates are equal in both input streams). The volume of the reservoir (R) is represented as the filled portion of the rectangle. The value of R can be read from the scale and associated digital value on the right side of the display; in Figure 13d the value of R is 68. The associated reservoir volume goal (G_1) is represented by the bold horizontal dashed line (approximately 85). The flow rate of the mass leaving the reservoir is represented by the horizontal bar

graph labelled "O" at the bottom of the display; the corresponding valve setting is represented by the bar graph labelled "V_3." These two bar graphs are also connected by a vertical line. The mass output goal (G_2) is represented by the bold vertical dashed line (approximately 55). The relationship between mass in ($I_1 + I_2$) and mass out (O) is highlighted by the bold angled line which connects the corresponding bar graphs.

Unlike the displays that have been discussed previously, this configural display integrates information from all levels of the abstraction hierarchy in a single representation, making extensive use of the geometrical constraints of equality, parallel lines, and colinearity. The general functions are related through a funnel metaphor with input (source) at the top, storage in the center, and output (sink) on the bottom. The abstract functions are related using the equality and the resulting colinearity across the bar graphs. For example, the constraints on mass flow (K_1, K_2, K_3) are represented in terms of equality of the horizontal extent of the bars labelled V_1/I_1, V_2/I_2, and $V_3/0$. In addition, the constraints relating rate of volume change and mass balance (K_4) are represented by the horizontal extent of I_1+I_2 relative to the horizontal extent of O, and these relationships are highlighted by the line connecting these bars. Thus, the mass balance is represented by the symmetry between the input bar graphs and the output bar graphs; the slant of the line connecting them should be proportional to rate of change of mass in the reservoir. Constraints at the level of functional purpose are illustrated by the difference between the goal and the relevant variable. For example, the constraint on mass inventory (K_5) is shown using relative position between the hatched area representing volume within the reservoir and the level marked G_1.

This configural display, while not a direct physical analog, preserves important physical relations from the process (e.g., volume & filling). In addition, it provides a direct visual representa-

tion of the process constraints and connects these constraints in a way to make the "functional" logic of the process visible within the geometric form. As a result, performance for both selective (focused) and divided (integration) tasks is likely to be facilitated substantially.

Integral displays. Figure 13e shows an "integral" mapping in which each of the process constraints are shown directly, providing information at the higher levels of abstraction. However, the low-level data must be derived. In addition, there is absolutely no information about the functional processes behind the display and therefore the display does not aid the observer to relate the higher level constraints to the physical variables. Because there would normally be a many-to-one mapping from physical variables to the higher order constraints it would be impossible for the observer to recover information at lower levels of abstraction from this display.

Figure 13f shows the logical extreme of this continuum. In this display, the process variables and constraints are integrated into a single "bit" of information, that indicates whether the process is working properly (all constraints are at their designed value) or not. It should be obvious that while these displays may have no divided attention costs, they do have selective attention costs and they also provide little support for problem solving when the system fails.

Summary. This section has focused on issues related to the quality of mapping between process constraints and display constraints. Even the simple domain that we chose for illustrative purposes has a nested structure of domain constraints: there are multiple constraints that are organized hierarchically both within and between levels of abstraction. The six alternative displays achieved various degrees of success in mapping these constraints. The principle of correspondence is illustrated by the fact that these formats differ in terms of the amount of information about the underlying domain that is present. The display in Figure 13f has the lowest degree of correspondence,

while the displays in Figure 13b and 13d have the highest degree of correspondence. These two displays are roughly equivalent in correspondence, with the exception of the two goals that are present in Figure 13d but absent in Figure 13b. Although these two displays are roughly equivalent in correspondence, it should be clear from the prior discussion that they are definitely not equivalent in terms of coherence. Figure 13d allows an individual to perceive information concerning the physical structure, functional structure, and hierarchically nested constraints in the domain directly, a capability that is not supported by the format in Figure 13b. The coherence of Figure 13d will be explored in greater detail in the following section. This section has also illustrated the duality of meaning for the terms integral, configural, and separable. In attention these terms refer to the relationship between perceptual dimensions, as described in section ?; in display design these terms more appropriately refer to the nature of the mapping between the domain and the representation.

20.4.5.2 Representation aiding for normal and abnormal operating conditions

In the previous section we outlined differences in correspondence and coherence that resulted from six alternative mappings for our simple domain. In this section we explore issues related to coherence in greater detail, focussing on Figure 13d and the implications of the mapping for performance under both normal and abnormal, or emergency operating conditions. To begin, we discuss the facilitating role that graphical constraints representing information in the abstraction hierarchy (in particular, abstract function -- the physical laws that govern normal operation) can play under normal conditions. Properly designed configural displays will provide a powerful representation for control: breaks in the domain constraints will generally be seen as breaks in display symmetries, and will suggest appropriate control inputs. This information is, perhaps, even more important for detecting faults (e.g., a leak). The possibility that these types of displays can change

the fundamental nature of the behavior that is required on the part of the operator will also be entertained. Finally, the implications for the reduction of errors (more likely to occur under abnormal or emergency conditions) will be discussed.

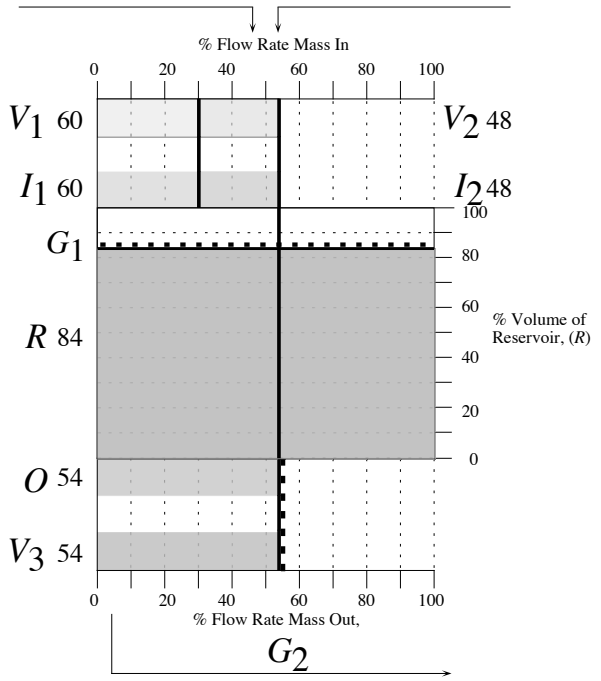
The mapping between domain constraints and geometrical constraints that is provided in the configural display shown in Figure 13d provides a powerful representation for control under normal operating conditions. In Figure 14a the display is shown with values for system variables indicating that all constraints are satisfied. The figure indicates that the flow rate is larger for the first mass input valve (I_1, V_1) than for the second (I_2, V_2) but that the two flow rates added together match the flow rate of the mass output valve (O, V_3). In addition, the two system goals (G_1 & G_2) are being fulfilled.

==========

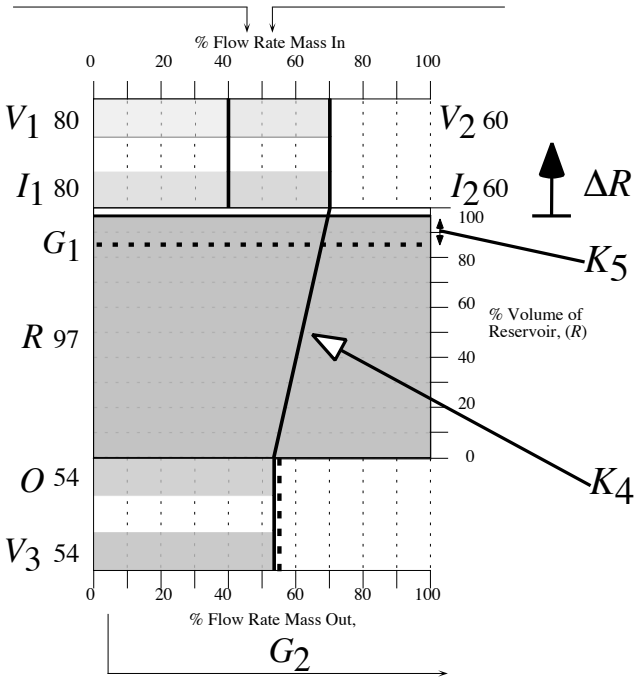
Insert Figure 14 about here

==========

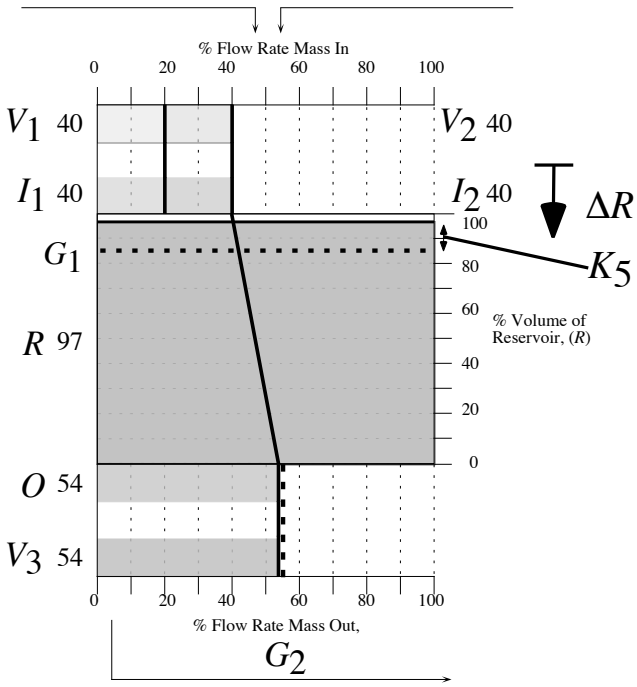
In contrast, Figure 14b, 14c, and 14d illustrates failures to achieve system goals. In these displays not only is the violation of the goal easily seen, but each system variable is seen in the context of the control requirements. Thus, in Figure 14b it is apparent that the K_5 constraint is not being met (the actual level of the reservoir is higher than the goal). It is also apparent that the K_4 constraint is broken. The orientation of the line connecting mass in (I_1 + I_2)and the mass out (O) utilizes the funnel metaphor to indicate that a positive net inflow for mass exists (mass in is greater than mass out). In essence, the deviation in orientation of this line from perpendicular is an emergent feature corresponding to the size of the difference. Under these circumstances control input is required immediately: an adjustment at Valve 1 and/or Valve 2 will be needed to avoid overflow from the reservoir. The observer can see these valves in the context of the two system goals; the



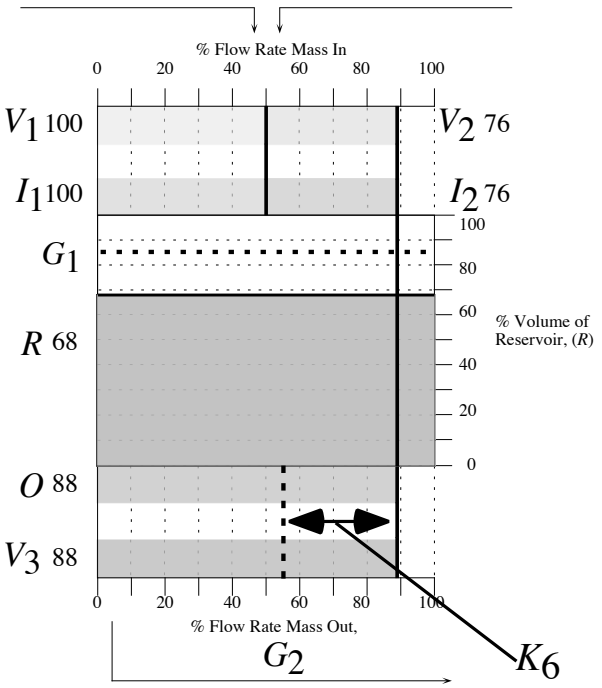
(*a*).



(*b*).



(*c*).



(*d*).

Figure 14. Illustration of the mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under relatively normal operating conditions.

representation makes it clear that these are the appropriate control inputs to make. For example, although adjusting Valve 3 from 54 to a value greater than 70 would also cause the reservoir volume to drop, it is an inappropriate control input because Goal 2 would then be violated.

In Figure 14c the situation is exactly the same, with one exception: there is a negative net inflow for mass, as indicated by the reversed orientation of the connecting line. Under these circumstances the operator can see that no immediate control input is required. Because mass in is less than mass out, the reservoir volume is falling, and this is exactly what is required to meet the G_1 reservoir volume goal. Of course, a control input will be required at some point in the future (mass will need to be balanced when the reservoir level approaches the goal). Similarly, in Figure 14d the observer can see that the K_5 and K_6 constraints are broken, and that an adjustment to Valve 3 (a decrease in output) is needed to meet the output requirements (G_2) and the volume goal (G_1).

Thus, in complex, dynamic domains it is the pattern of relationships between variables, as reflected in the geometric constraints, that determines the significance of the data that is presented. It is this pattern that ultimately provides the basis for action, even when the action hinges upon the value of an individual variable. When properly designed, configural displays will directly reflect these critical data relationships, and suggest the appropriate control input.

A similar logic applies for operational support under abnormal, or emergency conditions. As in the previous figure, Figure 15a represents a configuration with all system constraints being met. In Figure 15b the first constraint (K_1) is broken. There are two aspects of the display geometry indicating that the flow rate (I_1) does not match the commanded flow, or valve setting (V_1). First, the horizontal extent of the two bar graphs in the top left portion of the display are not equal and

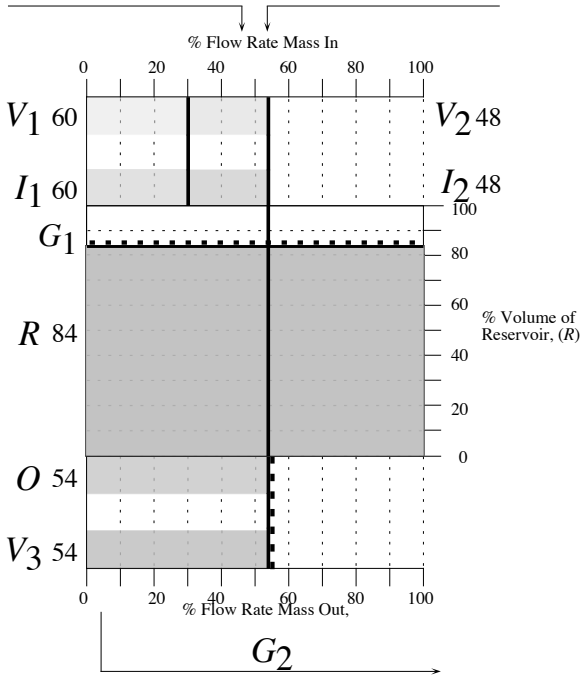
this relationship is emphasized by the bold line connecting the two graphs (similar to the connecting line for mass balance). There are a number of potential causes for this discrepancy, which include 1) a leak in the valve, 2) a leak in the pipe prior to the point at which the flow rate is measured, or 3) an obstruction in the pipe. In contrast, the fact that the line connecting V_2 and I_2 is not perpendicular (but is parallel to the first connector line) does not indicate that the K_2 constraint is broken. Instead, this is an indication that the commanded and actual mass flows in the second mass input stream are equal (and therefore that the discrepancy is isolated in the first mass input stream). A similar mapping between geometrical constraints and domain constraints is used to represent a fault in the K_3 constraint, as illustrated in Figure 15c.

==========

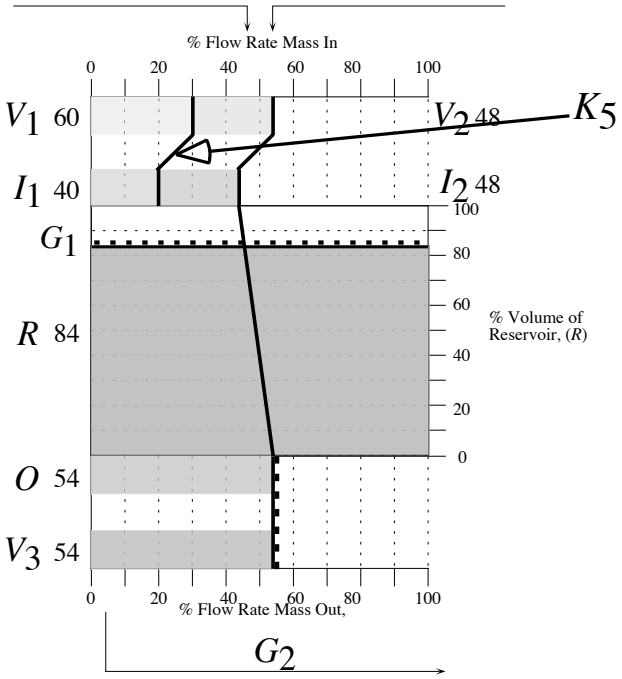
Insert Figure 15 about here

==========

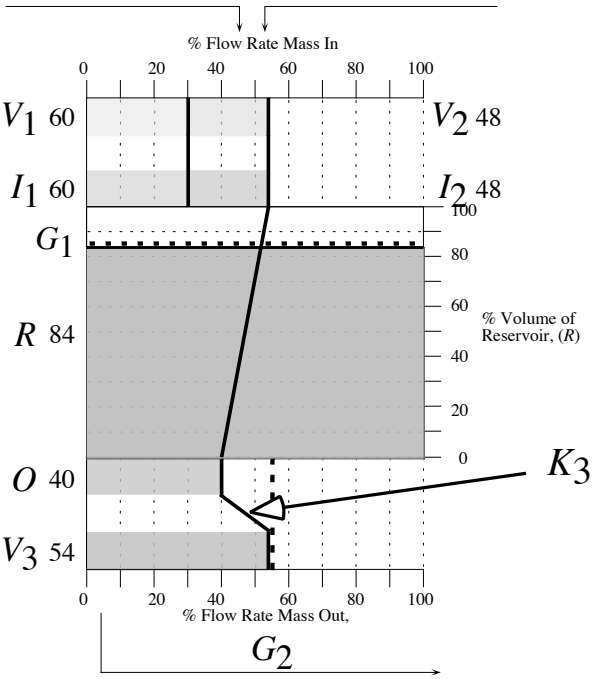
Figure 15d illustrates the geometrical constraints associated with a break (a fault) in the mass balance constraint (K_4). In this example there is a positive net inflow of mass, which is normally associated with an increase in the volume of the reservoir (again, suggested by the funnel metaphor). However, in this case the mass inventory is falling, as we have indicated in the diagram by the downward-pointing arrow located near the ΔR symbol (this is difficult to represent in a static diagram, but would be clearly seen on a dynamic display). Again, there are several potential explanations for this fault. The most likely explanation is that there is a leak in the reservoir itself, however, there could be a leak in the pipe between the reservoir and the point at which the flow measurement is taken. It should be noted that while the nature of the fault can be seen (e.g., leak or blockage in feedwater line) this representation would not be very helpful in physically locating the leak within the plant (e.g., locating Valve 1).



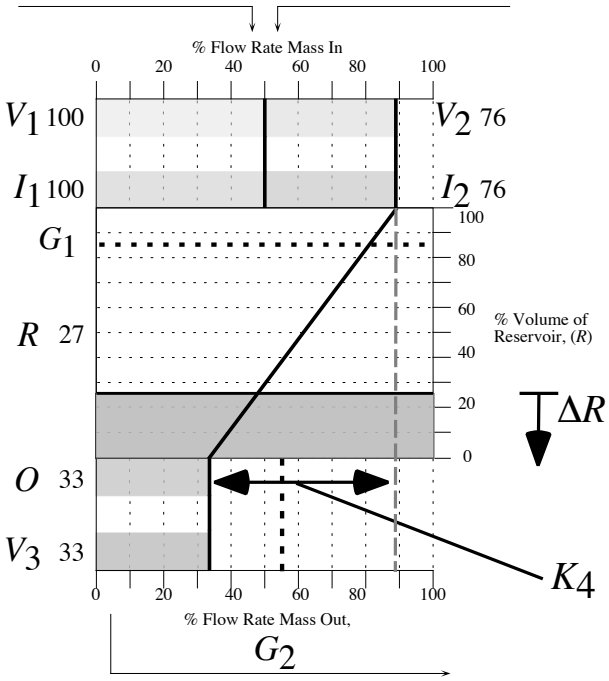
(*a*).



(*b*).



(*c*).



(*d*).

Figure 15. Illustration of the mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under abnormal or emergency operating conditions.

These examples illustrate that properly designed displays can change the fundamental type of behavior that is required of an operator under both normal and abnormal operating conditions. With separable displays (e.g., the separable configurations illustrated in Figure 13) the operators are required to engage in knowledge-based behaviors. As previously noted, operators must rely upon internal models of system structure and function (and therefore use limited capacity resources -- working memory) to detect, diagnose, and correct faults. As a result, the potential for errors is increased dramatically. In contrast, properly designed configural displays present externalized models of the system structure, function, state through the nature of the geometric constraints. This allows operators to utilize skill-based behaviors (e.g., visual perception and pattern recognition) that do not require limited capacity resources. As a result, the potential for errors will be dramatically decreased. As Rasmussen and Vicente (1989) have noted, changing the required behavior from knowledge-based behavior to rule-based or skill-based behavior is a goal for display design.

Properly designed configural displays will also reduce the possibility of "under-specified action errors" (Rasmussen and Vicente, 1989). In complex, dynamic domains individuals can form incorrect hypotheses about the nature of the existing problem if they do not consider the relevant subsets of data (Woods, 1988). Observers may focus on these incorrect hypotheses and ignore disconfirming evidence, showing a kind of "tunnel vision" (Moray, 1981). Observers may also exhibit "cognitive hysteresis," and fail to revise hypotheses as the nature of the problem changes over time (Lewis and Norman, 1986). Configural displays that directly reflect the semantics of a domain can reduce the probability of these types of errors by forcing an observer to consider relevant subsets of data.

20.4.6 Practical guidelines.

In conclusion, we believe that the application of the this approach to display design will improve overall human machine performance through the development of configural displays that support normal control, as well as fault detection, diagnosis, and repair. The potential for errors will be dramatically decreased, because the critical information for control is represented directly in the interface. This in turn dramatically reduces the requirement for knowledge-based reasoning on the basis of internalized models. To summarize, we offer three general heuristics for graphic design.

1. Each relevant process variable should be represented by a distinct element within the display. If precise information about this variable is desirable, then a reference scale or supplemental digital information should be provided.

2. The display elements should be organized so that the emergent properties (symmetries, closure, parallelism) that arise from their interaction correspond to higher order constraints within the process. Thus, when process constraints are broken (i.e., a fault occurs) the corresponding geometric constraints are also broken (the display symmetry is broken).

3. The symmetries within the display should be nested (from global to local) in a way that reflects the hierarchical structure of the process. High-order process constraints (e.g., at the level of functional purpose or abstract function) should be reflected in global display symmetries; lower order process constraints (e.g., functional organization) should be reflected in local display symmetries.

20.5.0 Challenges Of Complex Systems

The simple process described above is convenient for a tutorial introduction to some of the im-

portant decisions that must be made when designing a graphical representation. However, this example, greatly underestimates the complexity seen in many advanced, human-technological systems (e.g., nuclear power, air traffic control, advanced tactical aviation, command and control centers for managing military and space operations, minimally invasive and remote surgery, etc.). These systems typically have multiple modes of operation (each with different constraints and boundary conditions) and require multiple windows into the process. In these systems, the goal remains the same, to make the real constraints of the work process (at all levels of abstraction) visible to the human operator. The designer must still address the problems of correspondence, so that all relevant process constraints are represented in the interface, and coherence, so that the representation is comprehendible to the human operator. For these complex systems, however, it will not be possible to achieve both correspondence and coherence with a single graphic display. Thus, the added problem of navigating through multiple views (i.e., windows, screens, pages) must be addressed.

A principal threat to these complex systems is "mode error" (Woods, 1984). A mode error occurs when the operator loses track of dynamic changes in the operating constraints governing a process. The operator responds to one set of constraints (i.e., mode), when a different set of constraints (mode) is, in fact, governing the process. The design challenge is to coordinate the multiple windows necessary for a complete representation with the changing operational modes. In simple terms, how can the interface be designed to insure that the appropriate window is always coupled with the appropriate mode; to insure that the important information is salient at the appropriate times.

Two classes of solutions might be considered for dealing with the navigation problems that typ-

ically lead to mode errors --- computational and graphical. Computational solutions or adaptive interfaces include an inference engine that automatically manages the representation. This computational engine automatically adjusts the representation based on inferences about the state of the system and the state of the operator. Projects such as the "pilot's associate" are examples of attempts to design automatic systems to aid operators to navigate through the representations associated with a complex work domain. However, the focus of this chapter is on graphical solutions. For this reason, we will use the remaining space to briefly consider some graphical approaches to this problem.

Woods (1984) introduced the term "visual momentum" to refer to the cognitive costs associated with switching from one reference frame to another. If visual momentum is high, then the cost of switching views is low. In this case, the new display is consistent with expectations created by the prior display. If visual momentum is low, then there is a high cost of switching. That is, the new display is not consistent with expectations and the cognitive system must effectively recalibrate before information can be extracted from the new display. To insure high visual momentum, the design of each graphical display must be considered relative to the other displays that operators may be using. Are the graphical conventions (e.g., coordinates, scales, directions, motions, colors, S-R mappings, etc.) used in one display consistent with those in another?

A graphical device that Woods (1984) has suggested to increase visual momentum is the use of landmarks. Landmarks are graphical elements that provide an orientation point that relates one display to another. Just as a tall building or mountain that is visible from many different parts of the landscape might help a person to orient to the geography, graphical landmarks can be designed with the objective of aiding the operator to orient within the functional landscape of the work do-

main. For example, Aretz (1991) used a shaded wedge within an electronic map display as a landmark to specify the region within the map that corresponded to the head-up forward view of the pilot.

Another graphical device to help operators navigate across multiple display pages is a map or overview display. This display can be implemented as a separate window or as an embedded landmark in all windows. This overview might use a flow diagram or hierarchical tree structure to show functional links among the multiple display pages.

The BookHouse interface designed by Goodstein and Pejtersen (1989) uses a spatial metaphor in which rooms in a "house" are set-up for different categories of users. This spatial metaphor allows the operator to apply natural abilities for navigating in three-dimensional spaces to the task of navigating in the more abstract space of a library database. In the BookHouse the three dimensional space is implemented in a two-dimensional display. Virtual Reality systems now offer the possibility for effective 3-dimensional representations. With these systems, designers have the opportunity to maximize the transfer of natural human ability to orient and navigate in 3-D environments to more abstract environments; and to combine natural 3-D representations with imagery obtained by advanced sensor systems. For example, virtual displays for minimally invasive surgery are being designed that integrate the 3-D image of the patient's anatomy with information obtained by MRI scans and other advanced imaging technologies. Thus, virtual 3-dimensional spatial metaphors might provide another technique for integrating complex information from distributed sensors into a coherent representation.

When designing displays to support operators in complex systems, most of the important functional constraints, as might be characterized in the abstraction hierarchy, are either explicitly or im-

plicitly determined as a result of design decisions. The challenge then is to make all of these constraints explicit within the display representation. However, for many complex systems, the constraints are not completely known by the designers and engineers. In fact, the possibility of unexpected contingencies is one important reason to include humans in control loops. Can displays be designed to allow operators to respond to constraints that are not understood at the time of the design? In this case, where the operator must "discover" previously unknown constraints, the display must support exploration. Here Shneiderman's (1987; 1993) concept of "direct manipulation" may be very important. Direct manipulation is a term that was introduced to characterize interfaces that create a sense of immersion in which the operator experiences a sense of direct control through the interface. Shneiderman lists several attributes of these interfaces. Perhaps the most important of these for exploration is the following: "rapid, incremental, reversible operations whose impact on the object of interest is immediately visible" (1993, p. 30).

To support exploration, an interface should allow the operator to make direct manipulations, or in other words, the interface must allow the operator to experiment, to actively test hypotheses. Here the key may be flexibility to move around within the problem space, to adopt multiple perspectives, to zoom in and out, to act on the environment in a way that produces immediate feedback. A good example is seen in the development of exploratory data analysis tools that are now being incorporated in many statistical packages. These visualization tools help the scientists to discover constraints in their complex data sets. Shneiderman (1993) describes several interfaces such as the Dynamic Home Finder that are designed to allow flexible exploration of a complex data set (e.g., homes available in the vicinity of Washington, DC).

So, in many complex systems the display designer must consider both the problem of direct

perception (allowing the operator to see known constraints) and the problem of direct manipulation (allowing the operator to explore and discover unknown constraints).

The central theme of this chapter is that problem solving can be critically influenced by the nature of visual representations. Building effective representations requires designers to go beyond the simple psychophysical questions of data availability to the more complex questions of information availability. Where information refers to the specification of domain constraints and boundary conditions. This specification depends both on the mapping from display to human (i.e., coherence) and on the mapping from display to domain (i.e., correspondence).

20.6.0 Acknowledgments

The authors would like to thank Brian Tsou for discussions and comments on earlier drafts, and the helpful comments provided by anonymous reviewers. Funding in support of this work was provided to Kevin Bennett by the Ohio Board of Regents (Wright State University Research Challenge Grant # 662613). John Flach was partially supported by a grant from the Air Force Office of Scientific Research during the preparation of this manuscript. Opinions expressed are those of the authors and do not represent an official position of any of the supporting agencies.

20.7.0 References

Adelson, E.H. (1993). Perceptual organization and the judgement of brightness. Science, 262, 2042-2044.

Alexander, K.R., Xie, W., Derlacki, D.J. (1994). Spatial frequency characteristics of letter identification. J. Opt. Soc. Am. A, 11, 2375-2382.

Arend, L.E., and Reeves, A. (1986). Simultaneous color constancy. J. Opt. Soc. Am. A, 3, 1743-1751.

Aretz, A. J. (1991). The design of electronic map displays. Human Factors, 33(1), 85-101.

Beltracchi, L. (1987). A direct manipulation interface for heat engines based upon the Rankine cycle. IEEE Transactions on Systems, Man, and Cybernetics, 17 (3), 478-487.

Beltracchi, L. (1989). Energy, mass, model-based displays, and memory recall. IEEE Transactions on Nuclear Science, 36 (3), 1367-1382.

Bennett, K. B. (1993). Encoding apparent motion in animated mimic displays. Human Factors, 35(4), 673-691

Bennett, K. B., and Flach, J. M. (1992). Graphical displays: Implications for divided attention, focused attention, and problem solving. Human Factors, 34(5), 513-533.

Bennett, K. B., and Madigan, E. (1994). Contours and borders in animated mimic displays. International Journal of Human-Computer Interaction, 6(1), 47-64.

Bennett, K. B., and Nagy, A. (In Press). Spatial and temporal considerations in animated mimic displays. Displays.

Bennett, K. B., Toms, M. L., and Woods, D. D. (1993). Emergent features and configural elements: Designing more effective configural displays. Human Factors, 35(1), 71-97.

Boynton, R. M. (1990). Human color perception. In K.N. Leibovic (Ed.), Science of vision (pp. 211-253). New York, NY: Springer-Verlag,

Boynton, R.M. (1992). Human color vision. Washington, D.C.: Optical Society of America.

Brainard, D.H., and Wandell, B.A. (1992). Asymmetric color-matching: How color appearance depends on the illuminant. J. Opt. Soc. Am. A, 9, 1433-1448.

Buttigieg, M. A., and Sanderson, P.M. (1991). Emergent features in visual display design for two types of failure detection tasks. Human Factors, 33(6), 631-651.

Cannon, M.W., and Fullenkamp, S.C., (1991). A transducer model for contrast perception. Vision Research, 31, 983-998.

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

Cleveland, W. S. (1985). The elements of graphing data. Belmont, CA: Wadsworth.

DeValois, R.L., and DeValois, K.K. (1990). Spatial vision. New York, NY: Oxford University Press.

D'Zmura, M., and Lennie, P. (1986). Mechanisms of color constancy. J. Opt. Soc. Am. A, 3, 1662-1672.

Garner, W. R. (1970). The stimulus in information processing. American Psychologist, 25, 350-358.

Garner, W. R. (1974). The processing of information and structure. Hillsdale, NJ: Lawrence Erlbaum Associates.

Garner, W. R., and Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. Cognitive Psychology, 1, 225-241.

Gilchrist, A., S. Delman, and Jacobson, A. (1983). The classification and identification of edges as critical to the perception of reflectance and illumination. Percept. and Psychophys., 33, 425-436.

Ginsburg, A. (1986). Spatial filtering and visual form perception. In K.R. Boff, L. Kaufmann and J.P. Thomas (Eds.), Handbook of human perception and performance, (pp 34/1 - 34/41). New York, NY: John Wiley.

Goodstein, L.P. & Pejtersen, A.M. (1989). The BOOK HOUSE: System --- functionality and evaluation (Riso-M-2793). Roskilde, Denmark: Riso National Laboratory.

Graham, N. V. S. (1989). Visual pattern analyzers. New York., NY: Oxford University Press,

Grum, F., and Bartleson, C.J. (1980). Optical radiation measurements, Volume 2: Colorimetry. New York, NY: Academic Press.

Hollan, J. D., Hutchins, E. L., and Weitzman, L. (1984). Steamer: An interactive inspectable simulation-based training system. The AI Magazine, Summer, 15-27.

Hollan, J. D., Hutchins, E. L., McCandless, T. P., Rosenstein, M., and Weitzman, L. (1987). Graphical interfaces for simulation. In W. B. Rouse (Ed.), Advances in man-machine systems

research (Vol. 3, pp. 129-163). Greenwich, CT: JAI Press.

Hunter, R.S., and Herold, R.W. (1987). The measurement of appearance. New York, NY: John Wiley and Sons.

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1986). Direct manipulation interfaces. In D. A. Norman, and S. W. Draper (Eds.), User centered system design (pp. 87-124). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelly, D.H. (1974). Spatio-temporal frequency characteristics of color-vision mechanisms. J. Opt. Soc. Am., 64, 983-990.

Klein, G. A., Orasanu, J., and Zsambok, C. E. (Eds.). (1993). Decision making in action: Models and methods. Norwood, NJ: Ablex Publishing Corp.

Lewis, C., and Norman, D. A. (1986). Designing for error. In D. A. Norman, and S. W. Draper (Eds.), User centered system design. Hillsdale, NJ: Lawrence Earlbaum Associates.

Maloney, L.T., and Wandell, B.A. (1986). Color constancy: A method for recovering surface reflectance. J. Opt. Soc. Am. A, 3, 29-33.

Millodot, M. (1982). Image formation in the eye. In H.Barlow and J.D. Mollen (Eds.), The senses (pp. 46-61). New York, NY: Cambridge University Press.

Moray, N. (1981). The role of attention in the detection of errors and the diagnosis of failures in man-machine systems. In J. Rasmussen and W. B. Rouse (Eds.), Human detection and diagnosis of system failures (pp. ??-??). New York, NY: Plenum Press.

Moray, N., Lee, J., Vicente, K.J., Jones, B.G., and Rasmussen, J. (1994). A direct perception interface for nuclear power plants. Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting (pp. 481-485). Santa Monica, CA: Human Factors and Ergonomics Society.

Mullen, K. (1985). The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings. J. Physiol., 359, 381-400.

Nagy, A.L., and Kamholz, D. (1995). Luminance discrimination, color contrast, and multiple

mechanisms. Vision Research, 35, 2147-2155.

Nassau, K. (1983). The physics and chemistry of color. New York, NY: John Wiley and Sons.

Newell, A., and Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice Hall.

Noorlander, C., and Koenderink, J.J. (1983). Spatial and temporal discrimination ellipsoids in colour space. J. Opt. Soc. Am., 73, 1533-1543.

Olzack, L., and Thomas, J. (1986). Seeing spatial patterns. In K. R. Boff, L. Kaufmann, and J.P. Thomas (Eds.), Handbook of human perception and performance (pp. 7/1-7/56). John Wiley, New York.

Pejtersen, A.M. (1992). The Book House. An icon based database system for fiction retrieval in public libraries. In B. Cronin (Ed.), The marketing of library and information services 2. (pp. 572 - 591). London, UK: ASLIB.

Poirson, A.B., and Wandel, B.A. (1993). The appearance of colored patterns. J. Opt. Soc. Am. A, 12, 2458-2471.

Pokorny J. and Smith, V. (1986). Colorimetry and color discrimination. In K. R. Boff, L. Kaufmann, and J.P. Thomas (Eds.), Handbook of human perception and performance (pp. 8/1-8/51). John Wiley, New York.

Pomerantz, J. R. (1986). Visual form perception: An overview. In H. C. Nusbaum and E. C. Schwab (Eds.), Pattern recognition by humans and machines (Vol. 2, Visual Perception, pp. 1-30). Orlando, Florida: Academic Press.

Pomerantz, J. R., and Pristach, E. A. (1989). Emergent features, attention, and perceptual glue in visual form perception. Journal of Experimental Psychology: Human Perception and Performance, 15(4), 635-649.

Pomerantz, J. R., Sager, L. C., and Stoever, R. J. (1977). Perception of wholes and of their component parts: Some configural superiority effects. Journal of Experimental Psychology: Human Perception and Performance, 3, 422-435.

Post, D. (1992). Colorimetric measurement, calibration, and characterization of self-luminous displays. In H. Widdel and D. L. Post (Eds.), Color in electronic displays (pp. 299-312). New York, NY: Plenum Press.

Rasmussen, J. (1986). Information processing and human-machine interaction: An approach to cognitive engineering. New York, NY: Elsevier Publishing Co., Inc.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994). Cognitive systems engineering. New York, NY: John Wiley & Sons., Inc.

Rasmussen, J., and Vicente, K. (1989). Coping with human errors through system design: Implications for ecological interface design. International Journal of Man-Machine Studies, 31, 517-534.

Sanderson, P. M., Flach, J. M., Buttigieg, M. A., & Casey, E. J. (1989). Object displays do not always support better integrated task performance. Human Factors, 31(2), 183-198.

Sekiguchi, N., Williams, D.R., and Brainard, D.H. (1993). Aberration free measurements of the visibility of isoluminant gratings. J. Opt. Soc. Am. A, 10, 2105-2117.

Shafer, S.A. (1985). Using color to separate reflection components. Color Res. and Appl., 10, 210-218.

Shneiderman, B. (Ed.). (1993). Sparks of innovation in human computer interaction. Norwood, NJ: Ablex.

Shneiderman, B. (1986). Designing the user interface. Reading, MA: Addison-Wesley.

Solomon, J.A., and Pelli, D.G. (1994) The visual filter mediating letter identification. Nature (London), 369, 395-397.

Treisman, A. M. (1986). Properties, parts, and objects. In K. Boff, L. Kaufmann, and J. Thomas (Eds.), Handbook of perception and human performance (pp. 35/1 - 35/70). New York, NY: Wiley.

Tufte, E. R. (1990). Envisioning information. Chesire, Connecticut: Graphics Press.

Tufte, E. R. (1983). The visual display of quantitative information. Chesire, Connecticut: Graph-

ics Press.

Vicente, K. J. (1991). Supporting knowledge based behavior through ecological interface design. (Tech. Report EPRL-91-1). Urbana Champaign, IL: Engineering Psychology Research Laboratory and Aviation Research Laboratory. University of Illinois.

Wandel, B.A. (1995). Foundations of vision. Sunderland, Mass.:Sinauer.

Wertheimer, M. (1959). Productive thinking. New York. NY: Harper & Row.

Whittle, P. (1986). Increments and decrements: Luminance discrimination. Vision Research, 26, 1677-1692.

Whittle, P. (1992). Brightness, discriminability, and the "Crispening Effect." Vision Research, 32, 1493-1508.

Wickens, C. D., and Carswell, C. M. (In Press). The proximity compatibility principle: Its psychological foundation and its relevance to display design. Human Factors.

Widdell, H., and Post, D.L. (1992). Color in electronic displays. New York, NY: Plenum Press.

Woods, D. D. (1984). Visual momentum: A concept to improve the cognitive coupling of person and computer. International Journal of Man-Machine Studies, 21, 229-244.

Woods, D. D. (1988). Coping with complexity: The psychology of human behavior in complex systems. In L. P. Goodstein, H. B. Andersen and S. E. Olsen (Eds.), Mental models, tasks and errors: A collection of essays to celebrate Jens Rasmussen's 60th birthday (pp. ??-??). New York, NY: Taylor Francis.

Woods, D. D. (1991). The cognitive engineering of problem representations. In G. R. S. Weir and J. L. Alty (Eds.), Human-computer interaction and complex systems (pp. 169-188). London, UK: Academic Press.

Woods, D. D. (1994).

Woods, D. D., and Roth, E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), Handbook of human-computer interaction (pp. 1-41). Amsterdam: Elsevier Science Publishers

B. V. (North-Holland).

Worthy, J.A., and Brill, M.H. (1986). Heuristic analysis of color constancy. J. Opt. Soc. Am. A, 3, 1708-1712.

Wyszecki, G., and Stiles, W.S. (1982). Color Science, (2nd ed.) New York, NY: Wiley.

Zachary, W. (1986). A cognitively based functional taxonomy of decision support techniques. Human-Computer Interaction, 2, 25-63.

20.8.0 Figure Captions

Figure 1 Figure 1a shows idealized spectral reflectance curves for the ink, the page, and the wall in the example described in the text. Figure 1b shows the relative energy at each wavelength in sunlight.

Figure 2 The CIE 1924 photopic luminosity curve.

Figure 3 The diagram illustrates the calculation of contrast ratios for the page, the text, and the wall. The values of "r" indicate the reflectances of the three surfaces in the figure. The symbol "I" in the equations represents the illumination level which is identical for all three surfaces in the figure and therefore cancels out of the equations.

Figure 4 Plots showing the variation in luminance for sinusoidal patterns. Figure 4a illustrates a spatial frequency of 1 cycle/degree at a contrast of 100%. Figure 4b illustrates a spatial frequency of 2 cycles/degree at a contrast of 100%. Figure 4c illustrates a spatial frequency of 1 cycle/degree at a contrast of 50%.

Figure 5 The diagram illustrates the calculation of visual angle as described in the text.

Figure 6 A typical plot of a contrast sensitivity function for a human observer. Based on data given in DeValois and DeValois (1990).

Figure 7 A spectral reflectance curve for red ink.

Figure 8 The CIE 1931 chromaticity diagram. Plotted from data given in Wyszecki and Stiles (1982).

Figure 9 A typical plot of contrast sensitivity for isoluminant chromatic gratings. Based on data given in Mullen (1985).

Figure 10. Six alternative mappings. Figure 10a and10 b represents alternative versions of a separable (bar graph) format that are less effective (10a) and more effective (10b) mappings. Similarly, Figure 10c and 10d represents alternative versions of a configural display format that are less (10c) and more effective (10d), primarily due to layering and separation. Figure 10e and 10f represents the least effective mappings.

Figure 11. The "cognitive triad": A cognitive systems engineering perspective. Any domain produces cognitive demands that must be met by the cognitive agents interacting with (or controlling) the domain. The cognitive agent possesses cognitive resources that must be used to meet these demands. The interface is the medium (or representation) through which the cognitive agent views and controls the domain. The effectiveness of an interface is determined by both correspondence and coherence.

Figure 12. A simple domain from process control that has a reservoir for storing mass, two input streams that increase the volume of mass in the reservoir, and a single output stream that decreases the volume. The low level data (the measured domain variables), the high-level properties (constraints that arise from the interaction of these variables and the physical design) and the domain goals (requirements that must be met for the system to be functioning properly) are listed.

Figure 13. Six alternative mappings for the domain constraints described in Figure12. The circles represent generic separable displays which could be bar graphs, pie charts, or digital displays. The data and properties outlined in Figure 12 have been placed in two categories for each mapping: "P" for data that can be perceived directly from the display and "D" for data that must be derived from the display by the observer. Figure 13a and 13b represents separable mappings, Figure 13c and 13d represents configural mappings, and Figure 13e and 13f represents integral mappings. These mappings illustrate how the terms separable, configural, and integral have a different meaning when applied to display design (as opposed to attention).

Figure 14. Illustration of the mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under relatively normal operating conditions.

Figure 15. Illustration of the mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under abnormal or emergency operating conditions.