Display Characterization by Eye:
Contrast Ratio and Discrimination Throughout the Grayscale

Jennifer Gille[1], Larry Arend[2], James Larimer[2]
[1]Raytheon ITSS, [2]Human Factors Research & Technology Division, [3]Army/NASA Rotorcraft Division
NASA Ames Research Center, Moffett Field, CA, 94035

ABSTRACT

We have measured the ability of observers to estimate the contrast ratio (maximum white luminance / minimum black or gray) of various displays and to assess luminous discrimination over the tonescale of the display. This was done using only the computer itself and easily-distributed devices such as neutral density filters. The ultimate goal of this work is to see how much of the characterization of a display can be performed by the ordinary user in situ, in a manner that takes advantage of the unique abilities of the human visual system and measures visually important aspects of the display. We discuss the relationship among contrast ratio, tone scale, display transfer function and room lighting. These results may contribute to the development of applications that allow optimization of displays for the situated viewer / display system without instrumentation and without indirect inferences from laboratory to workplace.

1. INTRODUCTION

The ultimate goal in the characterization of displays is the assurance of high quality rendering of content for human viewers. Depending on the application, "high quality rendering" can mean that the viewer's perception of the information content is accurate, that task performance is optimized, that the content has a pleasing appearance, or all three. In any case, the most important issue is human usability rather than device physics. In this paper we focus on display characterization in the workplace; that is, in common, everyday imaging settings.

1.1 Display characterization by instrument and by eye

There are important differences between the characterization of displays in the laboratory and characterization in the workplace.

In the laboratory, characterization of displays is usually based on photometric and colorimetric measurement of the light output of a display as a function of digital data input. Complete characterization of a physical display for design and manufacturing purposes or for technical imaging can involve a substantial battery of measures that describe color output, geometry, spatiotemporal performance (especially resolution), artifacts, and other measures. Visual observations, if any, usually play a secondary role.

In the imaging workplace, characterization of users' displays usually involves more limited goals. These include such things as testing for acceptance of new equipment, understanding the capabilities of new equipment, guidance for display adjustment by the user, indication of needed maintenance, and color management.

With more immediate, local goals, characterization in the workplace tends to involve a reduced set of physical measurements. A typical set might include the CIE xyY of the white point, the chromaticities of the primaries, and the curvature parameter, "gamma", for an assumed power-law digital-data-to-luminance transfer function.

In actual applications, display characterization must be done in situ, on the user's equipment and in the user's lighting environment. The display card in the user's computer, adjustments to the display such as brightness and contrast, and the user's visual system are components of the situated display system along with the display itself. The viewing environment is also part of the system. Reflected light, specular or diffuse, on the emissive area itself

or even the near surround, can dramatically lower the luminance contrast and color in local regions, over the whole display, or both.

In contrast to the laboratory environment, photometric and colorimetric characterization of displays in the workplace has several limitations that are barriers to routine, widespread characterization:

Instrumentation costs. Physical measurement of display characteristics requires instruments capable of measuring the chromatic and luminance variables sufficiently accurately, procedures that correctly capture the influence of the viewing environment on usability, and a user with some expertise in light measurement. It involves the expenses of acquiring and maintaining proper instruments, development of uniform procedures, and training of users. Although meters specifically designed for users to measure their displays have become more affordable and easier to use, they don't capture the substantial effects of the reflections of environmental lighting.

Standard display models. The individual situated display may not be well described by the assumed physical models of display characteristics. Manufacturing variations, user adjustments of controls, and the viewing environment are all potential sources of error. Also, measurement of physical parameters of the presumed models may give ambiguous results (Gille and Larimer, 2001.)

Indirect inferences. An even more important limitation of characterization by physical measurement is that it requires indirect inferences from the physical measures to the perceptual performance by a particular user in the workplace. Even with perfect physical measurements, conclusions about usability require arguments based on psychophysical models that may not accurately describe the particular observer in the particular workplace.

Given these problems with photometric measurement in the workplace, we decided to further investigate direct visual characterization of the performance of the display/user/environment system. A number of researchers have investigated using the human eye to characterize various aspects of display performance (Gille and Larimer, 2001; Latvin, et al, 1999; MacDonald, 2000; Patterson, 2004.)

Design of an effective battery of visual measures is challenging because of the properties of human vision. We have photographic light meters because human vision is poor at judging absolute luminances. On the other hand, human vision has some strengths relative to photometric instruments. Vision is extremely sensitive to differences of luminance in certain patterns, over a very large range of absolute luminances. The common visual-system strategy of using a difference signal to convey information greatly reduces noise and enables comparison judgments. Visual assessments also possess face validity. The user is looking at and assessing an image under visual conditions similar to the normal working environment.

1.2 Display transfer function, contrast ratio, ambient light and image quality

High quality displays should make efficient use of digital bandwidth with minimal visual artifacts. With respect to rendering of tonescale, each change of digital count should produce a visible change but one small enough that smooth spatial gradients of digital data produce smooth visual gradients.

In current practice most images are encoded with an inverse power-law transfer function, and the digital-count-to-luminance transfer function for displaying images follows the corresponding power function. On newer high-contrast displays the power function produces visible artifacts and inefficiencies of use of digital bandwidth because it is not an accurate description of the visual system's contrast discrimination properties. In the middle part of the digital range the luminance steps of the power function that correspond to single digital steps are larger than the visual threshold for detection of luminance differences. As a consequence images with smooth spatial gradients in the middle of the tonescale will likely show visible edges at each digital step. On the other hand, in the high and low parts of the digital range the luminance steps of the power function corresponding to single digital steps are small relative to the visual threshold for detection of luminance differences. In these ranges digital resolution is wasted. Differences in the digital data produce no corresponding visible differences.

There are historical arguments for using a growth function instead of a power function for both encoding and display. Weber's Law for luminance discrimination and Fechner's insight into its implications for the "logarithmic" nature of perception in threshold judgments both suggest that the transfer function should be a growth function:

$$dy/dx = ky \rightarrow y = Ce^{kx} \; .$$

Under conditions in which the law holds this would provide equal perceptual steps from digital count to digital count. Equal perceptual steps ensure the most efficient use of pixel grayscale bits in encoding, transmitting and displaying images. (This property is often wrongly attributed to a power function with gamma 2.2.)

Growth functions have their own problems as display transfer functions. Growth functions increase their curvature as overall luminance contrast (L_{max}/L_{min}) increases. The shape of a power function, on the other hand, is invariant as the overall luminance contrast of the display is changed. Also, Weber's law does not hold at the lower output levels achievable by some displays when viewed in the dark.

In film photography, it is well understood that image quality depends on the interactions between tonescale and contrast ratio. In digital imaging this interaction was largely ignored, in part due to the fixed transfer function and low contrast of early CRT displays. The luminance contrast ratio has been reported mainly as a parameter that should be as large as possible, without examining how high contrast can generate tonescale artifacts, nor how it is affected by ambient light. Now that higher contrast CRT and LCD displays are available these issues affect image quality and can no longer be ignored.

In actual work environments ambient light reflected from the display reduces the accuracy of either a power-law or a growth-function model by adding a constant luminance independent of digital data level. This luminance typically includes a relatively static component (e.g., artificial lighting reflected off static surfaces) and a variable component (e.g., daylight from windows, specular reflection of light-colored clothing). In the light, the contrast ratio of the display will be reduced, and the transfer function altered. This also means that a bright display with a relatively poor contrast ratio in the dark may have an excellent contrast ratio under ordinary viewing conditions. Conversely, a dim, very-high-contrast (in the dark) display may have a poor contrast ratio in the light. Proper display of high-quality images requires that the performance of the system in actual use be known.

1.3 The test battery

The tonescale and contrast issues described above help define requirements for a complete battery of visual measures. Display technology is changing rapidly and the required measurements may change as a result. We have already seen this in relation to LCDs. Several years ago, there were severe viewing-angle dependencies that made display characterization difficult by any method. Today the viewing-angle dependence has been greatly reduced in high-quality LCDs.

For high-quality imaging, users need to know where in the tonescale the artifacts and inefficiencies lie so they can adjust their image display strategies accordingly. At the moment the users' options are usually confined to adjusting whatever analog display controls are provided and correcting problems with reflected environmental light. The ordinary user seldom has access to controls that will alter the transfer function of current LCD displays.

Current LCDs have at least two potential problems that make it desirable to examine every digital count of the tonescale. First, in some LCDs there are local anomalies of grayscale, with some digital steps producing no luminance change and others producing unusually large luminance changes. Second, some LCDs have problems with gray tracking, with the gray at different digital counts varying sufficiently in chromaticity to produce visible color differences (Marcu and Chen, 2002, Marcu 2004). On the two high-quality LCDs used in this study, gray tracking was found to be excellent. Our observers judged the color uniformity of grays throughout the tonescale on all three displays, but no important variation was noted. In addition to LCDs, other non-CRT display technologies are under development, with their own contrast-ratio and transfer-function characteristics. Our visual test battery therefore needs to characterize displays independent of any particular physical display model.

1.4 Specific goals

Our ultimate goal is to develop a complete battery of visual measurements that

1) can be used by ordinary image users to evaluate their own equipment in their own workplace,
2) reveals in detail the capability of a display to present images with high perceptual quality, and
3) produces information that will allow a rendering system to tailor its output for highest image quality on the particular, situated display.

We want to be able to make these measurements in such a way that reasonable user effort allows widespread use in actual viewing environments. By reasonable effort, we mean that the procedures should be quick, use only easily-obtained, inexpensive, small devices, and require no special skills of the observer.

Our initial set of measures characterize the situated display system's tonescale performance. The three measures were measurement of contrast detection thresholds at every digital level in the dark and in the light, overall luminance contrast in the dark, and local curvature of the transfer function (gamma) in the dark and in the light.

We demonstrate that these measures can capture perceptually important content of the photometrically measured tonescale. In several respects the results were better than characterization based on an assumed model of display characteristics with indirect inferences to usability.


2. METHODS

2.1 Displays

We tested our procedures using three high-quality displays: an IBM T221 204-dpi LCD, an Apple Cinema HD 98-dpi LCD, and an IBM P97 114-dpi CRT display. The LCDs were brighter than the CRT. The CRT had a much higher contrast ratio in the dark than the LCDs, largely due to the very good black that it achieved. The measured diffuse ambient light reflected off the CRT was about double that reflected off the LCDs.
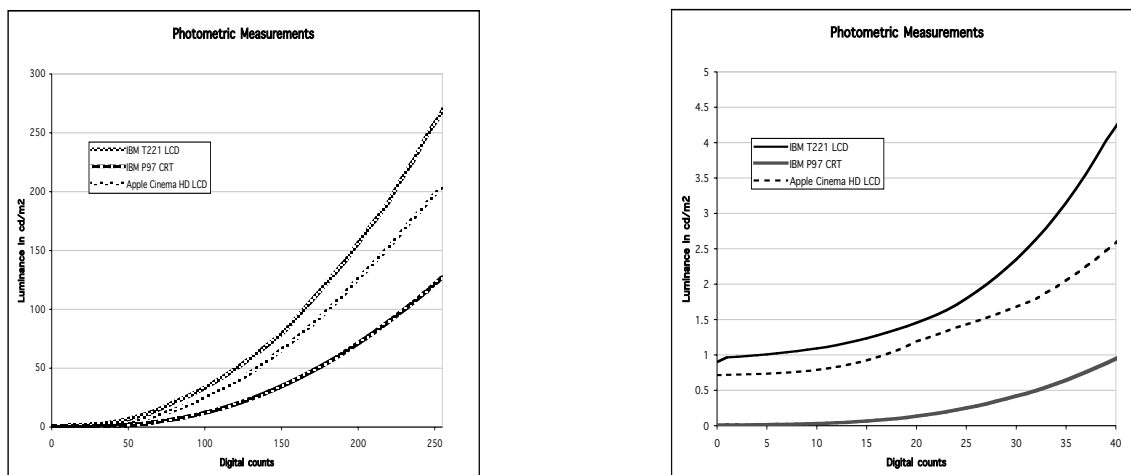
Figures 1A, 1B. Display transfer functions, measured by photometer.

Figure 1A shows the transfer functions for the three displays in the dark, with their typical power-law shapes. Figure 1B shows a detail of the same functions at the low end. Notice the larger-than-expected step between digital counts of 0 and 1 on the T221.
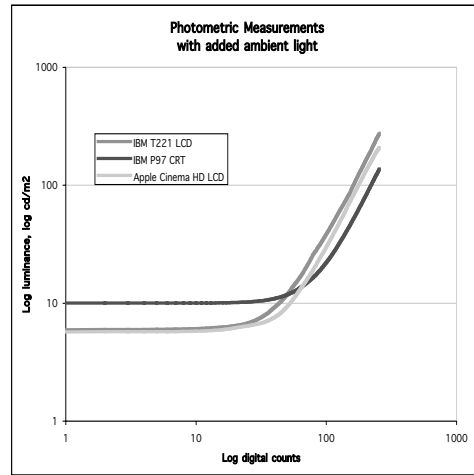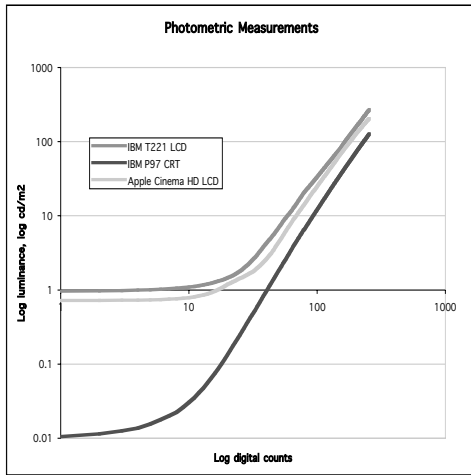
Figures 1C, 1D. Log-log plots of display transfer functions, in the dark and in the light.

Figures 1C and 1D are log-log plots of the three transfer functions in the dark and in the light, respectively. If the function were a simple power law,

$$\left(\frac{L}{L_{\max}}\right) = \left(\frac{dc}{dc_{\max}}\right)^{gamma}$$

each graph in Figure 1C would be a straight line on the log-log plot. However, a zero black is never achieved, and therefore the curves flatten out at the low end. The CRT, with the best black in the dark, elbows at a lower point than the other two displays.

At the points where the log-log graphs flatten out in the dark, at digital counts of about 25 for the T221 and the Cinema, 10 for the CRT, the measured display luminances in the dark are 1.8, 1.4, and .03 cd/m^2 respectively. These values are in the mesopic range for human vision, and therefore outside the Weber function region.

Figure 1D, again, is the log-log plot of the transfer functions for the three displays in the light. The values in the light are obtained by adding 5 cd/m^2 luminance to each LCD characterization and 10 cd/m^2 to the CRT; these are typical values within the measured range for each display in our workplace environment. The elbows for each display have moved; they are now at digital counts of about 40 for the T221 and the Cinema (luminance equal to 9.2 and 7.6 cd/m^2 respectively), and 65 for the CRT (13.8 cd/m^2), a reversal in the order.

	IBM T221 LCD	Apple Cinema HD LCD	IBM P97 CRT
Maximum luminance	270 cd/m^2	204 cd/m^2	127 cd/m^2
Contrast ratio in the dark	300:1	285:1	13,000:1
Contrast ratio in the light	54:1	41:1	13:1

Table 1. Maximum luminance and contrast ratios for the three displays, measured photometrically.

Table 1 shows maximum luminance for the three displays, and the overall contrast ratios in the dark and in the light. Notice that the huge CRT contrast ratio in the dark becomes the smallest in the light. This follows from the lower maximum luminance and greater reflectivity of ambient light of the CRT screen.

2.2 Observers and environment

Of the five observers in this study, two were in their twenties and three were over fifty. All but one required corrective lenses in order to make the judgments.

Viewing was arranged to approximate normal office and laboratory desktop working conditions. Viewing distance was not controlled, but observers sat at an ordinary working distance of about half a meter from the displays. Some judgments were made in the dark (the lights were turned off in the windowless room), and others were made with the lights turned on. Lighting was from ceiling fluorescent fixtures. The displays were in typical office positions, on a desk at a height comfortable for office work, behind keyboard and mouse. Observers had visually adapted to the lighting environment, lights on or off, for at least 5 minutes prior to observations.

2.3 Observer tasks

We assembled a battery of observer tasks that we felt would capture most of the information about the display's grayscale that is relevant to image quality. The three tasks were chosen to be practical for use by an individual in the workplace, but with some elaborations for our research purposes.

2.3.1 Luminance-contrast detection throughout the tonescale.
This task was designed to give us detailed information about visibility of differences in digital image values throughout the range of digital values. Figure 2 shows a portion of our test image. Large circles were placed on background vertical strips that sampled the entire range of digital values. The circles had digital values ranging from background+1 to background+8. Observers reported the smallest detectable incremental digital count on each background strip, both in the dark and in the light, on all three displays.
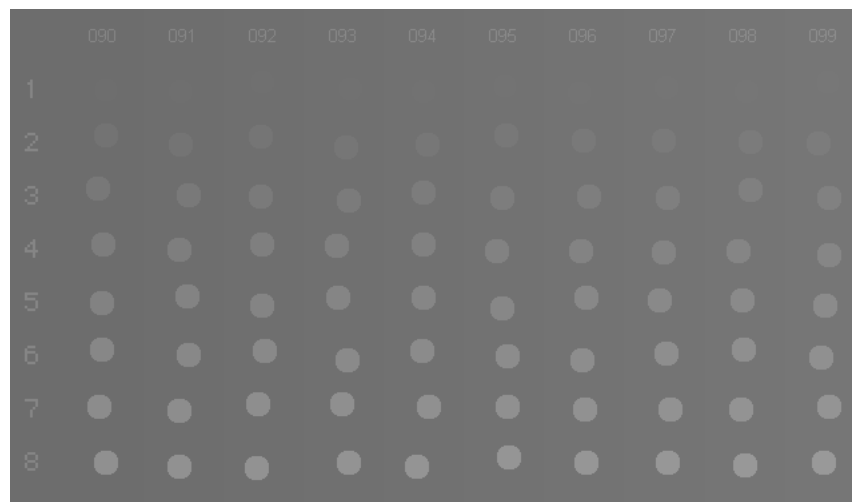

Figure 2. Example of a contrast-detection task screen, with exaggerated contrast

In pilot work we used square test patches that were aligned both vertically and horizontally, but found that phantom squares from subjective contours made the judgments difficult. Changing the patches to circles and slightly misaligning them randomly made the task much easier.

The size of the test patches governs which aspects of image quality will be tested. For this study we chose to use a large-sized patch because it reveals the banding artifacts that occur in smoothly-shaded parts of images when steps of one digital count are too large. We know from prior research that this will overestimate the perceptibility of small details in parts of the tonescale where luminance steps are too large for smooth shading. For this reason the data

should be considered a lower bound estimate of the detectability of details. We intend to investigate the visibility of small details in future work.

The background strips ranged in digital counts from 0 to 254. The strips were presented in 31 screens of 10 adjacent strips (there was a 2-strip overlap with previous and succeeding screens at either end of a screen) and 1 screen of 7 strips, in ascending order. This degree of detail served our research goal of analyzing the information captured by the task and proved useful in detecting local anomalies. Pilot work also showed that a pattern with just 16 background values takes little time and effort and captures much of the overall information about the tonescale.

The size of the screens in this task varied somewhat from display to display, because the presentation application was set to display full screen, and the physical sizes of the full screens varied. Since viewing distance was not controlled, observers were free to adjust their view as needed for optimum performance. Therefore, the retinal sizes of the patches could vary. The patches were large enough, however, that judgments were equivalent across displays and under the free-viewing conditions.

Knowing the photometric characterizations of a given display, or assuming a power-law transfer function and a good contrast ratio, we expected to find that the increment detection judgments would not be uniform, but would be low in the midrange of digital counts (banding artifacts) and would increase at the high end (wasted levels). We also expected wasted levels at the low end, worse in the light than in the dark. We expected the interference of the ambient light on judgments at low digital counts to become negligible at some point, so that the judgments in the light and in the dark would become the same, and the effects of the ambient could be disregarded. These general expectations follow from a simple percent luminance change calculation, as discussed below.

Our expectations for the specific displays of this study also included capturing the relatively sharp drop in luminance at dc = 0 for the T221, and the greater influence of the ambient light on the CRT compared to the LCDs up to about dc = 100.

The contrast detection task used in this study is not a criterion-free method. That is, an observer's willingness to judge that they see a small difference is not separated from their sensitivity to differences; an observer may be conservative, choosing an increment level where the difference is clearly visible, or more liberal, willing to judge that they see a difference that is quite borderline. Also, an observer's criterion may shift as they progress through the screens that make up the test.

2.3.2 Gamma Measurement.
We employed a widely-used matching task to estimate the gamma of the displays in both the light and in the dark (Figure 3). Observers chose which of several uniform grays matched the brightness of a black-and-white halftone pattern when viewed from a distance that optically blurred the halftone to a uniform appearance. Assuming a power-law transfer function, there is a functional relationship between the exponent, gamma, and the digital count required to produce the luminance of the blurred halftone. The digital count that corresponds to the actual gamma of the display produces the same luminance as the blurred halftone. Digital counts for incorrect gammas produce higher or lower luminance grays.
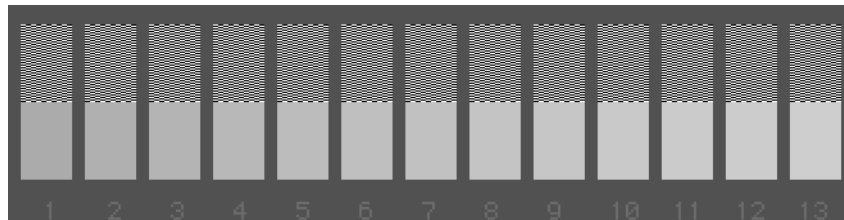

Figure 3. Example of a gamma tester, at .50 luminance.

If the transfer function were exactly a power function, the halftone could be assigned any ratio of black pixels to white pixels, provided only that the digital counts for the various gammas be sufficiently separated to allow the

visual judgment. Prior tests have used a halftone with a ratio of three black pixels for each white pixel, i.e., at a normalized display luminance of 0.25. The curves for various gammas are widely separated at the 0.25 point, which should allow accurate and consistent judgments.

Since our prior work showed that transfer functions are typically not exactly power functions, the matching task may be thought of as providing a statistic describing the curvature at the 0.25 point. We decided to evaluate two other points on the transfer function as well as the 0.25 point. Our normalized luminances were 0.25, 0.50, and 0.75, with ratios of black pixels to white pixels of 3:1, 1:1 and 1:3. Since the transfer functions are less separated horizontally at 0.50 and 0.75, the uniform grays in the test patterns were closer together in luminance, which should make the judgment more difficult.

We also used two physical methods for deriving estimates of gamma from the measurement of the transfer functions of the display. The first was to fit the measured transfer function to the power-law equation

$$L = (L_{max} - L_{min})(dc/dc_{max})^{\gamma} + L_{min}$$

and the second was to fit a line to the linear portion of the log-log plot of the transfer function (the slope of the line is an estimate of gamma).

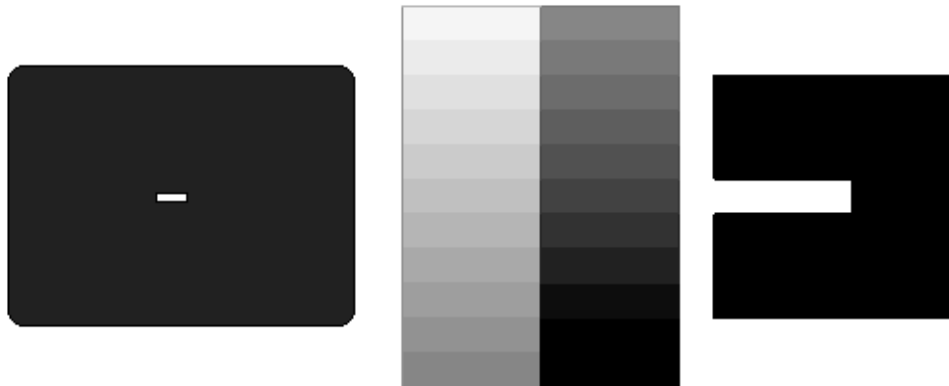2.3.3 Contrast ratio measurement

Figure 4ABC. White rectangle on screen, step wedge and background mask (screen not at the same scale)

We used a photographic step wedge with a series of densities (Stouffer transmission projection step wedge, a series of neutral density filters), placed nominally 1/2 f-stop apart; i.e. each step divided the light further by $\sqrt{2}$ (Figure 4B). The wedge was mounted in a black cardboard tube that reduced reflections from the front of the filter. The observer held the wedge against the display face, with a single step covering a white rectangle of the same size and shape (Figure 4A), and compared its brightness to that of an adjacent unfiltered rectangle (Figure 4C). The unfiltered background area was masked by an opaque cardboard aperture to make a rectangle of the same shape and size as the filtered rectangle. The observer slid the various steps of the wedge filter over the white rectangle to find the filter giving the best brightness match to the unfiltered background. The task was repeated with several unfiltered background levels, providing luminance ratios, gray:white, for several gray levels on the display's transfer function.

The actual physical densities of the wedge steps were measured by placing them against the white rectangle on the T221 display and measuring the resulting luminances with a Minolta LS-100 photometer.

2.4 Bootstrapping: Reconstructing the transfer function of the display using the contrast detection data and the contrast ratio estimates.

We wanted to know how much of the information that we get from a full photometric characterization of the display can be captured using only our battery of visual tasks. One method is to try to reconstruct the photometric transfer function from the visual data. We attempt this here, but it should be noted that this reconstruction is not part of evaluating the visual quality of the display. The reconstruction is for research analysis only. In practical use the visual measures themselves describe the visual quality of the display.

Using an argument based on Weber's Law, we devised a simple bootstrapping method for reconstruction of the display's transfer function using only the data from our contrast detection and contrast ratio estimation tasks. The contrast threshold task provides a measure of the contrast threshold (in digital count) at each output level (also in digital count) of the display. The contrast ratio estimates provide a measure of the luminance range spanned by the digital count range. If each Just Noticeable Difference (JND, in digital count) corresponds to a known constant proportion of the luminance at that point in the digital count range, we can construct the normalized luminance curve by multiplying up from 1.0 JND-by-JND. The contrast ratio estimates provide the known constant proportion, p, by the following argument:

A luminance contrast detection judgment of 1 digital count = 1 JND between adjacent digital counts; a contrast detection judgment of 2 digital counts = 0.5 JND between adjacent digital counts, etc. Therefore the total number of JNDs, J, over the full range of digital counts is:

$$J = \sum \frac{1}{t(d)} \ , \ \ d = 0 \text{ to } 255$$

where J is the total number of JNDs and $t(d)$ is contrast threshold in digital count increments (the observer's judgment) at each digital count d.

By Weber's Law, each JND represents a constant percent increase, p, in luminance, so that each JND step is a factor of $(1 + p)$. If we normalize the minimum luminance of a display to a value of one, the maximum relative luminance will equal the contrast ratio, C. Since the maximum relative luminance also represents J JND steps above the minimum, one, the following relationship must hold:

$$C = (1 + p)^J .$$

Solving for p, we derive:

$$p = e^{\frac{\ln(C)}{J}} - 1 .$$

Thus we can use our contrast estimation and contrast detection tasks to estimate C and J, respectively, and to derive an estimate of p. For the five observers, estimates of p ranged from 1.5% to 3%, consistent with classic luminance difference detection data.

The relative luminances for other levels can be derived through iteration, once we have an estimate for p:

$$\lambda(d_i) = \lambda(d_{i-1}) \left(1 + \frac{p}{t(d_{i-1})} \right)$$

where $\lambda(d)$ is the relative luminance at digital count d. We can evaluate this approximation of $\lambda(d)$ by comparing it to the normalized transfer function from our photometric measurements.

Several factors will contribute error to our bootstrapping procedure:

    1) Weber's Law doesn't hold at low luminances; threshold contrast is greater than at higher luminances.

    2) Observer's criterion may not be constant over the entire contrast detection task.

    3) The contrast estimation task has coarse steps (1/2 f-stop = 40% increase).

    4) The contrast detection task can't measure thresholds smaller than one digital count.

3. RESULTS

Our tasks are intended to eventually be used by individual observers, in single sessions, to characterize their display system at that moment, in their work setting. Accordingly, we are interested in whether results for individuals (as opposed to averages over observers) capture the important aspects of display performance.

3.1 Luminance-contrast detection throughout the tonescale.

All of our observers, both experienced and naive, found the contrast detection task easy to perform under all of the conditions. Younger observers differed from older mainly in setting higher criteria for differences (this was an unexpected result). The pattern of results was the same for all observers; three examples for a single observer are shown in Figure 5.
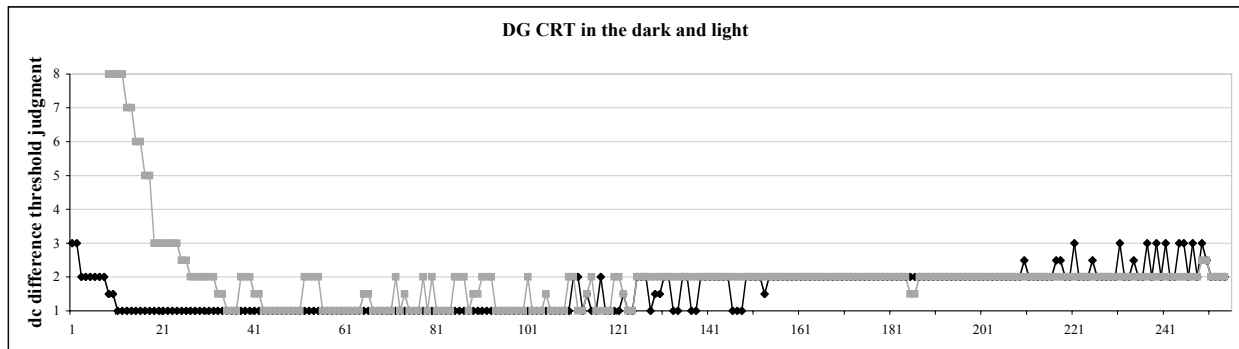

Figure 5A. Threshold differences of digital count as a function of background digital count. Gray symbols: lighted room. Black symbols: darkened room. IBM CRT.
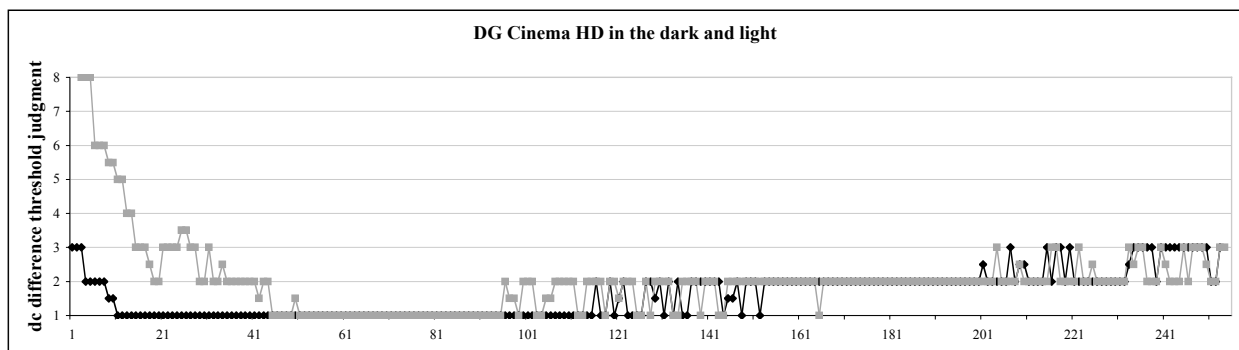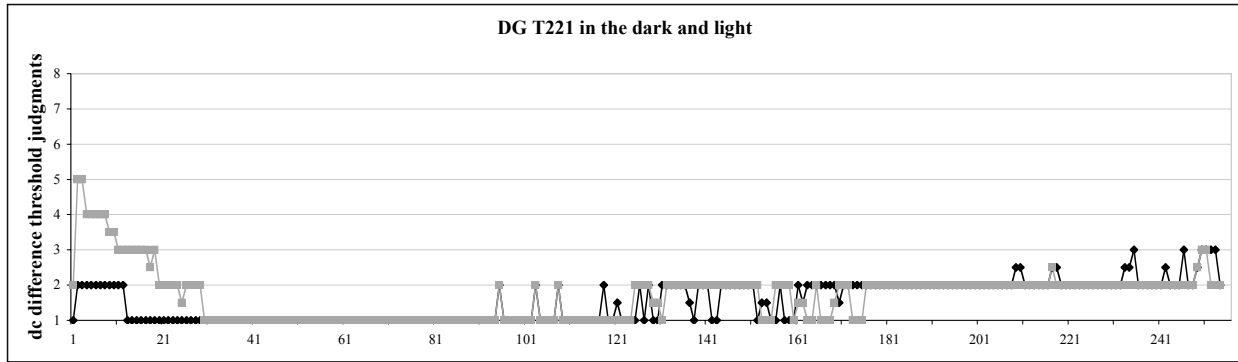

Figure 5B. Same legend. Apple Cinema HD.

Figure 5C. Same legend. IBM T221.

Low JND numbers indicate large perceptual steps (increments) between adjacent digital counts, and high numbers indicate small steps. While quick and easy, the task was sensitive enough to show many of the differences we predicted.

1) Examining the graphs in detail, we can see the predicted effects of the perceptual non-uniformity of the power-law transfer function. Single count differences were visible through the middle part of the range, but multiple counts were required at the low and high counts. This means that smooth gradients in the middle of the tonescale will likely show visible edges at each count. Conversely, digital resolution is wasted at the top and bottom of the tonescale: differences in the digital data produced no corresponding visible differences. It is likely that steps of even less than one count could have been detected through part of the midrange had our stimuli included halftones.

2) The unusually large luminance increment between digital-counts zero and one on the T221 display was easily detected (Figure 5C). As a feature, it is much more prominent in contrast detection than a quick examination of the physical measurement of the transfer function would indicate.

3) Reflected light had the predicted effects. Thresholds were higher in the light than in the dark but only at low luminances. The data show that the ambient illumination had no effect on contrast detection above digital counts of about 40 for the LCDs, and extending somewhat further for the CRT.

4) The CRT display showed larger effects of reflected light than the LCD displays.

3.2 Gamma Measurement.

As in our previous paper (Gille and Larimer, 2001), we found perceptual estimates of gamma that were consistent across observers and viewing conditions (Table 2). The task was easy and gave consistent estimates for both the 0.25 and 0.50 normalized luminance patterns. All observers complained that the judgment for the 0.75 normalized luminance patterns was too hard, as there was little or no visible difference among the comparison grays above the level corresponding to gamma = 2.2. Estimates of gamma from physical measurements were less consistent than the perceptual judgments.

	Average perceptual judgment		Gamma estimated from simple power-law fit	Gamma estimated from slope of log/log plot
	dark	light		
IBM T221	2.2	2.2	2.3	2.3
Apple Cinema	2.2	2.2	2.1	2.2
IBM CRT	2.3	2.3	2.4	2.5

Table 2. Gamma estimates for the three displays using the perceptual judgment in the dark and in the light, and two methods based on the photometric data.

3.3 Contrast ratio measurement

Our results using our contrast ratio device were mixed (Figure 6). For the LCD displays, observers were able to do the task with good consistency and agreement with the photometric measurements. This was true for all five levels (luminances) of the unfiltered area. For the CRT, the visual estimates were lower than the photometric measurements, especially for the darker grays.

The comparison steps were coarse (40% difference between steps) by basic research standards, and judgments were more consistent when the contrast ratio fell at a particular step rather than between steps. Nevertheless, on the LCDs, the judgments provided information that we were able to use for reconstructing the relative transfer function. To meet our standards of usability in the workplace, the contrast ratio test needs further development.
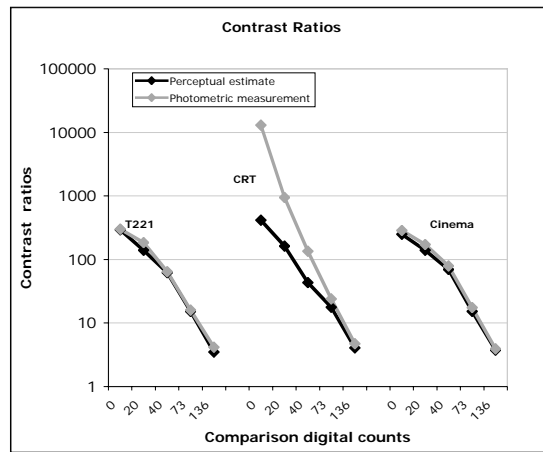

Figure 6. Contrast ratios measured visually and by photometer.

We did not systematically investigate why the CRT measurements were less accurate than the LCD measurements, but one obvious visual difference between the two types of display was substantial blurring of the edges of the white bar on the CRT when viewed through the filter. This scatter may have reduced the actual photometric contrasts when viewed through the neutral density filter.

3.4 Bootstrapping: Reconstructing the transfer function of the display using the contrast detection data and the contrast ratio estimates.

We compared the normalized transfer functions reconstructed as described above from the contrast detection and contrast estimation tasks to the corresponding normalized photometric transfer functions. The results matched quite closely when the contrast ratio was accurately judged. This was in spite of the error factors listed above. For some of these reconstructions, the transfer function was closely recoverable, with good agreement among observers (Figure 7).

If the contrast ratio estimate was inaccurate, as with the data in Figure 8, the relative transfer function could not be recovered. When the inaccurate contrast ratio was the only problem, the shape nevertheless was correct. For the T221 using digital counts 73 to 252 there was again good agreement among observers.

In Figure 9, the estimated transfer functions had a different problem. The contrast threshold task judgment scale (background+1 to background+8) was too coarse. All the observers made judgments of "one" throughout the range from digital counts 10 to 100, but comparison with the photometric curve reveals that the increments were much larger than one JND. That is, the shape of the transfer function for the observers is distorted in that region, and the distortion is propagated throughout the function. However, there is still good agreement among observers for this condition.
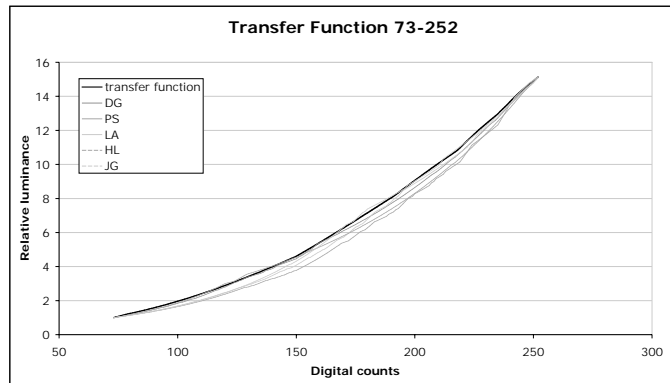
Figure 7. Normalized transfer functions derived from the visual contrast detection and contrast estimation data (dotted lines) and from photometric measurements (solid line). Data from the T221 display; judgments on digital counts from 73 to 252.
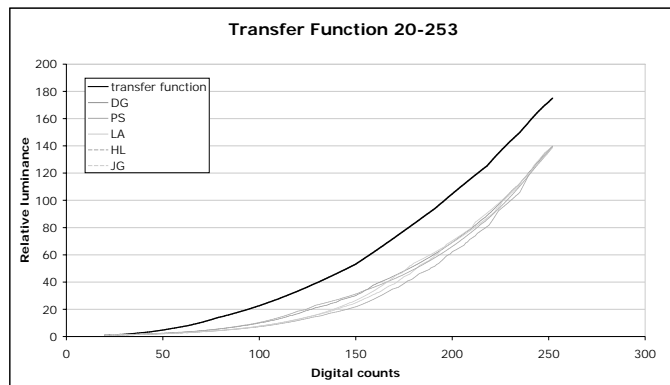

Figure 8. Same as Figure 7; judgments on digital counts from 20 to 253.
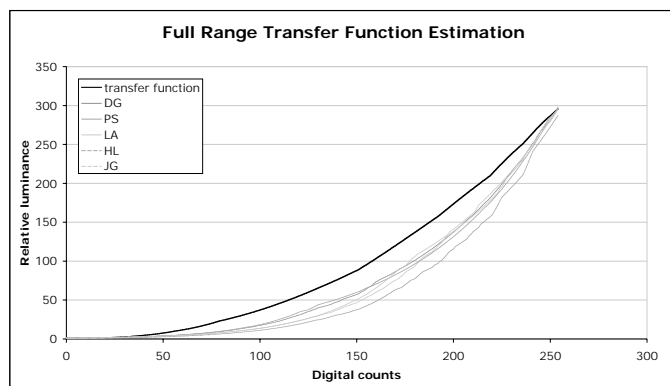

Figure 9. Same as Figure 7; judgments on digital counts from 0 to 254.

3.5 Estimating the gamma of the display using the contrast detection data and the contrast ratio estimates.

The gamma of a display with a power-law transfer function can also be estimated from our bootstrapped relative transfer function derived as above. The slope of the linear portion of the log/log plot is an estimate of gamma. An example using the perceptually derived function from the CRT is plotted in Figure 10; the gamma estimate of 2.35 derived from a linear fit is in accord with the perceptual judgments for the display.
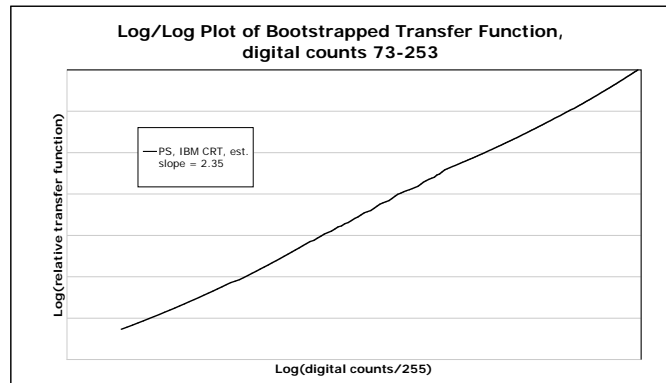
Figure 10. Log-log plot of one of the bootstrapped transfer functions; slope = 2.35, an estimate of gamma.

3.6 Summary

The luminance-contrast detection task throughout the tonescale is simple to perform. It gives data sufficiently detailed to show regions of inefficient use of bandwidth, regions with probable banding artifacts, local anomalies of the tonescale, and the effects of reflected ambient light.

The visual gamma measurements confirmed that the visual task gives results at least as reliable as those derived from photometric measurements and without the complications of device modeling.

The contrast ratio measurement is a work-in-progress, giving good results under some conditions and inadequate under others. Different filters may solve some of the problems.

The bootstrap reconstruction of the digital-data-to-luminance transfer function from our visual measures showed that they are capable of capturing all of the shape information contained in photometric measurements, provided that two issues can be resolved. (1) Minor improvements of the contrast sensitivity task will allow measurement of thresholds of less than one digital count. (2) The contrast ratio measurement needs more work: it needs to give accurate results under all conditions.

Together the results show that this set of tasks can provide adequate characterization of the tonescale of displays once the above problems are solved.

4. DISCUSSION

Image quality is judged by eye. It depends on the properties of the source material and the encoding and rendering of that material. Encoding and rendering almost always result in a loss of information, and it is of course desirable that such losses are not *visible*. The rendering step is constrained by what comes before it, but certainly one would hope to have a visually efficient rendering, and to avoid the introduction of new artifacts caused by display characteristics.

We have argued that, for the user in the workplace, a direct visual measurement of display characteristics will necessarily be better than one based on an instrument measurement coupled with indirect inferences from psychophysical models, even if one can be had. A direct visual measurement can simultaneously account for display anomalies, the working environment, and user characteristics. If the direct visual measurements are such that they can be coordinated with the rendering intent that guided the encoding, a superior image must be the result.

Image encoding schemes for electronic displays have traditionally been tightly coupled to an understanding of the properties of those rendering machines and to storage and transmission issues (file size and channel bandwidth). Historically, the transfer function for displays was set to be a power function, for several reasons. A power function was easy to generate in the hardware of the CRT display and provided a convenient manipulation to enhance image contrast on early, dim, low-contrast CRTs (partial gamma correction). For many years, eight-bit grayscale encoding based on a power-law scheme was accepted for most purposes on most displays.

However, as our contrast detection data showed, on current CRTs and LCDs the eight-bit power-law transfer function produces banding artifacts at mid-range digital counts, and wasted bits at the high and low ends. This is another argument in favor of current activity in the imaging standards community to rethink the number and luminance spacing of bits required for high-quality image encoding.

Even if an image is perfectly encoded (no loss of information), it is necessary to have a characterization of the rendering display that is complete enough to allow the system or the user to adjust settings and perform image processing (such as halftoning or contrast enhancement) in order to achieve the desired result. Essential elements to a complete characterization include the relative shape of the transfer function, the perceptual dynamic range, and local anomalies in the tonescale.

The relative shape of the transfer function, or tonescale, for displays conforming to the power-law transfer function is usually summarized by the parameter gamma. Gamma can be estimated by eye, as this and other studies have shown. However, the power-law shape as realized in actual systems also requires an offset parameter that is not part of the "gamma" measurement, and varies with the lighting conditions. This is the reason for the flattening out of the log-log transfer function plots in Figures 1C and 1D. Thus, although estimating gamma provides some information about the tonescale, it is not a complete specification of the relative shape of the transfer function.

The maximum brightness and the overall contrast ratio (in the dark) are often cited in display specifications. Neither of these is a direct measure of perceptual dynamic range, although they are correlated with it, and have value in the comparison of displays. In addition, there is currently no widespread, simple method of estimating either of these parameters by eye. They are important for tracking display changes over time, for predicting regions that will have banding artifacts (when combined with tonescale), and for image processing such as contrast adjustment when the encoded image originated with a rendering intent different from what is native to the display.

Local anomalies in the tonescale can only be assessed locally. Idealized parameters such as gamma cannot characterize them.

In this study, we were successful in finding simple tests that can be used by ordinary image users to evaluate their own equipment in their own environment and that produce information that would allow a rendering system to tailor its output for highest image quality. Our contrast detection and contrast ratio tasks produce information about the relative shape of the transfer function throughout the entire range of the display, and incorporate the effects of the lighting conditions, allowing for the proper mapping of the encoded image to the display. Banding artifacts are identified directly. The gamma estimate as it would be measured by eye can be derived directly from the contrast detection and contrast ratio task data. Local anomalies are revealed by the detection judgments, although identifying non-monotonicities would require an astute observer.

Our next step is to refine the current tasks, and then to identify new tasks that can add important independent information about display characteristics. The first refinement needs to address the problem that the one-digital-count steps in the contrast detection stimuli were too coarse throughout much of the tonescale. Some judgments of "1" were true threshold values, the dots being just visible against the background (1 JND); others represented very obvious, easy-to-see differences (3 or more JNDs). This difficulty can be overcome easily by using a simple halftoning method to create dots that are midway in luminance between their component levels. Second, now that the step-wedge contrast ratio judgments have been shown to be viable measures of actual contrast ratios, a more systematic method for choosing the levels at which to test, based on contrast detection results both in the light and in the dark, needs to be developed.

An important dimension of perceptual display performance is the visual quality of small image features. Information about the relationship between feature size and visibility can be derived by adding dot size as a factor to the contrast detection task. For smaller dots thresholds will be larger than those measured here (Graham and Bartlett, 1940; Blackwell, 1946; van Nes and Bouman, 1967).

One of the strengths of the current tests is that they can identify display problems for the user. Some problems, such as excessive reflections of ambient light or poor settings of the display's controls, can be corrected by the user. Others, such as an inherently poor transfer function shape, must be addressed by the software, or ultimately in display manufacture. Our visual characterization tasks provide tools that can deliver information to the user for managing the aspects of image quality determined by the transfer function. Simple, reliable visual tests of display performance support the development of applications that allow the optimization of displays in the workplace.

5. REFERENCES

Blackwell, H.R. (1946). Contrast thresholds of the human eye. *J. Opt. Soc. Amer.*, 36, 642-643.

Gille, J., & Larimer, J. (2001). Using the human eye to characterize displays. *Proceedings of the SPIE, 4299*, 439-454.

Graham, C.H., and Bartlett, N.R. (1940). The relation of size of stimulus and intensity in the human eye: III. *J. exp. Psychol.*, 27, 149-159.

Latvin, Y., Silverstein, A., & Zhang, X. (1999). Visual experiment on the web. *Proceedings of the SPIE, 3644*, 278-289.

MacDonald, L. W. (2000). Assessment of monitor calibration for internet imaging. *Proceedings of the SPIE, 3964*, 162-167.

Marcu, G. G. (2004). Gray tracking correction for TFT-LCDs. *Proceedings of the SPIE, 5293*.

Marcu, G. and Chen, K. (2002). Gray tracking correction for TFT-LCDs. *Proc. IS&T/SID Tenth Color Imaging Conference*, 272-276.

Patterson, D.R. (2004). Personal communication. In the 1990s the National Information Display Laboratory, Princeton, NJ, developed Softrak, a program that allowed users to quickly measure aspects of their CRT display performance and store the results for comparisons over time. The measurement tasks included resolution at various contrasts, and coarse measurement of contrast detection through the tonescale.

Van Nes, F. L., & Bouman M. A. (1967). Spatial modulation transfer in the human eye. *J. Opt. Soc. Am.* , 57:401-406.