Comparing Signal Detection Between Novel High-Luminance HDR and Standard Medical LCD Displays

M. Dylan Tisdall, Gerwin Damberg, Nhi Nguyen, Yan Tan, M. Stella Atkins, Hiroe Li, and Helge Seetzen

Abstract

DICOM displays linearly space digital data values over the range of just-noticable differences (JNDs). To increase the number of JNDs available we must increase the display's contrast. However, operating over too wide a range may cause human observers to miss contrast in dark regions due to adaptation to bright areas or, alternatively, miss edges in bright regions due to scattering in the eye.

Dolby Inc.'s new high dynamic range (HDR) LCD display has a maximum luminance over 2000 $\mathrm{cd/m^2}$; bright enough to produce significant in-eye scatter. The display combines a spatially varying backlight allows a low-resolution 8-bit "backlight image" with a high-resolution 8-bit LCD panel, approximating a 16-bit greyscale display. Alternatively, by holding the backlight constant at 800 $\mathrm{cd/m^2}$, a standard medical LCD display can be simulated.

We used two-alternative forced choice (2AFC) signal-detection experiments to quantify display quality. We explored whether the full-power HDR display's optical characteristics (scattering and low resolution backlight) have a negative effect on signal detection in medical images compared with standard medical LCDs. We used 8-bit test images derived from high-field MRI data combined with synthetic targets and synthetic Rician noise.

Initial results suggest that the HDR display allows signal detection comparable to a standard medical LCD.

M. D. Tisdall, N. Nguyen, Y. Tan, and M. S. Atkins are with the School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6 Canada e-mail: mtisdall@cs.sfu.ca.

G. Damberg, H. Li, and H. Seetzen are with Dolby Inc., Vancouver, BC, Canada

Manuscript received x; revised y.

Comparing Signal Detection Between Novel High-Luminance HDR and Standard Medical LCD Displays

I. Introduction

GREYSCALE medical image displays rely on the observer's sensitivity to spatially varying luminance in order to communicate a 2D array of digital values. Given an LCD display that can produce a finite set of greyscale luminance values, the DICOM standard formalizes a function for selecting the appropriate luminance for each pixel to best represent some digital value in a medical image [1]. The core principle behind the choices suggested by the DICOM standard is that equal differences in digital values should be represented by equal perceptual differences. Thus, it proposes units of just-noticeable differences (JNDs) as the perceptual equivalent of the digital data's units. A mapping between luminance and JNDs is established in the DICOM standard based on previous human experiments. Using this relationship, we can convert our digital values into luminances by ensuring that equal steps in the digital domain are mapped to equal steps in the JND space and determining the relevant luminances from the desired JND values.

In practice LCD-based displays can achieve a finite range of luminances. The ratio of the maximum to minimum luminances is generally referred to as the display's contrast ratio. Further they have limited discrete luminance values inside this range that are available for display. The base-2 logarithm of the number of discrete luminance settings is called the display's greyscale bit depth. The minimum and maximum brightness provide an upper-limit on the number of JNDs that a human could perceive on a perfectly controllable display. The bit depth determines how well we can approximate this ideal display. Previous work has suggested that, for regular medical displays with maximum and minimum luminances of approximately 900 cd/m^2 and 1.5 cd/m^2 respectively, there is little value in producing monitors with more then 12-bit greyscale bit depth [2].

Dolby has demonstrated a new high-dynamic range (HDR) LCD-based technology that allows displays to have effectively infinite contrast ratio by having the minimum luminance of the display become very close to zero. Medical LCD displays normally use a uniform backlight that provides approximately equal illumination to the back of the LCD panel at every pixel. The LCD panel is then used to filter this light. However current LCD technology cannot block all the light, even when the LCD is set to full black. Thus, on a normal LCD the minimum luminance level is some value greater than zero. The new Dolby display technology uses a spatially varying backlight to illuminate a standard LCD panel. Because the backlight can vary spatially, it is possible to turn it off completely in regions where the image should be black, making for regions with effectively zero luminance. Furthermore, the Dolby technology relies on high-power light-emitting diodes (LEDs) for the backlight, making the maximum luminance of displays in the thousands of cd/m^2.

However, the Dolby HDR LCD also introduces some compromises compared to a standard LCD. The spatially varying backlight system cannot be controlled individually at each pixel in the image. Instead, a low-resolution array of backlight LEDs is used and the illumination behind the LCD at any location is the sum of the contributions from all the LEDs whose point spread functions (PSFs) extend to that location. Thus while the backlight LEDs each individually have 8-bits of luminance depth and the LCD panel also has 8-bits of greyscale depth, the resulting display does not have 16-bits of independent greyscale depth at every pixel. Instead, we have an approximation to a 16-bit display where neighbouring pixels' luminance values are coarsely correlated.

The low-resolution backlight is partially justified by the imperfect nature of the human optical system. In particular, light scattering in the media of the eye causes bright regions to be blurred [3]. This is commonly observed as a "blooming" or "halo" effect where a bright region with a sharp edge abutting a dark region will have a halo that extends over the edge. In practice, this scattering-induced halo will be larger than the PSF of the LED, meaning that the approximation artifacts from the Dolby technology are less than the dominant source of error in the human eye [4]. However, since this blooming effect can obscure fine details and edges, it may be that there is still an effective upper limit on the brightness that is useful in medical displays.

We were interested in determining whether the artifacts introduced by the low-resolution Dolby backlight, combined with the potential effect of scattered light, would impact the use of these screens in a medical context. To this end, we have conducted an experiment based on a two-alternative forced-choice (2AFC) signal known exactly (SKE) signal-detection task. To control the effects of the varying backlight and scattered light, we tested the Dolby display in two configurations. In the first, we made full use of the brightness and spatial variation available from the backlight. In the second configuration we set the backlight to be spatially uniform and produce a maximum display luminance of 800 cd/m^2; approximately the same luminance as a high-end medical display. We then compared task performance between these two conditions.

In the section II we will describe the process used to produce our stimulus images, provide a description of the 2AFC task

that our subjects performed, and provide more details about the display and how we used it. In section III we present the results of our experiments, and discuss their implications for use of the Dolby display. Finally, in section IV we present our conclusions.

II. Methods and Materials

A. Stimulus Images

Our stimulus images were generated using a similar methodology to previous work on the evaluation of MRI reconstruction [5]. Our goal in using anatomical MRI backgrounds was not to simulate a realistic pathology, but instead to provide a realistic background that would stimulate the contrast sensitivity of the observer in the same way a real medical image would. This provides a visual distraction effect similar to that of real medical images.

We began with several 16-bit magnitude-reconstructed 3D inversion recovery head MRI volumes of healthy volunteers acquired on a 3 T Philips Gyroscan Intera scanner. The volumes were sliced along the three major axes to produce a corpus of full-head images. From the full-size images, 128×128 pixel images were constructed along the three major axes by selecting 128×128 pixel regions randomly from the full-size images. To verify that our small images contained anatomy in the central part of the image, we computed the average intensity in the central 64×64 pixel sub image and ensured it was above a sufficient threshold. Images that were over the threshold were normalized to the range $(0, 1)$ and kept as backgrounds.

Our backgrounds were randomly divided into target-present and target-absent sets. Images in the target-present set were summed with an anti-aliased circular target signal defined by the function

$$S(\boldsymbol{x}) = \begin{cases} b & \text{if } |\boldsymbol{x} - \boldsymbol{z}|_2 \leq w \\ b(1 - \|\boldsymbol{x} - \boldsymbol{z}\| + w) & \text{if } w < \|\boldsymbol{x} - \boldsymbol{z}\| < 1 + w \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\boldsymbol{x}$ is a 2D coordinate in image space, b is the amplitude of the target signal, $\boldsymbol{z}$ is the index of the image center, $\|\cdot\|$ is the Euclidean norm, and w is the width of the feature. We set $w = 3$ which was approximately equivalent to a 6 mm feature in the anatomy.

To simulate Rician-distributed thermal MRI noise [6] in our target-present and target-absent images, we produce two random samples from a Gaussian distribution $\mathcal{N}(0, \sigma)$ for each pixel in each of our synthetic images. Let $B(\boldsymbol{x})$ be the intensity of a given anatomical background image at location $\boldsymbol{x}$, $S(\boldsymbol{x})$ be the intensity of the target signal at location $\boldsymbol{x}$, and $Q_1(\boldsymbol{x})$ and $Q_2(\boldsymbol{x})$ be the two samples from the Gaussian distribution at location $\boldsymbol{x}$. We can then write the final target-present image with simulated thermal noise as

$$I(\boldsymbol{x}) = \left[\left(B(\boldsymbol{x}) + S(\boldsymbol{x}) + Q_1(\boldsymbol{x}) \right)^2 + Q_2(\boldsymbol{x})^2 \right]^{-1/2}, \tag{2}$$

and target-absent images are simulated with

$$I(\boldsymbol{x}) = \left[\left(B(\boldsymbol{x}) + Q_1(\boldsymbol{x}) \right)^2 + Q_2(\boldsymbol{x})^2 \right]^{-1/2}. \tag{3}$$

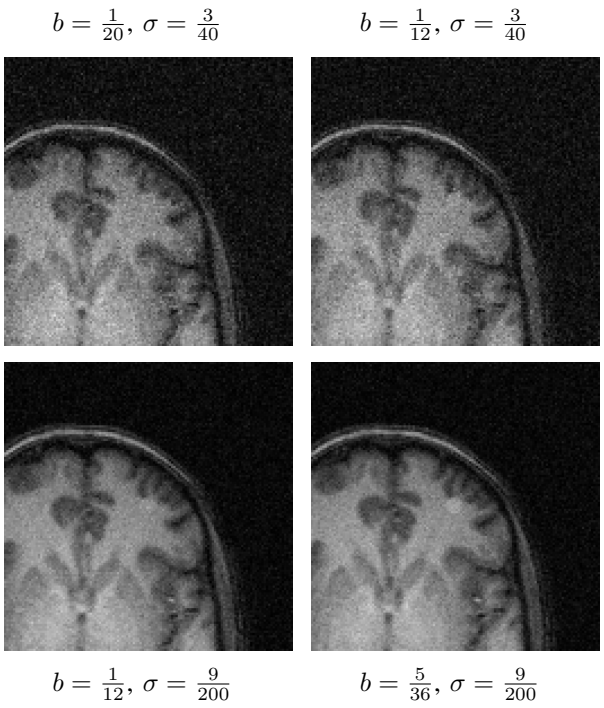

Fig. 1. Example of one anatomical background in all four target-present conditions. The target signal is the small circle visible just under the cortical folds, in the center of the images.

When adding signals and noise, we subdivided our images into four conditions representing four different target signal-to-noise ratios (SNRs). Using the variables specified above, these four conditions were $(b = \frac{1}{20}, \sigma = \frac{3}{40})$, $(b = \frac{1}{12}, \sigma = \frac{3}{40})$, $(b = \frac{1}{12}, \sigma = \frac{9}{200})$, and $(b = \frac{5}{36}, \sigma = \frac{9}{200})$ which give target SNRs of approximately $\frac{2}{3} \simeq 0.667$, $\frac{10}{9} \simeq 1.111$, $\frac{50}{27} \simeq 1.852$, $\frac{250}{81} \simeq 3.086$. In practice, because our target was summed on top of background anatomy, it was usually far brighter than indicated by these SNR values. However, these values are indicative of the degree of contrast between the target and the anatomy it was summed with relative to the simulated thermal noise. Additionally, it is important to note that there is some real thermal noise already present in our background images $B(\boldsymbol{x})$. However, because of the quality of the scans used, this noise's variance is far less than that of our simulated thermal noise, and thus we do not expect it had any impact on the final results.

Once the signal and noise was added, the entire image corpus was normalized so that the darkest pixel over all the images was set to 0 and the brightest pixel over all the images was set to 255. Thus, most images spanned some slightly smaller range of values. We then stored the final images as 8-bit values. An example of one anatomical background in all four target-present conditions is shown in figure 1

Reducing our data to 8-bit values could be seen as missing the point of using a high-contrast display. Having increased the luminance range that the display can provide, we now have more JNDs available and can thus afford to show more than 8 bits of greyscale information simultaneously. Our reason for choosing 8-bit information as the final digital output for our experiment stems from the fact that, at this point, we are

interested only in testing the effects of the Dolby display's optical design and brightness. Thus, by using 8-bit data we can ensure that our images can be presented without further data reduction in both of our display conditions, which are described in a later section.

B. 2AFC Task

Signal-detection tasks have a long history in measuring the quality of imaging systems, including the evaluation of medical imaging modalities and image reconstruction algorithms [5], [7]–[9]. Our particular experiment structure – the 2AFC experiment – has also been used previously in the evaluation of medical LCD displays [10].

Our seventeen volunteers were all non-radiologists with no previous medical image reading experience. All subjects had fully corrected vision and were graduate students or university graduates. The age range (mid-20s to mid-40s) was well below the age significance threshold in the CIE General Disability Glare Equation [3] indicating that the age variation should not be a significant factor in the quantity of scatter in their eye, and thus their perception of the "blooming" effect. Similarly, eye colour of the subjects was not an important factor as the experiment was setup to have viewing angles smaller than the CIE threshold of 30 degrees at which eye color becomes significant [3]. The entire experiment was conducted in a fully darkened room, with the display being the only source of illumination. Subjects were seated on-axis both vertically and horizontally with the display and approximately 3 times the height of the display away from the screen as this is considered the optimal distance for HDTV viewing (our prototype screen was based on a restricted region of an HDTV screen).

The display was shrouded in heavy black cloth to cover reflected light from the frame and ensure that participants saw only the portion of the screen containing the interface and a border of approximately 1 inch of screen around it. Inside of this region, two images were displayed in a vertical orientation, with a gap between them in which we displayed the target feature for the trial. In each trial the two images were chosen so that one was target-present and the other target-absent. Subjects were told that, if the target was present in an image, the circular target would sum with the background to make the region brighter. They were then directed to compare both images with the target feature displayed in the center of the screen and select their best guess for which of the two images was target-present. To ensure that there was no confusion about the location of the target, we superimposed cross hairs on the images. These cross hairs could be toggled on and off by the users so that visual distraction could be minimized when desired. The interface is illustrated in figure 2. Since the users were shown both the target and where it would be located if it were added, this is a 2AFC SKE task.

Users were given 10 minutes of training in the darkened room in order to allow for eye adaptation to the lighting conditions. Users were then shown the display configured either in uniform or spatially varying backlight mode (odd-numbered subjects saw the spatially varying display first, even-numbered subjects were initially presented with the uniform display). The subjects were asked to perform the task for 128 image pairs. The display was then toggled into the opposite mode and the 128 image pairs were repeated. In each of the two trials the order of the image pairs was randomized for every subject.

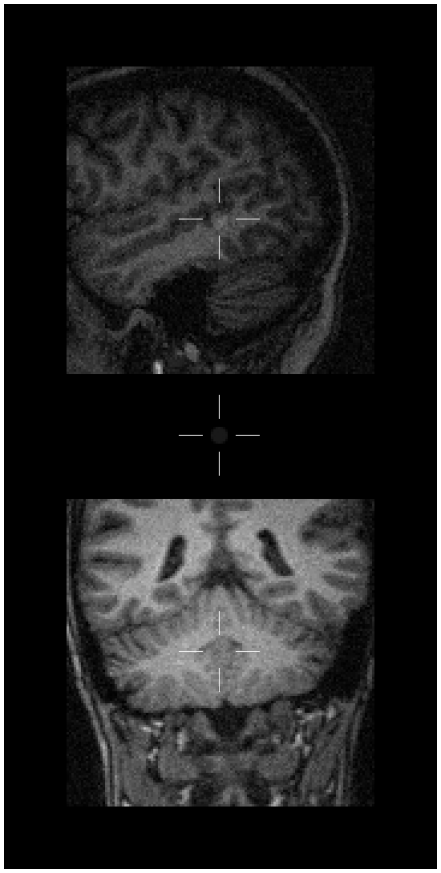

Fig. 2. Example of the 540×1080 pixel user interface. In this image cross hairs have been turned on to reduce localization errors. The target feature is located in the top image in the center of the cross hairs.

C. Dolby Display

The display we used in our experiment was a prototype Dolby display that had a portrait-shaped usable region with a resolution of 540×1080 pixels. The display consisted of an 8-bit color LCD panel backlit by an array of LEDs with 8-bits of luminance control. The LEDs are laid out in a hexagonal grid such that behind each of the images in the experimental interface there were approximately 110 LEDs to provide illumination, with the remainder in the center or around the periphery of the interface.

Using this setup we can simulate a normal, uniformly backlit LCD-based display by simply turning all the LEDs on to the same drive level and using only the LCD panel to modulate the brightness of the display at each pixel. In this case contrast is limited to the available contrast of the LCD panel.

To make full use of the Dolby display, we need to vary the drive levels of the LEDs as well as the LCD panel to produce a spatially varying backlight. We have used the in-house algorithms developed by Dolby to calculate the desired

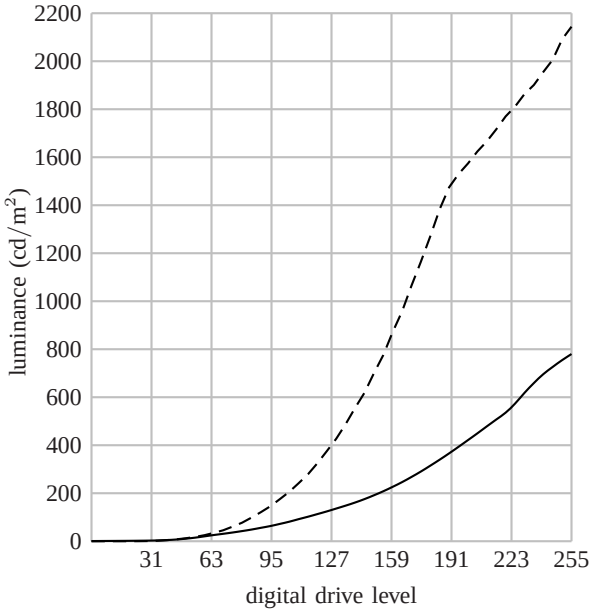

Fig. 4. Luminance of the display in uniform (solid) and spatially varying (dashed) backlight modes for each digital drive level.

LED and LCD drive levels from an 8-bit image. The basic principle though, is that the LEDs display a low-frequency image derived from the 8-bit input and the LCD is used to display a high-frequency correction to the LEDs [11]. This idea is illustrated in figure 3.

To measure the luminance of the display in each of the conditions we used a region the size of one of the images used in our 2AFC task. We recorded the luminance at the center of the image region as the digital drive level was increased in steps of 5. The choice of using a region of this size instead of varying the drive level of the whole screen is based on the nature of the backlight employed. The spatially varying backlight system is affected both by a limit on the power the system can safely draw and the fact that the brightest possible luminance value the screen can produce results from summing the overlapping light emissions of neighbouring LEDs. Given the nature of the system, we felt the most realistic description of display luminance for our task was to vary an "image" of the same size as our data from minimum digital drive level to maximum. The results of these measurements are plotted in figure 4.

As we can see in figure 4, the maximum luminance of the display uniform backlight setup was approximately the same as in a medical-grade LCD display. Luminance varied from 780 cd/m^2 to 0.706 cd/m^2, giving a contrast ratio of approximately 1100:1. Although the display had luminance properties similar to a medical grade LCD display, our prototype display was lacking substantially in resolution. However, given that we were displaying only 128×128-pixel images, the resolution constraint did not affect our simulation of uniformly backlit medical LCD display.

In comparison to the uniform condition, the spatially varying backlight mode allows effectively infinite contrast, with luminance varying from 0 to a maximum of 2140 cd/m^2. In this respect, the performance of the Dolby display with spatially varying backlight resembles the luminance and contrast properties associated with film displayed on light boxes.

Despite the similarities between the Dolby display and film (refer to figure 4) there is a significant deviation from the performance of film when the Dolby display attempts to show high-contrast edges. Edges going from full white to full black in the digital data cannot be physically produced by the display. To understand this, note that essentially the same amount of backlight is shone on two neighbouring pixels as the PSFs of the LEDs are far wider than two pixels. Thus, in order to go from full white to full black in the display would require the LCD to block all the light from the backlight. Of course, if this were feasible there would be no need for spatially varying backlights in the first place, and so we might suspect that the Dolby display is not useful for medical images that contain many edges.

However, as we noted previously, a great deal of scattering occurs in the eye when observing bright objects [3]. This scattering causes neighbouring regions to appear brighter than they actually are, regardless of the light emitted by the display in the dark regions. This effect is known under many names; *blooming*, *veiling luminance*, and *disability glare* are the more common. Based on calculations of this effect, the Dolby display is setup such that the scattering in the eye will produce a "halo" around bright regions that is larger than the mismatch caused by the PSF of the LED backlights [4]. Thus, the display's imperfect ability to represent high-contrast edges can be disregarded as the errors in the display are usually subsumed by the errors in the observer's eye.

In fact, we expect that the same scatter effect would occur for observers of film on light boxes as well. The range of luminance available on mammography light boxes provides at approximately four orders of contrast [12], which suggests that sharp edges on these displays should be equally obscured by scatter in the eye. However, while these effects have apparently not been a substantial detriment to film reading, we felt it was possible that the scattering induced by the prototype Dolby display would be different and detrimental to signal detection, necessitating our present evaluation.

Another deviation in our experiments from normal medical displays was our decision to use the native relationship between digital drive and luminance, instead of the more standard DICOM calibration. One reason for this choice is that it is unclear how a display using the Dolby system can be made to comply with the DICOM standard, given that the available range of luminance available at any pixel is dependent on the luminance of the neighbouring pixels. In fact, this problem holds for any attempt to calibrate the display. If we used a transfer function that linearized the measured values when we used the full-sized square image, we would almost certainly end up with highly non-linear response in regions of medical images.

As we were interested mostly in the veiling luminance due to scatter and the approximations being made by the Dolby variable backlight system, we felt that the transfer function was not likely to be a significant contributor to error between displays as long as it was of a reasonable shape. Ideally we

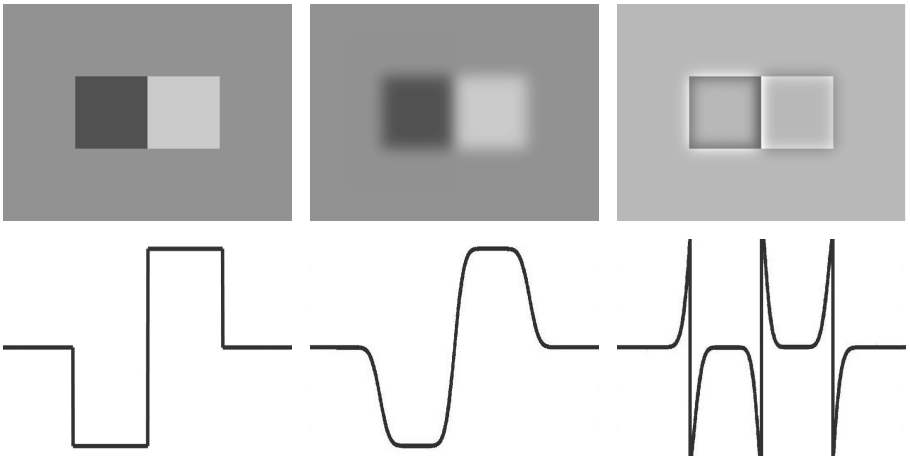

Fig. 3. An example of the LED and LCD drive levels calculated for a single square wave. The left column displays the true image (*top*) and a cross-sectional plot from the true image (*bottom*). The middle column provides the same visualizations of the LED image and the right column provides the visualization of the LCD image. In our experimental pipeline the true square wave would be presented as an 8-bit image. The calculated LED image cannot accurately represent the edges of the wave due to the PSF of the LEDs, so the difference must be corrected by the LCD.

would pick some calibration standard (e.g., luminance-linear, DICOM) and ensure both backlight modes were consistently calibrated. However, since it was unclear how to usefully calibrate the variable backlight mode, we could not setup such equivalent environments. Instead, we left the display with its default transfer function and, having measured the display and viewed many images on it, were satisfied that it provided a usable display for our experiments.

III. Results and Discussion

For any individual subject, we can compute an estimate of the area under the ROC curve (AUC) for the combined diagnosis system of subject and display in a given target/noise power combination by simply calculating the percentage of correct guesses they made in this configuration [13]. Noting this connection between percentage correct and AUC, for the remainder of this paper we will discuss hypothesis tests on percentage correct values, but note that we could equally claim to be performing our tests on estimates of AUC.

Using this method, we have plotted the first, second, and third quartiles of the subjects' percentage correct guesses in figure 5. While figure 5 is useful for illustrating the variability across readers in our experiment, we were most interested in determining whether the two display systems were equivalent for our task. In the remainder of this section we present our analysis of these results using the methods presented by Gallas *et al.* [14] to compute the necessary values for paired t-tests of our hypotheses.

A. Means and Variances of Percentage Correct

As a first step towards this goal we computed the mean percentage correct for each of the two displays and the mean difference in percentage correct between the two displays in each of the four target/noise power conditions. Since our study uses a fully-crossed design where all readers saw every case, we can safely compute our average percentage correct across all readers and cases in a given display and condition as [14]

$$\widehat{P_{d,c}} = \left\langle s_{r,d,c(i)} | d, c \right\rangle \ , \tag{4}$$

TABLE I
Mean Percentage Correct Values

	Spatially Varying	Uniform	Difference
$b = \frac{1}{20}$, $\sigma = \frac{3}{40}$	0.557	0.577	-0.020
$b = \frac{1}{12}$, $\sigma = \frac{3}{40}$	0.621	0.623	-0.002
$b = \frac{1}{12}$, $\sigma = \frac{9}{200}$	0.697	0.744	-0.048
$b = \frac{5}{36}$, $\sigma = \frac{9}{200}$	0.930	0.926	0.004

where $s_{r,d,c(i)}$ is a binary-valued function with 1 for a correct guess and 0 for an incorrect guess when the r^{th} subject looked at the i^{th} image selected from the set of images with target/noise power condition c using display d. We use $c(i)$ to accentuate the fact that the i^{th} image in one target/noise power condition is not the same as the i^{th} image in the other three conditions. Inside of a given condition images are i.i.d. while between conditions images are merely assumed to be independent. We use the notation $\left\langle s_{r,d,c(i)} | d, c \right\rangle$ to indicate that we are taking the mean over i and r with d and c held fixed.

The mean difference in percentage correct between the two displays is then very similar

$$\widehat{P_c} = \left\langle s_{r,d_v,c(i)} - s_{r,d_u,c(i)} | c \right\rangle \ , \tag{5}$$

where d_v indicates the spatially varying backlight and d_u indicates the uniform backlight, which are treated as constants and thus not varied in taking the means. The values we computed for these variables from our experimental data are shown in table I.

We also need to compute the variances of these values. The variance of the percentage correct with one display and condition held constant is given by [14]

$$V_{d,c} = k_1 \left\langle s^2_{r,d,c(i)} | d, c \right\rangle + k_4 \left\langle \left\langle s_{r,d,c(i)} | r, d, c \right\rangle^2 \right\rangle + k_5 \left\langle \left\langle s_{r,d,c(i)} | d, c, i \right\rangle^2 \right\rangle + k_8 \left\langle s_{r,d,c(i)} | d, c \right\rangle^2 \ , \tag{6}$$

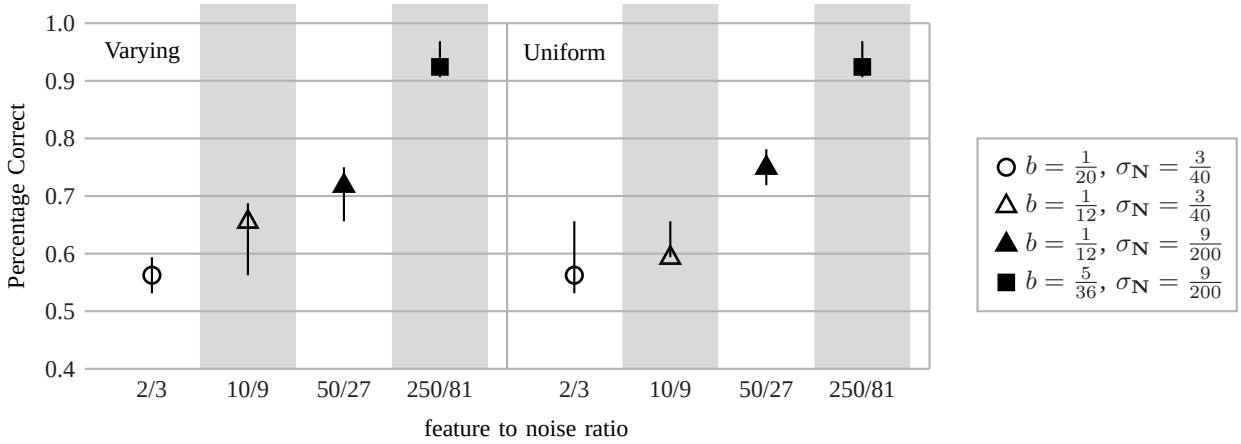

Fig. 5. Plot of the first, second, and third quartiles of subject percentage correct in each of the 8 possible conditions (4 combinations of target and noise power, and two choices of backlight system). The plot is divided vertically into two halves. The left half shows results for the Dolby spatially varying backlight. The right half shows results for the uniform backlight. Each shape represents the median of the percentage correct in one target/noise power configuration. The vertical bars extending from the shape represent the first and third quartiles of the percentage correct, over all the subjects.

where we have slightly modified the notation of Gallas *et al.* and define

$$k_1 = 1/(N_r N_i) \quad (7)$$
$$k_4 = (N_i - 1)/(N_r N_i) \quad (8)$$
$$k_5 = (N_r - 1)/(N_r N_i) \quad (9)$$
$$k_8 = \left[(N_r - 1)(N_i - 1) - N_r N_i\right]/(N_r N_i)\,, \quad (10)$$

where N_r is the number of subjects and N_i is the number of images displayed in one of the four target/noise power conditions. In our data $N_r = 17$ and $N_i = 32$. Using these constants, we can also write the covariance of the two displays' percentage correct values in a specific target/noise power condition as

$$\begin{aligned} V_c = k_1 \left\langle (s_{r,d_v,c(i)} - s_{r,d_u,c(i)})^2 | d, c \right\rangle + \\ k_4 \left\langle \left\langle (s_{r,d_v,c(i)} - s_{r,d_u,c(i)}) | r, d, c \right\rangle^2 \right\rangle + \\ k_5 \left\langle \left\langle (s_{r,d_v,c(i)} - s_{r,d_u,c(i)}) | d, c, i \right\rangle^2 \right\rangle + \\ k_8 \left\langle (s_{r,d_v,c(i)} - s_{r,d_u,c(i)}) | d, c \right\rangle^2 . \end{aligned} \quad (11)$$

In computing these variance estimates it is important we use unbiased estimators of the various means. Following the example of Gallas *et al.* we compute the estimates as follows [14]

$$\left\langle s^2_{r,d,c(i)} | d, c \right\rangle = \frac{1}{N_r N_i} \sum_{r=1}^{N_r} \sum_{i=1}^{N_i} s^2_{r,d,c(i)} \quad (12)$$

$$\left\langle \left\langle s_{r,d,c(i)} | r, d, c \right\rangle^2 \right\rangle = \frac{1}{N_r N_i (N_i - 1)} \sum_{r=1}^{N_r} \sum_{i=1}^{N_i} \sum_{i' \neq i}^{N_i} s_{r,d,c(i)} s_{r,d,c(i')} \quad (13)$$

$$\left\langle \left\langle s_{r,d,c(i)} | d, c, i \right\rangle^2 \right\rangle = \frac{1}{N_i N_r (N_r - 1)} \sum_{r=1}^{N_r} \sum_{r' \neq r}^{N_r} \sum_{i=1}^{N_i} s_{r,d,c(i)} s_{r',d,c(i)} \quad (14)$$

TABLE II
VARIANCE OF PERCENTAGE CORRECT VALUES

	Spatially Varying	Uniform	Covariance
$b = \frac{1}{20}$, $\sigma = \frac{3}{40}$	2.14×10^{-3}	2.32×10^{-3}	0.056×10^{-3}
$b = \frac{1}{12}$, $\sigma = \frac{3}{40}$	2.58×10^{-3}	1.89×10^{-3}	0.25×10^{-3}
$b = \frac{1}{12}$, $\sigma = \frac{9}{200}$	2.42×10^{-3}	2.22×10^{-3}	0.37×10^{-3}
$b = \frac{5}{36}$, $\sigma = \frac{9}{200}$	0.39×10^{-3}	0.50×10^{-3}	0.18×10^{-3}

$$\left\langle s_{r,d,c(i)} | d, c \right\rangle^2 = \frac{1}{N_r (N_r - 1) N_i (N_i - 1)} \sum_{r=1}^{N_r} \sum_{i=1}^{N_i} \sum_{r' \neq r}^{N_r} \sum_{i' \neq i}^{N_i} s_{r,d,c(i)} s_{r',d,c(i')} \quad (15)$$

The results of performing the variance and covariance computations on our experimental data are shown in table II.

B. Hypothesis Tests

The structure of our experiments naturally admits the use of t-tests under the assumption that the mean differences between displays are normally distributed in each condition. Since we are interested in determining the relative performance of the two displays we will test both for difference and equivalence. The difference test will use the standard paired t-test for difference of means while the equivalence of means test will be made using the two-one-sided-test (TOST).

Both of these tests rely on the computation of confidence intervals for the difference of the means. Based on our data, the 95% confidence interval's upper and lower bounds are computed as

$$CI_{\pm} = \widehat{P_c} \pm \frac{\sqrt{V_{d_v,c} + V_{d_u,c} - 2V_c}}{\sqrt{N_r N_i}} T_{0.975}(N_r N_i - 1)\,, \quad (16)$$

where $T_{0.975}(N_r N_i - 1)$ is the 0.975^{th} quantile of Student's t distribution with $N_r N_i - 1$ degrees of freedom. The resulting confidence intervals are given in table III along with the relevant values of the t statistic for each test.

TABLE III
t-VALUES AND 95% CONFIDENCE INTERVALS OF MEAN DIFFERENCE PERCENTAGE CORRECT VALUES

	$t(N_rN_i-1)$	CI_-	CI_+
$b=\frac{1}{20}$, $\sigma=\frac{3}{40}$	-8.16	-2.51×10^{-2}	-1.54×10^{-2}
$b=\frac{1}{12}$, $\sigma=\frac{3}{40}$	-0.68	-0.71×10^{-2}	0.35×10^{-2}
$b=\frac{1}{12}$, $\sigma=\frac{9}{200}$	-17.85	-5.31×10^{-2}	-4.25×10^{-2}
$b=\frac{5}{36}$, $\sigma=\frac{9}{200}$	3.76	0.18×10^{-2}	0.56×10^{-2}

TABLE IV
EQUIVALENCE BOUNDS ON MEAN DIFFERENCE PERCENTAGE CORRECT VALUES

	10% Bounds	Narrowest at 95%
$b=\frac{1}{20}$, $\sigma=\frac{3}{40}$	$\pm5.77\times10^{-2}$	$\pm2.51\times10^{-2}$
$b=\frac{1}{12}$, $\sigma=\frac{3}{40}$	$\pm6.23\times10^{-2}$	$\pm0.71\times10^{-2}$
$b=\frac{1}{12}$, $\sigma=\frac{9}{200}$	$\pm7.44\times10^{-2}$	$\pm5.31\times10^{-2}$
$b=\frac{5}{36}$, $\sigma=\frac{9}{200}$	$\pm9.26\times10^{-2}$	$\pm0.56\times10^{-2}$

Using the derived confidence intervals, the test for difference simply asks whether the value 0 falls within the 95% confidence interval (we could alternatively compute p-values from the reported t-values against the t distribution with N_rN_i-1 degrees of freedom). If not, then we can reject the hypothesis that the two displays are the same since the p-value of the hypothesis is less than 0.05. Referring to table III we see that in three of our four conditions we can reject the hypothesis that the displays are the same, but we do not consistently prefer one to the other across all conditions. Instead, there appears to be a slight preference for the spatially varying display when presenting images with high target SNR and a preference for the uniform backlight display in images with low target SNR.

The TOST suggests that we define some bound on the difference between displays inside of which we will declare them equal for practical purposes. Standard approaches to choosing this include appeals to domain specific knowledge of the field being tested or accepting an error of less than 10% (some authors prefer 20%) of the mean of the reference condition (in our case the uniform backlight) [15], [16]. Since we have no domain specific knowledge for this test that allows us to define meaningful bounds we will compute the 10% bound for the four conditions from the uniform backlight mean in table I. The resulting bounds are shown in table IV.

According to the 95% TOST with 10%-of-reference bounds, we conclude that the means are equivalent if the confidence 95% confidence interval falls within the stated bounds [17]. There are additional methods for constructing symmetric confidence intervals that may allow us to better fit inside the agreed-upon bounds of equivalence [17]. However, consulting tables III and IV, we can see that with our data the confidence interval is already inside the equivalence bounds in all conditions. Thus we can say that, with 10%-of-reference bounds of equivalence, the displays are equivalent.

In fact, we can conclude that they are equivalent with substantially tighter symmetric bounds in some conditions. Based on the TOST methodology it is clear that the smallest symmetric bounds of equivalence that are acceptable are the range symmetric around zero that contain the complete confidence interval. Any definition of practical equivalence that uses a wider symmetric bound than this will similarly be accepted given our data while any that is more restrictive will fail. These narrowest symmetric bounds are listed in table IV.

C. Discussion

Based on the means and variances we computed, our t-tests have demonstrated detectable differences in 3 of our 4 target/noise power conditions, although there was not one consistent display preference across all conditions. However, we have also been able to show equivalence given what is normally considered a reasonable bound on the difference in performance. In fact, as shown in table IV we could use substantially stronger definitions of equivalence in most conditions and still reach the same conclusion. That we have shown both detectable difference and effective equivalence is not a contradiction in our testing methodology. Instead this result indicates that we had a sufficiently large study to detect small differences, but that the differences we found were small enough to be negligible.

Considering table III, there is a clear trend from small differences in favour of the uniform backlight when in very low target SNR images (top of table) to small differences in favour of the spatially varying backlight when in very high target SNR images (bottom of table). It is important consider if this effect derives from either veiling luminance from scatter in the eye or from the low resolution of the spatially varying backlight. We suggest this cannot be the case, as these errors should become more significant effects as the noise is reduced, but we instead see the opposite trend in our data.

Instead, we suspect the trend in performance results from the trade-offs in the algorithm used to produce the LED and LCD images (see figure 3). The algorithm we used was an experimental one that we had tuned specifically for MRI data. However, we have noticed, based on qualitative evaluation, that many of the experimental algorithms we considered have a tendency to accentuate the thermal noise in the MRI images, and we suspect our MRI-specific algorithm still has this property, although to a lesser extent. It is important to note that this accentuation is a result of the algorithm used to produce the LED and LCD images and not the optics of the Dolby display. As such, we expect that with further work on developing medical-imaging-specific LED/LCD image algorithms this effect could be further reduced or completely nullified.

IV. CONCLUSION

We have presented a 2AFC SKE experiment for the exploration of the Dolby spatially varying backlight technology in medical LCD displays. The use of real MRI data as backgrounds ensured that our experiment used images with realistic contrast and structure. Our targets and noise power

were chosen to cover the full range of difficulties from forcing users to essentially guess to making the task almost obvious. By comparing using the same display with a spatially uniform backlight as a simulation of a medical-grade LCD we have minimized the number of possible confounds in our experiment design.

The results of our experiment suggest that the detection of small low-contrast features in complicated, high-contrast backgrounds is possible on displays using the spatially varying backlight. This result confirms our suspicion, based on the years of clinical experience with film light boxes, that the veiling luminance caused by scatter in the observers' eyes would not be a substantial impediment to signal detection. Additionally, our results indicate that the use of the approximations introduced by the low-resolution LED backlight display are not detrimental to signal detection in this context. More generally, we suggest that the displays with the Dolby spatially varying backlight system are useful platforms for further study of high-contrast displays in medical imaging.

We see two significant areas of future work related to using the Dolby display with spatially varying backlighting in a medical context. First, our results suggest it is necessary to reduce the noise-enhancing properties of the current LED/LCD image-generation algorithms. Second, the development of a method for DICOM calibration will be an essential prerequisite to performing further validation studies using real medical data.

We also suggest that further experiments are needed to verify that veiling luminance is not a substantial impediment to detection, Despite the long-standing use of film light boxes with brightness and contrast sufficient to induce veiling luminance via scatter in the eye, we are still concerned that this effect may play a role in hiding small, low-contrast lesions. To thoroughly test this hypothesis detection experiments like this one could be run with the low-contrast target being located in regions calculated to be obscured by veiling glare based on a model of the display and the veiling luminance effect [3].

ACKNOWLEDGMENT

Removed for peer review.

REFERENCES

[1] K. A. Fetterly, H. R. Blume, M. J. Flynn, and E. Samei, "Introduction to grayscale calibration and related aspects of medical imaging grade liquid crystal displays," *Journal of Digital Imaging*, Mar. 2007.

[2] T. Kimpe and T. Tuytschaever, "Increasing the number of gray shades in medical display systems – how much is enough," *Journal of Digital Imaging*, vol. 20, no. 4, pp. 422–432, Dec. 2007.

[3] J. J. Vos, B. L. Cole, H.-W. Bodmann, E. Colombo, T. Takeuchi, and T. J. T. P. van den Berg, "CIE equations for diability glare," CIE, Tech. Rep. 146, 2002.

[4] H. Seetzen and L. A. Whitehead, "A high dynamic range display using low and high resolution modulators," in *SID 03 Digest*, 2003, pp. 1450–1453.

[5] M. D. Tisdall and M. S. Atkins, "Using human and model performance to compare MRI reconstructions," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1510–1517, Nov. 2006.

[6] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magn. Reson. Med.*, vol. 34, pp. 910–914, 1995.

[7] J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," *J. Opt. Soc. Am. A*, vol. 9, no. 5, pp. 649–658, May 1992.

[8] M. P. Eckstein, C. K. Abbey, and J. S. Whiting, "Human vs model observers in anatomic backgrounds," in *Medical Imaging 1998*, ser. Proceedings of the SPIE, vol. 3340, 1998, pp. 16–26.

[9] M. P. Eckstein, C. K. Abbey, F. O. Bochud, J. L. Bartoff, and J. S. Whiting, "The effect of image compression in model and human performance," in *Proc. SPIE*, vol. 3663, 1999, pp. 243–252.

[10] A. Badano and B. D. Gallas, "Detectability decreases with off-normal viewing in medical liquid crystal displays," *Academic Radiology*, vol. 13, pp. 210–218, 2006.

[11] M. Trentacoste, "Photometric image processing for high dynamic range displays," Master's thesis, University of British Columbia, 2006.

[12] E. Siegel, E. Krupinski, E. Samei, M. Flynn, K. Andriole, B. Erickson, J. Thomas, A. Badano, J. A. Seibert, and E. D. Pisano, "Digital mamography image quality: Image display," *Journal of the American College of Radiology*, vol. 3, pp. 615–627, 2006.

[13] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating function," *J. Opt. Soc. Am. A*, vol. 15, no. 6, pp. 1520–1535, Jun. 1998.

[14] B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B70–B80, Dec. 2007.

[15] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, vol. 15, no. 6, pp. 657–680, 1987.

[16] S. Wellek, *Testing Statisical Hypotheses of Equivalence*. Chapman and Hall, 2003.

[17] W. J. Westlake, "Symmetrical confidence intervals for bioequivalence trials," *Biometrics*, vol. 32, no. 4, pp. 741–744, 1976.