COLORIMETRICALLY QUANTIFIED VISUAL TOLERANCES FOR PICTORIAL IMAGES

Mike Stokes, Mark D. Fairchild, and Roy S. Berns*

Abstract: A large body of research exists on tolerances for perceptible and acceptable color matches. This work is exemplified by data such as the well-known MacAdam ellipses. A common theme throughout this type of research is the use of simple fields (*e.g.* uniform patches on uniform neutral backgrounds) and expression of results via device-independent color designations such as CIE color spaces. A separate body of research exists on color tolerances in image reproduction. Common themes in this work include the use of complex images as stimuli and the expression of results using device-dependent color designations such as density or dot percentage units. One aim of the current research was to combine the best aspects of the two bodies of knowledge to determine device-independent color tolerances for complex stimuli. This paper describes psychophysical experiments for the measurement of colorimetrically specified perceptibility and acceptability tolerances in pictorial images. The images could be manipulated in CIELab by a multiplicative factor of 0.93 and power of 1.10 and 0.92 in lightness, a multiplicative factor of 0.92 and powers of 1.12 and 0.88 in chroma, and hue angle offsets of +5.2° and -4.6° before 50% of the observers detected a change. Scene content was found to be unimportant for perceptibility tolerances, while acceptability tolerances were dependent upon scene content. Color difference formulas were also investigated. These results provide useful tools for evaluating color reproduction techniques and testing color appearance models for use in color reproduction.

INTRODUCTION

Digital color-image reproduction continues as an active area of research despite the recent prevalence of imaging software that gives the impression of image portability and color consistency between color peripherals. Recurring research themes include device independence and characterization, chromatic adaptation and appearance models, color gamut mapping, color preferences, and image compression. Only when these issues are adequately resolved will imaging systems produce and transfer visually acceptable color images between devices and systems. The principal research goal is to determine a methodology for each issue that produces the best visual results. Accordingly, there is a need for a metric based on physical measurements that correlates effectively with visual observations of color images.

* Munsell Color Science Laboratory, Rochester Institute of Technology

In classical colorimetry, the metric is a color difference calculation. Today, these include color difference formulas based on the CIELab and CIELUV color spaces (CIE, 1986), and extensions of CIELab such as the CMC (McDonald, 1988) and MCSL (Berns, *et al.*, 1991) equations. These metrics have been successfully employed in the paint, textile, and polymer industries, to name a few. The major advantage of CIE-based metrics is their device independence based on the human visual system. A disadvantage is that these equations were optimized for differences between uniform fields of color. It is unknown whether these formulas will adequately model color differences between pictorial images.

For pictorial images, an entirely different set of metrics has been created. The most popular include using device digital counts, status densitometry, dot areas, and equivalent neutral densities (Evans, 1953). Significant testing has been done in order to successfully relate each of these methods to perceptible color tolerances for pictorial images. However, none of these methods are device independent; each tolerance depends on the particular colorant set and the measurement devices have system spectral responsivities that are not easily related to colorimetry.

There is a need to develop a visual data base of tolerance judgments of pictorial images that can be used to test colorimetry-based metrics. This paper describes two experiments that were performed to generate this data base. The first experiment was designed to derive perceptibility and acceptability tolerances, determine the impact of scene content on tolerances, and determine if the CIELab, CMC, and MCSL color difference formulas adequately model pictorial image tolerances (Stokes, 1991). The second experiment was designed to measure the repeatability and robustness of the first experimental results and to determine if color charts adequately model pictorial images. Ideally, if color difference formulas and color charts adequately model pictorial image tolerances and scene content does not impact the results, then one can simply use color charts, measure color differences between images generated using various imaging modalities, and apply the results to pictorial images.

EXPERIMENTAL TECHNIQUES

Image Display

A Sony GDM-1950 color monitor controlled by a Pixar II image computer was used as the stimulus generator in both experiments. It was colorimetrically characterized using an LMT C1200 colorimeter and the calibration technique of Berns, Gorzynski and Motta (1991a, 1991b, and 1988). First, the neutral point was set to D65 at various luminance levels using the monitor's internal adjustments. This ensured that the monitor would track neutrals correctly. The maximum luminance was 85.0 cd/m^2. To convert between linear red, green, and blue tristimulus values and XYZ tristimulus values, a 3x3 matrix was derived from the maximum red, green, and blue phosphor outputs. These readings were normalized to the Y value of the white point digital counts to produce relative tristimulus values. Next, the XYZ tristimulus values for five luminance levels of equal digital counts (neutrals) were transformed into red, green and blue tristimulus values using the above matrix. A non-linear regression model was used for each channel to obtain the appropriate parameters to transform the nonlinear digital counts into linear red, green, and blue tristimulus values. The accuracy

of the calibration technique was verified by comparing the measured and predicted XYZ tristimulus values throughout the color gamut with a sample of 125 colors. The average CIELab E*ab was 0.40 with a standard deviation of 0.17.

Psychophysics

Forced-choice paired-comparison experiments were performed where three images were displayed sequentially within an adapting background image. Toggling between three keys allowed the observer to switch between a reference image, a standard image (which was identical to the reference image) and a manipulated image, all of which were stored in the frame buffer. The observer stopped on the image that appeared different from the reference image and then judged if it was an acceptable reproduction or not. After both judgements were made, the next triad of images was loaded into the frame buffer for presentation. The sequential overlaying of three experimental images in the same location was chosen to eliminate the effects of spatial monitor non-uniformity. A neutral image of identical luminance factor to the adapting background was displayed for 0.2 seconds between the three images.

The following instructions were read to each observer.

> In this experiment, you will be comparing three images at a time. A reference image will always be displayed first, the left button will recall this image at any time. The middle and right buttons toggle between an image that is identical to this reference image and one that has be manipulated in color. Switch between the images, STOPPING on the one that appears different from the reference. If they appear the same, you must still make a choice. If you can't decide, just guess. Once you have decided which image is different, you must decide if this is an acceptable difference or not. Press 'A' if the difference is acceptable and 'N' is it is not acceptable. For this experiment, we are defining acceptability to be 'a reproduction print that you would be expect to find in an expensive book of photographic reproductions.' Many pairs of images will appear identical, please do not let this frustrate you. You should make overall judgements and not compare very small image areas. There are a total of 426 image pairs. There is a five second delay between the images, and a bell will sound when the next image is ready, (please do not press any buttons between images). We will begin with six demonstration images to make sure you understand the directions.

One of the complexities encountered was obtaining both perceptibility and acceptability tolerances within a single experiment. Although the instructions were lengthy, most observers had no difficulties. The 426 image pairs were split between three one-hour observational sessions.

A background image was employed to make the experimental images appear more like reflection prints, as opposed to self-luminous images and to provide a constant adapting stimulus. This provides a better correlation to most color reproduction systems in which

output is a hardcopy image (Fairchild, 1991). The background images were created by photographing a pairs of hands holding a 5" by 7" white card on a Munsell N5 background. Two background images were created, one with the white card placed horizontally and one with it placed vertically. With the observer positioned approximately 18" to 24" from the display, the monitor displayed a background image of two hands on a neutral field holding a 5" by 7" white card. Centered within the card, the test images were displayed in a 4" by 6" field.

Image Manipulation

The first step in the color image manipulations was choosing appropriate color dimensions. The criteria for choosing color dimensions included visual uniformity, intuitiveness, industry standardization, and applicability to current color difference formulas. The CIELab dimensions of lightness, chroma, and hue angle were chosen as the most appropriate dimensions to meet these criteria. Munsell and others before him illustrated that the dimensions of lightness, chroma and hue are very intuitive (Munsell, 1979). Common spaces in current use, such as YIQ or HLS, provide no device independence and are significantly nonuniform. The Yxy space, while device independent, is not perceptually uniform. CIELUV, while meeting most of these criteria, was not as applicable to current color difference formulas in common use.

After the color dimensions were chosen, the types of manipulations to be performed within these dimensions were determined. Four transfer functions were used to manipulate the images. These functions are fundamental mathematical constructs and simulate common industry process transformations for color casts or shifts, gain, gamma, and contrast. The mathematical form of these four transfer functions are an additive offset, a multiplicative factor, a power function, and a sigmoidal function. An example of the simoidal function is shown in figure 1.

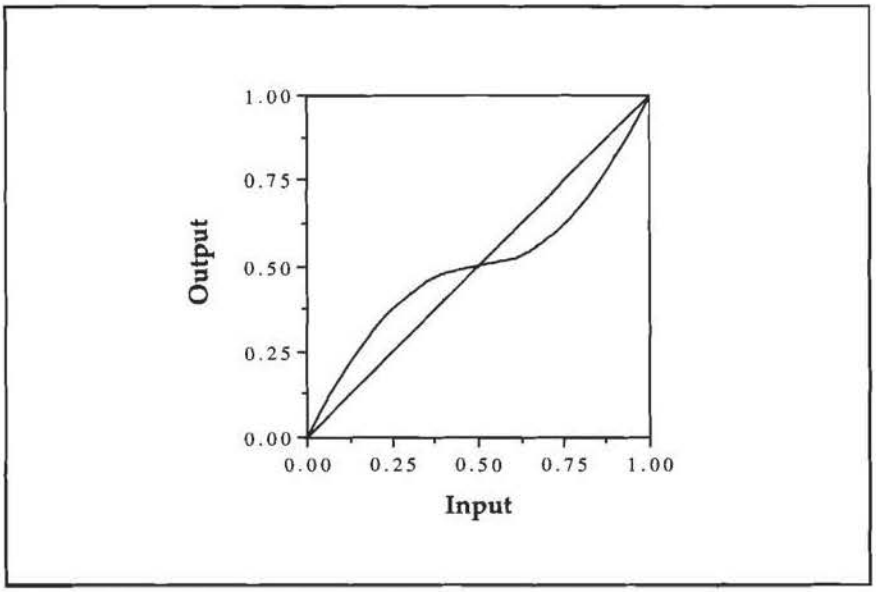

Figure 1 : Sigmoidal transfer function with a parameter level of 0.5.

Given three color dimensions and four transfer functions, a total of twelve combinations exist. Common sense dictated the elimination of several of these combinations. For example it made no sense to have a multiplicative function applied to hue or an additive offset applied to chroma. The final combinations used in this research were a multiplicative factor in lightness, a power function in lightness, a sigmoidal function in lightness, a multiplicative factor in chroma, a power function in chroma, and an additive offset in hue. Since symmetry could not be assumed for the additive offset, power and sigmoidal functions, each of these functions were divided into either high and low or positive and negative parameter levels. These divisions brought the total number of unique transfer function and color space dimension combinations to the 10 shown in table 1.

Some of the transfer functions required anchor points. Anchor points are points in a transfer function that are stable with respect to changes in the function's parameters and must be chosen within the context of the dimension being manipulated. For example, a power transfer function for L* has two canonical anchors at the values of 0 and 100. For chroma, the maximum value can be device, image, and hue-angle dependent. As a consequence, algorithms used to locate maximum chroma would result in transfer functions that vary throughout the experiment. The experimental objectives necessitated an alternate approach to achieve predetermined transfer functions. This was accomplished by evaluating the chroma of blue sky, green grass and flesh tones (Bartleson, 1962) and assigning the maximum chroma anchor as twice the chroma of these important colors. This was a C* of 65.

Transfer Function	Name	CIELAB Dimension	Parameter Values
Multiplicative Factor	LMF	Lightness	<= 1.0
Power	LPH	Lightness	>= 1.0
Power	LPL	Lightness	<= 1.0
Sigmoidal	LSH	Lightness	>= 1.0
Sigmoidal	LSL	Lightness	<= 1.0
Multiplicative Factor	CMF	Chroma	<= 1.0
Power	CPH	Chroma	>= 1.0
Power	CPL	Chroma	<= 1.0
Additive Offset	HOH	Hue Angle	>= 0.0
Additive Offset	HOL	Hue Angle	<= 0.0

Table 1 : Transfer-function/color-space-dimension combinations and abbreviations.

Images were then manipulated using each transfer function with the appropriate parameters set to different levels. The exact values and number of levels for each function were determined from a pilot experiment. Once the transfer-function parameter levels were established, the actual experimental images were created. The digitized images were transformed into CIELab lightness, chroma, and hue-angle dimensions and the appropriate colorimetric manipulations were performed. All image transformations and manipulations were done using floating-point variables to avoid quantization errors. After each image was manipulated, it was transformed into a displayable format for the Pixar image computer.

Statistical Analysis

The observational data were analyzed for goodness of fit, perceptibility and acceptability tolerances and uncertainty estimates for these tolerances. The individual scene results were compared against each other and the average results were compared against common color difference formulas.

Probit analysis (Finney, 1979) was used to analyze the data. Probit analysis is a maximum likelihood model relating experimental responses to occurrence probability estimates. With this model the frequency of observer responses are fitted to a cumulative normal distribution. The Pearson Chi-Squared test and its associated probability determined how well the data fit the cumulative normal distribution assumed in the probit analysis and therefore how homogeneous the data were. Previous visual experiments have shown that goodness of fit Chi-Squared probabilities of five percent or greater yield sound results. Estimation of tolerances and uncertainty of these tolerances were derived from this model. The color tolerance estimate is the median tolerance (T50) at a rejection or acceptance probability of 50%. The fiducial limits (approximately 95% confidence limits denoted LOWER and UPPER) were calculated to produce an estimate of uncertainty for the T50 results. Fiducial limits are expressions of the probability that, within a certain percentage, the estimate will fall in a particular range. Alman (1989), and Berns (1991) have provided detailed explanations of probit analysis and fiducial limits as applied to color tolerance experiments.

The tolerance median and fiducial limits were derived using the SAS probit analysis procedure (SAS, 1990). The SAS Logistic procedure (SAS, 1990) (using the probit model) was used to determine the actual Pearson Chi-Squared values, probabilities and the C statistic. The C statistic is a measurement of parameter sensitivity and thus model fit. These two separate SAS procedures were used since only the logit procedure provided discrimination parameters and only the probit procedure provided fiducial limits. For large populations, logit and probit models are equivalent (Finney, 1971).

The C statistics were used to isolate and eliminate the extreme 1.5 percent of the observations. Previous researchers have filtered data to reduce some of the visual noise and improve the statistical significance (Berns, 1991). Such filtering was justified in the current experiment by an estimated "key-stroke-error rate" of 4 percent. This error rate was derived from a one percent rate of "fail/standard" response combinations. This response would indicate that the observer judged the standard image to be an unacceptable reproduction of the reference image. This is obviously wrong, since these images were identical, so it was assumed that the observer miskeyed the response. Since this is one of four possible combinations, a total error rate of 4 percent was estimated. This error rate justified filtering out up to 4 percent of the data, although only 1.5 percent of the data were actually eliminated.

The perceptibility results were significantly improved by this filtering technique and strong statistical significance with a very low noise level was achieved. The acceptability tolerances were not significantly improved by filtering, indicating that the data were heterogeneous. This was not unexpected. In the post-experimental survey, several observers stated that they ignored the acceptability criteria and used their own criteria for most of the experiment. Such behavior would create multiple acceptance tolerances and thus heterogeneous noise in the acceptability data depending on each observer's acceptability criteria. The individual Chi-Squared statistics support this argument by showing less noise for the individual results than the grouped results. Since the acceptability data was not improved by filtering, the raw results were used in further analysis.

The above experimental details were followed for both experiments except where explicitly noted below.

EXPERIMENT 1

Forty-four color-normal observers with varied color analysis experience participated in the experiment. The observers varied in age from 20 to 49 and were tested for color vision using either a visual colorimeter or standard color deficiency plates. A survey was made after the experiment to determine any problems and suggestions for the overall experiment. The room was darkened during the entire experiment.

In order to determine the impact of scene content on tolerances, several concerns were reviewed. Three dominant concerns in analyzing pictorial images have been scene-content dependence (Jones, 1941), perceived object distance (Corey, 1983) and overall chroma levels (Bartleson, 1958). Six different images were used in order to examine these issues. The six images were divided into three scene-content types and two levels each for perceived object

distance and chroma content. The three scene types were man-made objects, people, and natural scenes.

Each image was judged for both perceptible and acceptable differences. A limit of three one-hour observation sessions per observer was set to avoid observer fatigue. The 3 hour time limit required that the average pair of perceptibility and acceptability judgements be made in under twenty seconds. The standard and sample images were loaded and displayed in about eight seconds; thus leaving an average of twelve seconds for each pair of judgements. The post-experimental survey indicated that observational duration was not a problem for the observers.

Perceptibility

The experimental perceptibility tolerance results analyzed by transfer function and across scene are given in table 2. The high Chi-Squared (χ^2) probabilities indicate that averaging across scenes produces homogeneous data, therefore scene content did not impact the tolerance results. Values greater than or equal to 0.05 indicate the probit model characterizes the visual results with a 95% probability. Notable are the tolerance values themselves, the tight fiducial limits, the apparent symmetry for all of the dual-sided functions (*e.g.*, LPH and LPL) and the hue-angle offset of 5 degrees. These tolerance values, because of their high statistical confidence as quantified by the tight fiducial limits, can be used to calculate whether a reproduced or manipulated image is perceptibly different than an original image. The symmetry implies experimental redundancy; thus future experiments can be significantly shortened with the same amount of information gathered. The hue angle tolerances of 5° were surprising because previous research (McDonald, 1988) indicated that observers were extremely sensitive to hue angle shifts. The 5° shift indicates that observers are less sensitive to shifts in hue angle of images than uniform fields by about a factor of two. This is verified in the CIELab analysis below.

Function	T50	LOWER	UPPER	Prob > χ^2
LMF	0.93	0.92	0.95	0.08
LPH	1.10	1.09	1.11	0.49
LPL	0.92	0.91	0.92	0.13
LSH	1.15	1.14	1.16	0.37
LSL	0.89	0.88	0.90	0.18
CMF	0.92	0.91	0.93	0.19
CPH	1.12	1.12	1.13	0.33
CPL	0.88	0.88	0.89	0.58
HOH	5.23	4.74	5.68	0.13
HOL	-4.62	-5.10	-4.08	0.19

Table 2 : Perceptibility results for transfer functions in experiment 1. (See Table 1 for a description of each function.)

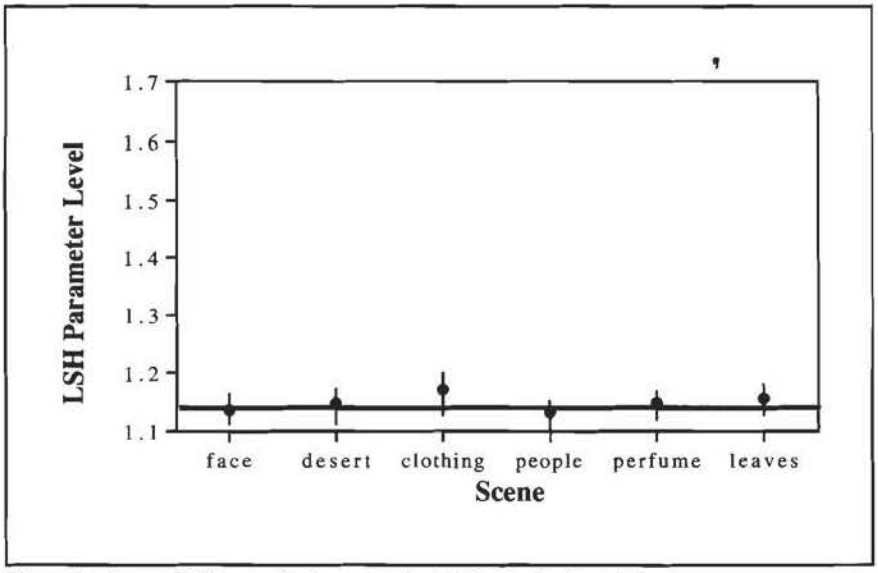

Figure 2 : Perceptibility results by scene for LSH transfer function.

The result that scene content did not affect perceptibility judgements was surprising. In general the opposite result is noted (Jones, 1941 and Corey, 1983). Details of the current results are shown in figures 2-4 for three representative color-dimension/transfer-function combinations. The median tolerance and its fiducial limits for the particular transfer function is plotted against each of the six scenes. If a horizontal line can be drawn intersecting the fiducial limits of all six images, they are not statistically different from each other at a 99% probability level. The plotted results are indicative of the remaining seven transfer functions. The "face" scene was a silhouette of an woman's face within a neutral background. The "desert" image was a typical southwestern scene containing red dirt and blue sky. The "clothing" scene was a collection of different textile materials. The "people" scene consisted of three women in brightly colored dresses. The "perfume" scene contained three bottles of colorful perfume on white ceramic tiles. The "leaves" image was a collection of fall-colored leaves on the ground.

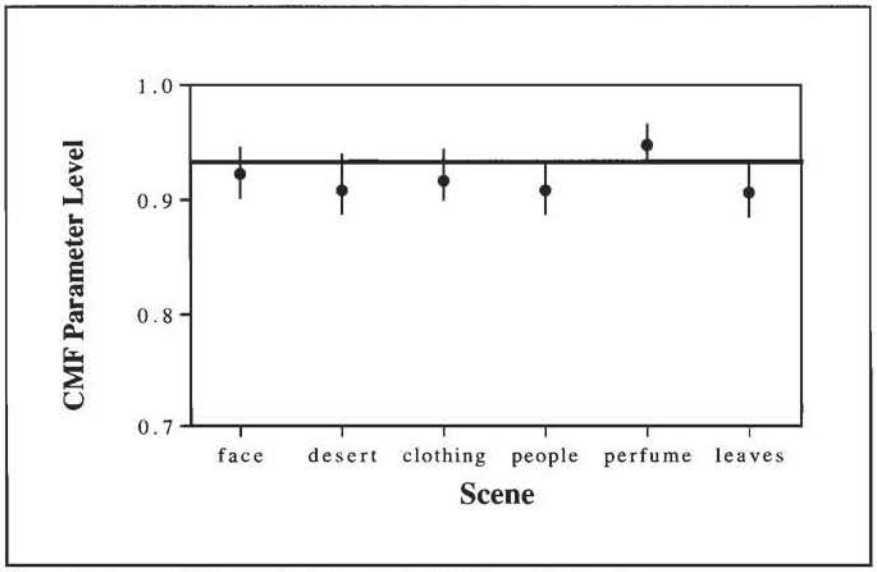

Figure 3 : Perceptibility results by scene for CMF transfer function.

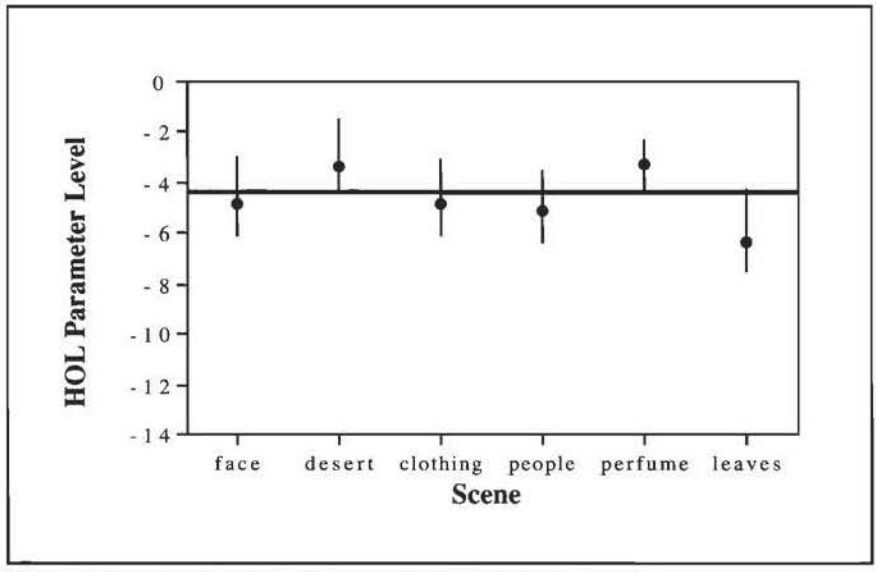

Figure 4 : Perceptibility results by scene for HOL transfer function.

Acceptability

The experimental acceptability tolerances are shown in table 3. Most importantly, the Chi-Squared probabilities indicate a poor model fit for all of the transfer functions. An analysis by scene and by transfer function indicated adequate model fit; as a consequence, the table 3 results indicate that scene content significantly impacts acceptability tolerances, unlike the perceptibility tolerances, in similar fashion to Jones (1941) and Corey (1983). The low Chi-Squared probabilities indicate random noise that can be attributed to the different observers using different acceptability criteria for different scenes. This is supported in the survey comments.

Function	T50	LOWER	UPPER	Prob > χ^2
LMF	0.90	0.89	0.91	0.00
LPH	1.18	1.17	1.20	0.00
LPL	0.88	0.87	0.89	0.01
LSH	1.32	1.28	1.35	0.00
LSL	0.84	0.83	0.86	0.00
CMF	0.86	0.85	0.88	0.00
CPH	1.19	1.18	1.21	0.00
CPL	0.81	0.79	0.83	0.00
HOH	8.43	7.33	9.66	0.00
HOL	-8.20	-9.46	-6.98	0.00

Table 3 : Acceptability results for transfer functions in experiment 1.

Despite the low Chi-Squared probabilities, the median T50 tolerances are very representative of analyses done by scene and by observer. (Analyses with better model fits by definition have smaller fiducial limits thus only the T50 values are representative.) The tolerances are significantly greater than the previous perceptibility tolerances. The lack of symmetry eliminates the possibility of reducing the number of functions, again indicating a significant difference between perceptibility and acceptability judgements. These results also indicate some significant difference between scenes. This was verified by analyzing the individual scene results.

The significant differences between perceptibility and acceptability judgments strongly indicate that one cannot simply scale perceptibility results to define acceptability tolerances. Tolerances based on acceptability judgments are probably scene dependent and certainly observer dependent. These results suggest that it is critical to carefully define the subjective criteria when designing visual experiments to scale image quality. If possible, physical anchors should be used rather than cognitive criteria based on a set of instructions or observer experience.

Color Difference Formulas

The CIELab, CMC, and MCSL color difference equations were evaluated in comparison with the perceptibility tolerances above. Each image was manipulated by the ten transfer functions with parameter levels equal to the T50, UPPER, and LOWER perceptibility values for each scene.

The color difference for each pixel was calculated and from these differences, an average color difference was computed for the entire image. Finally, the color differences for each scene were averaged together. These calculations are summarized in table 4. Median perceptibility tolerances ranged between CIELab color differences of approximately 1.5 and 2.5. This magnitude is at least twice that of perceptibility tolerances for solid color patches. With the exception of the sigmoidal transfer functions of lightness, CIELab was an excellent predictor of pictorial color differences for this experiment. The sigmoid transfer functions caused both positive and negative changes within an image whereas the other transfer functions caused either positive or negative changes but not both. As a consequence, observers were more sensitive to the sigmoidal changes resulting in smaller color differences.

The MCSL equation varies the chroma and hue weighting of color differences compared with CIELab. For this reason, both equations had the same color differences for the lightness functions but very dissimilar values for the chroma and hue transfer functions. The CMC equation varies lightness, chroma, and hue weightings compared with CIELab resulting in different values for all three types of transfer functions.

In order to compare the effectiveness of the three color difference formulas, the magnitudes were normalized by the average color difference for each equation and plotted in figures 5 - 7. Ideally, all of the normalized results would overlap at or near unity. The figures show that this is not the case for the CMC and MCSL color difference formulas and thus these color difference formulas do not adequately predict tolerances for pictorial images. The CMC and MCSL formulas are based on visual judgements of small color differences where as CIELab is based on large color differences exemplified by the Munsell color order system. Since CIELab is a better predictor of the visual results than the other formulas, perceptibility judgments of pictorial images produce sensory responses similar to large color difference judgments.

The mean acceptability tolerances were about 6 CIELab color difference units. This acceptability result agrees well with results derived in a dramatically difference fashion by Stamm (1981).

	CIELAB			CMC			MCSL		
FCN	T50	lower	upper	T50	lower	upper	T50	lower	upper
LMF	2.51	1.86	2.98	2.68	1.99	3.16	2.51	1.86	2.98
LPH	2.44	1.66	2.91	3.04	2.07	3.62	2.44	1.66	2.91
LPL	2.21	1.49	2.72	2.76	1.85	3.41	2.21	1.49	2.72
LSH	1.71	1.36	1.97	2.16	1.71	2.48	1.71	1.36	1.97
LSL	1.43	1.06	1.68	1.79	1.33	2.11	1.43	1.06	1.68
CMF	1.97	1.39	2.39	1.04	0.74	1.27	0.85	0.60	1.03
CPH	2.46	2.05	2.77	1.45	1.21	1.64	1.18	0.99	1.34
CPL	2.65	1.91	3.21	1.61	1.16	1.96	1.31	0.95	1.59
HOH	2.22	1.64	2.61	2.54	1.88	2.99	1.59	1.17	1.87
HOL	1.90	1.23	2.34	2.19	1.43	2.69	1.36	0.88	1.67
AVG	2.15	1.57	2.56	2.13	1.54	2.53	1.66	1.20	1.98
S.D.	0.39	0.32	0.48	0.65	0.43	0.80	0.50	0.33	0.61

Table 4 : Comparison of Raw Color Difference Formulas for experiment 1.

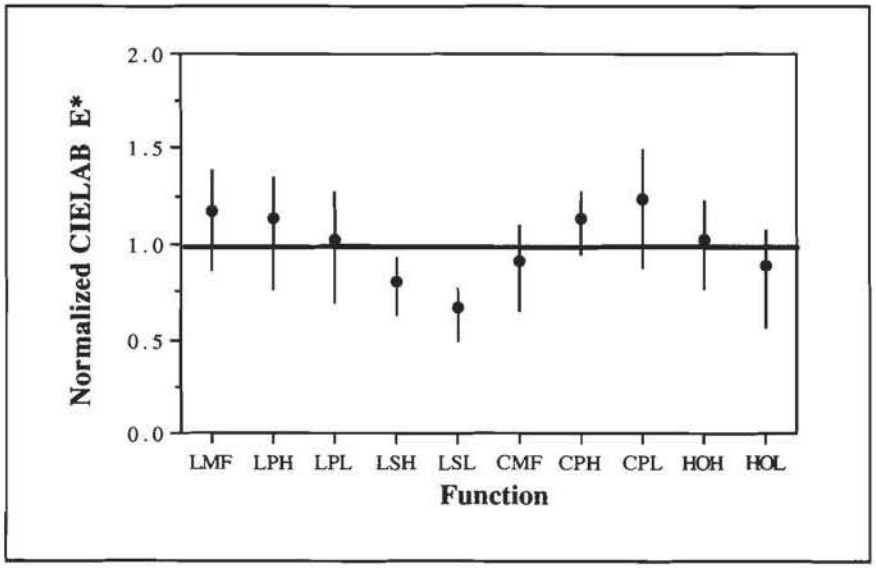

Figure 5: Normalized CIELab color difference results for experiment 1.

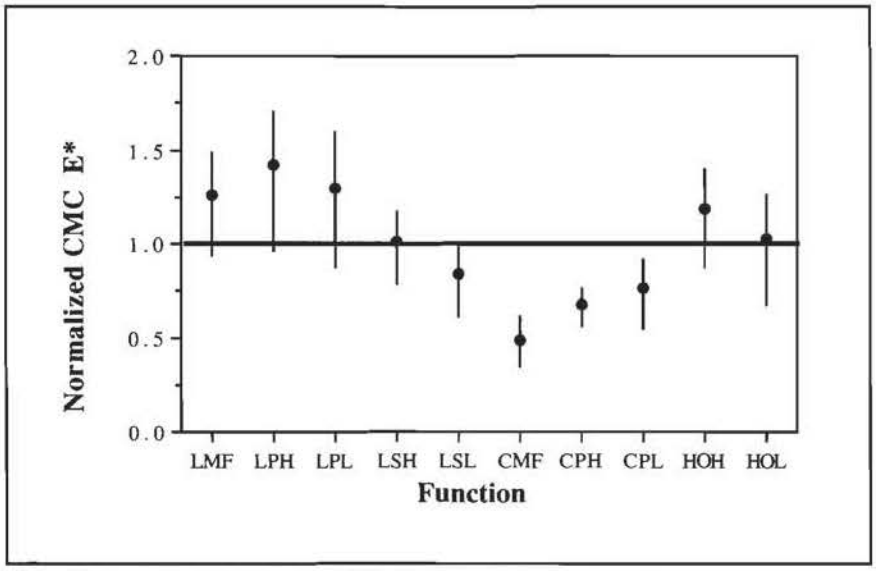

Figure 6: Normalized CMC color difference results for experiment 1.

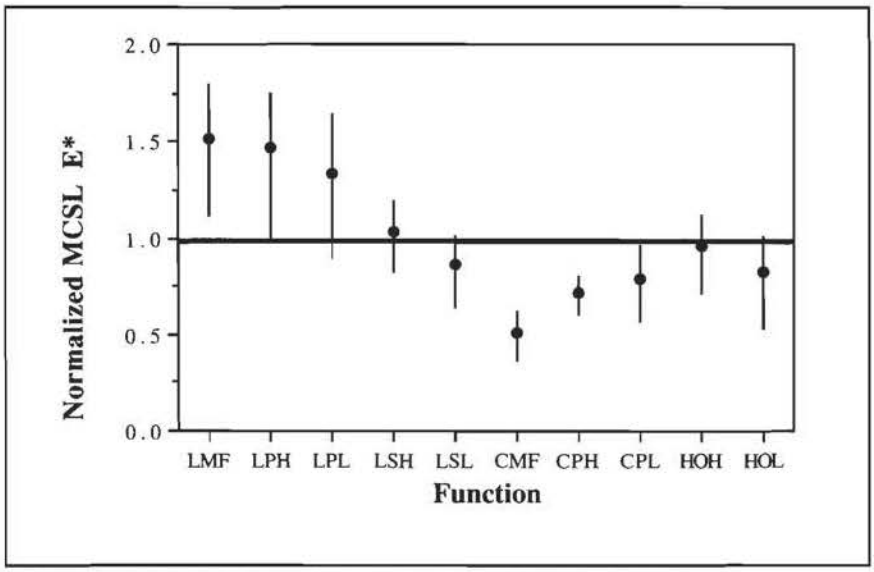

Figure 7: Normalized MCSL color difference results for experiment 1.

Practical Applications

The transfer-function tolerances can be used to evaluate other image manipulation routines such as gamut mapping or quantization. There are two possible methods of applying the experimental tolerances to derive such perceptibility and acceptability data without visually assessing the images.

The first method incorporates non-linear regression and is computationally intensive. A comparison can be performed between an original and a manipulated image by regressing with the various transfer functions operating on the manipulated image. The resulting regression parameter estimates can be compared directly to their respective tolerances. This comparison establishes whether the manipulated image is perceptibly or acceptably different from the original.

A second comparison method uses the CIELab color difference results. By using the minimum and maximum differences, limits can be established for classifying image differences. The CIELab minimum and maximum differences are 1.06 and 3.21 respectively. Differences are calculated by averaging the color differences for each pixel in an image. If the mean difference is below the minimum difference, then the manipulated image is not perceptibly different from the original. If the color difference is above the maximum difference, then the manipulated image is perceptibly different from the original. If neither of the above two cases are true, then this method does not yield conclusive results.

EXPERIMENT 2

A second experiment was performed to verify the first experiment and to determine if tolerances for color charts matched those for pictorial images. A simplified version of the previous experimental set-up and design was used. The simplifications included using only perceptibility judgements and three color-dimension/transfer-function combinations. In addition five of the six images were changed. A total of thirty-two observers performed this experiment with an average observational time of 45 minutes for 124 judgements.

The six images in the second experiment included three pictorial images and three different color charts. The desert image was repeated from experiment 1. The additional two pictorial images consisted of a scene of a mountain with a blue lake and a couple in a canoe on a pond. The three color charts were the Macbeth Color Checker (McCamy, 1976), a mosaic of randomly colored rectangles, and a chart patterned after research done by Luo, *et al. (1991)*.

The desert scene produced the same tolerances, illustrated in table 5, verifying experimental repeatability within a 99 percent confidence interval.

Function	Exp. #	T50	Low	High	Prob > χ^2
LSH	1	1.17	1.13	1.20	0.77
LSH	2	1.24	1.18	1.30	0.18
CMF	1	0.91	0.89	0.94	0.73
CMF	2	0.89	0.84	0.94	0.96
HOL	1	-3.32	-1.47	-4.27	0.71
HOL	2	-5.08	-2.96	-6.97	0.78

Table 5 : Comparison of desert scene perceptibility results for experiments 1 and 2.

The perceptibility tolerance results from this experiment are listed in table 6. These results do not agree with experiment 1 tolerances. The T50 tolerances are actually closer to the acceptability results and the chi-squared probability for the LSH function indicates heterogeneous data. Further investigation indicated that filtering the data would not alter these results. Therefore the scenes were analyzed individually.

Function	T50	Low	High	Prob > χ^2
LSH	1.28	1.21	1.38	0.01
CMF	0.84	0.83	0.86	0.45
HOL	-9.94	-8.65	-11.44	0.28

Table 6 : Perceptibility results for transfer functions in experiment 2.

Individual scene analyses shown in figures 8-10 yielded many fruitful results. The two new pictorial images were compared with the previous results to verify robustness of the previous results. While the lightness tolerances were in excellent agreement with the previous pictorial results, the chroma and hue tolerances were more difficult to analyze. Further image analysis indicated that the canoe and mountain images were very low in chroma content with average C* values of 9.7 and 10.9 compared with an average C* value of 20.8 for the desert scene. This possibly makes the chroma and hue shifts difficult to judge and was verified by interviewing the observers. After careful review of the individual responses, it was found that some of the inexperienced observers detected no change in these images with respect to chroma or hue and thus shifted the T50 values higher than expected. By eliminating these observers, the results were in complete agreement with the previous experiment, verifying the robustness of the previous results. Third, the various color charts were analyzed to see if the observers judged them differently from the pictorial images. Figures 9 and 10 show no significant differences for changes in chroma or hue, verifying the above chi-squared results. The lightness sigmoidal function illustrated in figure 8, shows observers are less sensitive to changes in the mosaic and to a lesser extent in the Macbeth color checker chart. Although not a strong result, this result does bring into doubt the feasibility of simulating pictorial images with color charts.

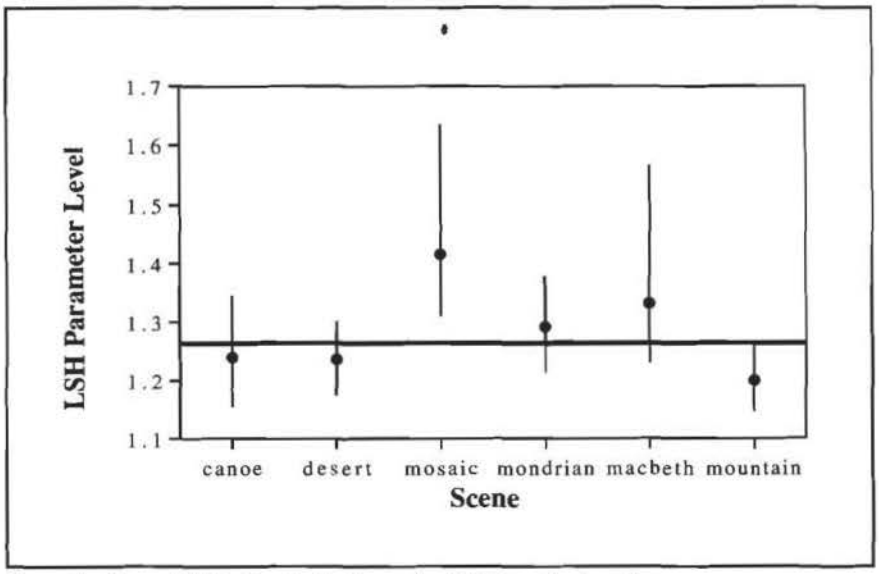

Figure 8 : Perceptibility Results by Scene for LSH transfer function.

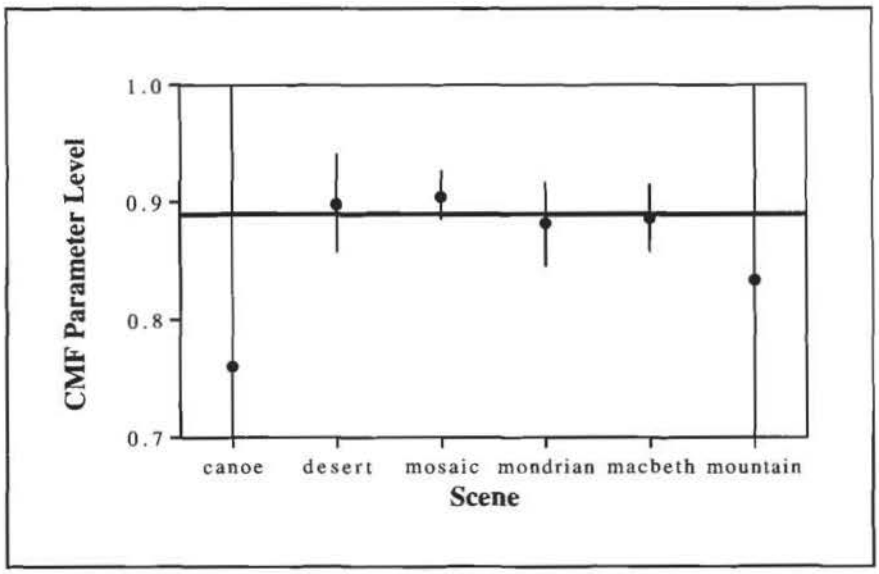

Figure 9 : Perceptibility Results by Scene for CMF transfer function.

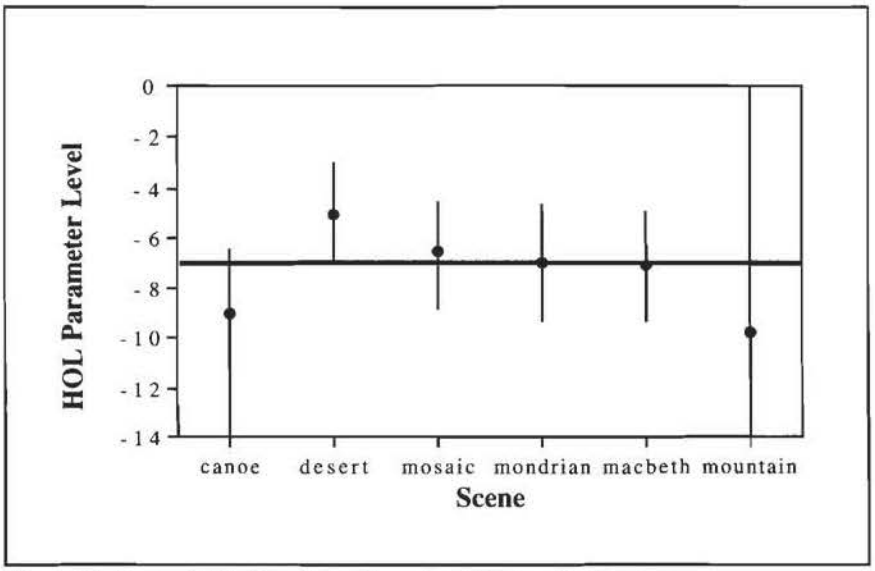

Figure 10 : Perceptibility Results by Scene for HOL transfer function.

Finally, table 7 and figure 11 illustrate that CIELab remains an adequate, although not perfect metric for measuring color differences for pictorial images. The average T50 values for experiment 2 are within the 95% fiducial limits of experiment 1, verifying the experimental repeatability.

	CIELAB			CMC			MCSL		
FCN	T50	lower	upper	T50	lower	upper	T50	lower	upper
LSH	2.38	1.61	2.85	3.02	2.03	3.62	2.38	1.61	2.85
CMF	2.19	1.23	3.41	1.16	0.58	1.54	0.94	0.48	1.26
HOL	2.01	1.11	2.66	2.30	1.37	2.84	1.43	0.85	1.77
AVG	2.49	1.54	3.29	2.54	1.62	3.17	1.98	1.26	2.48
S.D.	0.43	0.34	0.54	0.98	0.69	1.17	0.82	0.57	0.99
Nml'd									
AVG.	1.00	0.73	1.19	1.00	0.72	1.19	1.00	0.72	1.19
S.D.	0.18	0.15	0.22	0.31	0.20	0.37	0.30	0.20	0.37

Table 7 : Comparison of raw color difference formulas for experiment 2.

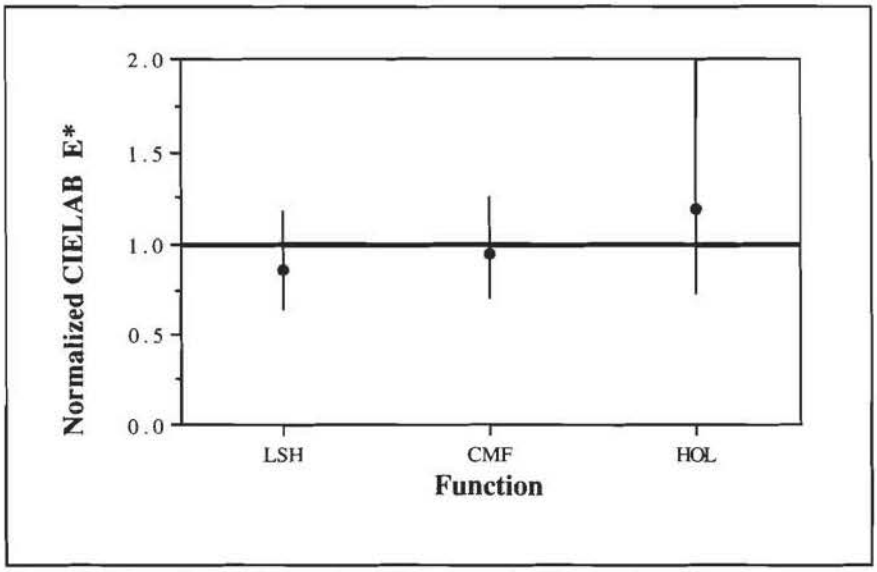

Figure 11: Normalized CIELab color difference results for experiment 2.

CONCLUSIONS

Four main conclusions can be drawn from the results discussed in this paper. Scene content does not affect perceptibility tolerances for pictorial images. The CIELab color difference metric, E^*_{ab}, is adequate for estimating these perceptibility tolerances. Acceptability tolerances are not linearly scaled values of perceptibility tolerances. Lastly, color charts cannot be assumed to be good simulators of pictorial image with respect to perceptibility thresholds.

LITERATURE CITED

Alman, D.H., Berns, R S., Snyder, G.D., and Larsen, W.A.,
1989. "Performance Testing of Color-Difference Metrics Using a Color Tolerance Dataset," **Col. Res. Appl., 14** 139-151.

Bartleson, C.J.,
1958. "Influence of Observer Adaptation on the Acceptance of Color Prints," **Pho. Sci. Eng., 2**, 32-39.

Bartleson, C.J., and Bray, C.P.,
1962. "On the Preferred Reproduction of Flesh, Blue-Sky and Green-Grass Colors," **Pho. Sci. Eng., 6**, 19-25.

Berns, R.S., Alman, D.H., Reniff, L., Snyder, G.D. and
Balonon-Rosen, M.R.,
1991. "Visual Determination of Supra-Threshold Color-Difference Tolerances Using Probit Analysis," **Col. Res. Appl., 16 No. 5,** 297-316.

Berns, R.S., Gorzynski, M.E., and Motta, R.J.,
1991a "CRT Metrology and Colormetric Characterization Techniques," **MCSL Technical Report**.

Berns, R.S., and Gorzynski, M.E.,
1991b "Characterizing the total uncertainty of the colorimetric calibration of color video displays," **proceedings of the 22nd Session of the CIE, part I,** 35-38.

Berns, R.S., and R.J. Motta,
1988 "Colorimetric calibration of soft-copy devices to aid in hard-copy reproduction," **Proceedings SPSE 41st Annual Conference**, 266-269.

CIE,
1986 "Colorimetry, Second Edition (Official Recommendations of the International Commission on Illumination)," **CIE Publ. 15.2**, Central Bureau of the CIE, Vienna.

Corey, G.P., Clayton, M.J., and Cupery, K.N.,
1983. "Scene Dependence of Image Quality," **Phot. Sci. Eng., 27**, 9-13.

Evans, R.M., Hanson, W.T., and Brewer, W. L.,
1953. **Principles of Color Photography**, Wiley, New York.

Fairchild, M D.,
1991. "Chromatic Adaptation and Color Constancy," **Advances in Color Vision Technical Digest 4** (OSA, Washington, D. C.) 112-114.

Finney, D.J.,
1971. **Probit Analysis**, 3d ed., Cambridge U. Press, Cambridge.

Jones, L.A., and Condit, H.R.,
1941. "The Brightness Scale of Exterior Scenes and the Computation of Correct Photographic Exposure," **J. Opt. Soc. Am., 31**, 651-678.

Luo, M.R., Clarke, A.A., Rhodes, P.A., Schappo, A., Scrivener, S.A.R., and Tait, C.J.,
1991. "Quantifying Colour Appearance.

Part I. LUTCHI Color Appearance Data,"
Col. Res. Appl., 16 No. 3, 166-180.

McCamy, C. S.,
1976. "A Color-Rendition Chart",
J. Appl. Phot. Eng. 2 No. 3, 95-99.

McDonald, R.,
1988. "Acceptability and perceptibility decisions using the CMC color difference formula,"
Tex. Chem. Col. 20 No. 6, 31-37.

Munsell, A. H.,
1979. **A Color Notation**, 13 E d., Munsell Color, Baltimore.

SAS Institute,
1990. **SAS/STAT User's Guide,** Version 6, Fourth Ed.

Stamm, S.,
1981. "An Investigation of Color Tolerance,"
Proc. Tech. Assoc. Graphic Arts, 156-173.

Stokes, M.,
1991. "Colorimetric Tolerances of Digital Images,"
Master's Thesis, Rochester Institute of Technology.