https://doi.org/10.2352/EI.2026.38..IQSP-247

Validation of Skin Tone Test Charts with Real Human Data

Megan Borek[1], Amelia Limbocker[1], Ellis Monk[2]
[1]Imatest LLC; Boulder, Colorado, USA
[2]Harvard University; Cambridge, Massachusetts, USA

Abstract

Photographic test charts for measuring color accuracy in cameras have historically included a limited number of skin tones, typically in the form of uniform color patches. Such charts are not representative of the wide range of skin tones found in humans, and do not test the behavior of modern automatic exposure, white balance, and focus (3A) algorithms that are commonly driven by facial detection in today's digital consumer cameras. We built upon our previous work on the development of printed skin tone charts featuring detectable faces by conducting a study with human participants whose skin tones approximately span the Monk Skin Tone Scale. Participants were photographed under a series of controlled lighting conditions, and each scene was then reproduced using a high-resolution inkjet print of the participant. Corresponding captures of the human subjects and the printed charts were quantitatively compared by calculating the CIEDE2000 color difference for regions of interest across the subject's face in the scene. This analysis evaluates how printed skin tones behave across exposure settings and lighting conditions relative to real skin, with the goal of determining whether printed charts provide a suitable solution for repeatable, lab-based image quality testing in face-present scenes. While not intended to replace final field testing with real human subjects, results indicate that face charts printed with sufficiently wide-gamut printers can provide an effective solution for lab testing and benchmarking of color accuracy and 3A behavior in a controlled and repeatable manner.

Background

The use of skin tones in image quality testing dates back to film color calibration in the mid-20th century. One of the first mainstream examples was Kodak's use of the "Shirley Card", introduced in the 1950s, which featured a female model with a light skin tone positioned with colored objects, clothing, or patches on a reference print used by photo labs to calibrate printing machines [1]. These remained the standard until the first multi-racial Shirley Card was introduced more than 40 years later in 1995 [1]. As a result, darker skin tones and objects were often poorly rendered, lacking contrast and detail, appearing underexposed, or exhibiting hue shifts.

Although the use of the Shirley card declined with the rise of digital photography, many of these issues have persisted in modern imaging systems, in part due to the continued lack of camera testing and image signal processor (ISP) tuning for scenes containing subjects with a wide range of skin tones.

The Calibrite ColorChecker, originally introduced as the Macbeth ColorChecker in 1976, is a classic 24-patch color target that remains widely used in modern image quality testing for measuring color reproduction accuracy and performing color correction [2]. The chart includes only two skin tone patches (light and dark), representing a limited subset of the wide range of melanin content and undertones found in human skin. This leaves the lightest, darkest, and many mid-tones underrepresented in the color characterization and calibration process. Wider-gamut charts, such as the 140-patch ColorChecker Digital SG, have been introduced for testing digital cameras and contain a larger variety of skin tones, but remain less widely used than the classic ColorChecker chart.

In the past decade, increased attention has been given to improving skin tone reproduction in consumer cameras across a broader range of tones. Examples include Google's Real Tone technology in Pixel smartphones [3], image quality benchmarking efforts such as the Valued Camera eXperience (VCX) standard [4], and the development of new chart designs by companies such as Imatest and Image Engineering that are being incorporated into camera evaluation workflows.

A key limitation of traditional patch-based charts is their inability to evaluate scene-dependent image processing. In contrast, charts that include detectable faces enable testing of automatic exposure, white balance, and focus (3A) algorithms that depend on scene content. Many modern cameras, including smartphones and webcams, use face detection to dynamically adjust ISP behavior. For example, when a face is detected, the camera may estimate facial lightness and adjust exposure accordingly, increasing exposure for darker skin tones or decreasing it for lighter skin tones relative to a global exposure. These behaviors cannot be fully characterized using uniform patch-based charts.

Printed face charts allow for controlled, repeatable testing of these face-dependent behaviors in a laboratory setting, while providing ground truth references and enabling automated workflows that are difficult to achieve with real human subjects.

Related Work

Recent work has highlighted limitations in traditional image quality testing for face-present scenes, particularly with respect to skin tone diversity and the behavior of automatic image processing algorithms. In prior work, we evaluated camera performance in scenes containing both light and dark skin tone mannequin heads using methods based on the VCX Webcam standard [4]. We demonstrated that automatic exposure and white balance algorithms can produce significantly different results depending on the skin tone of a detected face in the scene. These findings showed that commonly used patch-based color error metrics do not fully capture system behavior when face detection is active [5].

Subsequent work introduced printed face charts based on the Monk Skin Tone (MST) Scale. We investigated the applicability of the 10-tone scale, and the spectral and colorimetric differences between printed and real skin, highlighting challenges in accurately reproducing skin reflectance and exposure behavior using inkjet printing methods [6].

More broadly, industry efforts have begun addressing fairness in imaging systems through both algorithm development and benchmarking. The MST Scale has been adopted in computer vision and image research to better represent diverse populations [3], while standards such as VCX incorporate skin tone testing into rigorous objective evaluation protocols [4]. However, existing test

methodologies still rely heavily on a small number of skin tones or lack the ability to evaluate face-dependent processing in a controlled and repeatable manner. This work builds on these efforts by performing a validation study of printed face charts as a means to improve lab-based testing by comparing chart behavior to real human skin under matched conditions.

Methods

Data Collection

Validation of the printed charts was conducted by comparing their behavior across lighting conditions and exposure settings to that of real human subjects. The study included 10 participants (see Figure 1. Participants were selected to represent the 10 tones of the Monk Skin Tone Scale whose skin tones approximately span the tones of the Monk Skin Tone Scale from 1 (lightest) to 10 (darkest). Participant selection was guided by Dr. Ellis Monk, creator of the scale. Participants, aged 17-32, included 3 males and 7 females.

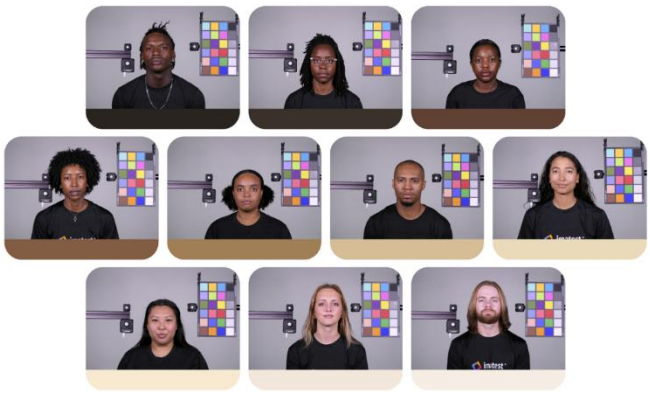


Figure 1. *Participants were selected to represent the 10 tones of the Monk Skin Tone Scale.*

Images were captured using three devices: a Panasonic Lumix DMC-FZ2500 (primary camera) with a Leica 8.8-176mm f/2.8-4.5 lens, a Samsung Galaxy S25+, and a Google Pixel 9 Pro. The Panasonic camera captured RAW and JPG images at a 50 mm equivalent focal length. The smartphones captured processed images using fully automatic settings.

Each participant was photographed in a landscape head-and-shoulders scene with a neutral gray background. Each scene included a Calibrite ColorChecker Classic chart, two lux meters positioned on either side of the participant's head, and a spectral illuminance color sensor. Participants were photographed under 12 different lighting conditions, including three light intensities (200 lux, 1000 lux, 3000 lux), and four correlated color temperatures (CCTs) (3200 K, 4000 K, 5000 K, 5500 K).

For each condition, the primary camera used two exposure techniques. The first was a gray card exposure, in which exposure was set according to a neutral reference target in the scene. With this approach, all participants were photographed using identical exposure settings for each lighting condition. The second method used spot metering, where exposure was determined from a manually selected region of the face, resulting in exposure settings that varied with subject skin tone. Exposure bracketing was applied in both cases (±1 stop in 1/3-stop increments; 7 images per condition). Due to overlap between the two methods, only results from the gray card exposures are shown in the results.

Spectral reflectance measurements were obtained using a Nix Spectro 2 spectrophotometer, with contact measurements taken from the cheek, forehead, and forearm.

Following image capture, a subjective evaluation was conducted. For each lighting condition and exposure method, participants selected their preferred image from each seven-image JPEG exposure bracket. These responses were recorded for analysis of exposure preference across skin tones for use in future studies.

Printing

A printed chart was made for each participant using the 3000 lux, 5000 K, face-based exposure RAW images captured with the Panasonic camera. The RW2 files (Panasonic RAW format) were converted to TIFF using LibRaw. This process included demosaicing and white balance correction based on camera metadata, but did not apply a gamma curve. A color correction matrix (CCM) was then computed using the ColorChecker patches within each scene. The CCM was derived using a least-squares optimization to minimize color error between measured patch values and reference CIELAB values.

Basic image processing, including denoising, sharpening, and contrast adjustment, was applied to the color-corrected images using Adobe Photoshop. Images were resized to life-size scale and exported in the Adobe RGB (1998) color space for compatibility with the printing workflow. These nominal corrected images then entered an iterative printing process (see Figure 2) to further tune the printed skin tones.

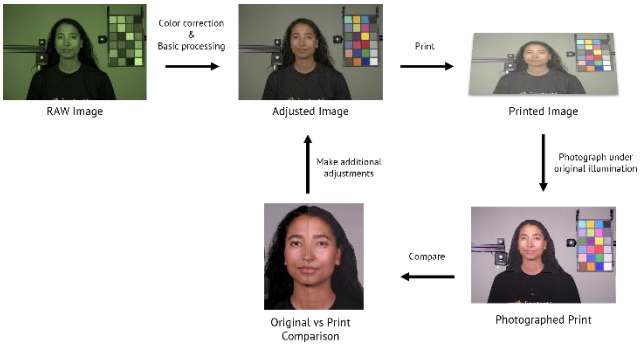


Figure 2. *Overview of iterative printing process used for tuning printed skin tones.*

During the iterative printing process, the adjusted images were printed on ultra-smooth, high color gamut matte photo paper using commercial extended-gamut printers. Two printers with different color gamut characteristics were used. Tones 1-6 were printed on a 9-color printer with improved performance at higher L* values, while tones 7-10 were printed on a 7-color printer with relatively better performance at lower L* values.

The tuning process was performed using the 3000 lux, 5000 K condition as the reference. This condition was selected because it provides high signal-to-noise ratio and stable illumination while remaining representative of common indoor lighting, making it well suited for establishing a reliable baseline for color and exposure matching. The goal of this process was to produce a single printed chart for each participant whose skin tone behaves as similarly as possible to the real subject across a range of conditions, rather than

optimizing separately for each lighting scenario. By tuning the prints under a single, well-controlled reference condition, we ensure that any observed differences under varying illumination and exposure settings reflect the inherent behavior of the printed charts, rather than scene-specific optimization.

After printing, each face was mounted and placed back into the original scene, where it was re-imaged using the Panasonic camera under all lighting conditions. A custom Python tool was developed to analyze and compare images of the human participants and their printed counterparts. The analysis workflow consisted of four main steps: (1) color correction, (2) image registration, (3) skin tone segmentation, and (4) ROI-based comparison.

Analysis

Color Correction

To reduce the impact of small illumination differences between captures of human participants and printed charts, each image pair is color corrected using the ColorChecker chart present in the scene. This step minimizes residual lighting and color balance differences prior to comparison. For each image, a global 3×3 linear color correction matrix (CCM) is estimated using weighted least squares and the 24 patches of the detected ColorChecker target. Mean RGB values are computed for each patch and linearized when processed images are analyzed. The CCM is then solved by minimizing the weighted squared error between corrected patch values and reference ColorChecker values in linear RGB space.

Patch weights control the relative influence of each ColorChecker patch during optimization. In this configuration, the two skin tone patches are assigned the highest weight to prioritize skin tone fidelity, neutral grayscale patches are given intermediate weight to improve gray balance, and remaining chromatic patches retain baseline weight to preserve overall color consistency. As a result, the fitted CCM is intentionally biased toward minimizing error in skin and neutral regions while maintaining a globally constrained solution.

Image Registration

Before segmentation, each image pair is spatially aligned to ensure that corresponding regions of interest (ROIs) represent the same anatomical locations. Image registration is performed by treating the original image as the moving frame and estimating a geometric transform into the reference image coordinate system. The pipeline supports ORB or SIFT feature detection, followed by descriptor matching using a nearest-neighbor ratio test. Outliers are suppressed by retaining only the best matches prior to robust model fitting. A homography is then estimated using RANSAC and applied via perspective warping, bringing both images into a shared spatial frame. This step is critical to ensure that downstream measurements such as skin masks, ROI sampling, and color difference calculations are computed from spatially corresponding regions rather than misaligned pixels. For efficiency, the implementation optionally reuses the homography from the first image pair when camera geometry is fixed, or when processing underexposed images where feature matching is less reliable.

Skin Tone Segmentation

Since skin tones are the primary target of comparison, they are segmented from surrounding image content using a combination of CIELAB thresholding and spatial ROI selection. This process removes non-skin regions such as hair, clothing, background, and the ColorChecker chart. CIELAB thresholds are manually defined for each participant and further refined for each lighting condition and image format (RAW and JPG). An elliptical region is additionally applied to isolate the facial area and exclude features not representative of skin tone.

ROI Comparison

After segmentation, a binary skin mask is generated and divided into 40 × 40 pixel ROIs, corresponding to approximately 5 mm × 5 mm areas on the physical face or printed chart. Only ROIs in which all pixels fell within the user-defined skin tone thresholds were used for analysis. For each valid ROI, color differences and related metrics are computed between corresponding regions in the aligned image pairs, and aggregate statistics are computed across ROIs for each condition.

Results

Results indicate that printed face charts can approximate the color behavior of real human skin under controlled conditions, particularly for light to medium skin tones. Across most lighting conditions and exposure levels, mean color differences remain within acceptable thresholds. Deviations from this behavior are primarily observed under low-light conditions, nonuniform illumination, and for darker skin tones, where limitations in printer gamut and substrate properties become more pronounced.

The primary metric used for comparison is CIEDE2000, or ΔE_{00}, which quantifies perceptual color differences based on lightness (L*), chroma (C*), and hue (h°). A ΔE_{00} of ~1 approximately corresponds to the just noticeable difference (JND) under typical viewing conditions, values between 1-3 are generally considered good, values between 3-5 acceptable, and values above 5 indicate clearly visible differences.

To quantify uncertainty in the reported mean facial color difference, we performed Monte Carlo propagation of residual color error estimated from the ColorChecker. After CCM correction, patch-level residuals were computed for each image relative to the same reference in CIELAB space, and the difference between the two residual sets was used to estimate a global mismatch distribution (mean vector and covariance). We then drew repeated samples from this distribution and, for each sample, applied the same LAB perturbation to all facial ROIs in one image to represent plausible capture-to-capture calibration drift. For every perturbation, per-ROI ΔE_{00} values were recomputed and averaged, yielding a sampling distribution of the mean ΔE_{00} across ROIs. From this distribution, we report the Monte Carlo mean, standard deviation, and 95% confidence interval, which indicate how sensitive the observed color-difference result is to residual calibration mismatch.

Performance Across RAW Images

The heatmap in Figure 3 shows the mean ΔE_{00} across all exposure levels for demosaiced linear RAW images, using gray card exposure. Most values fall below the $\Delta E_{00} = 5$ threshold, indicating acceptable agreement between real and printed skin across a wide range of lighting conditions. Higher errors are observed in several 200 lux conditions, particularly for the lightest skin tones.

Nonuniform lighting in the original scene in Figure 4 was unintended, but reveals an important distinction between real faces and flat charts. Human faces are three-dimensional, so their geometry interacts with lighting to produce spatial variation across the surface. In contrast, printed charts are flat and are intended to be used under uniform illumination. The aggregate heatmap reveals where certain lighting conditions cause the correlation between real and printed faces to break down, but it is also important to evaluate how this relationship varies across exposure.

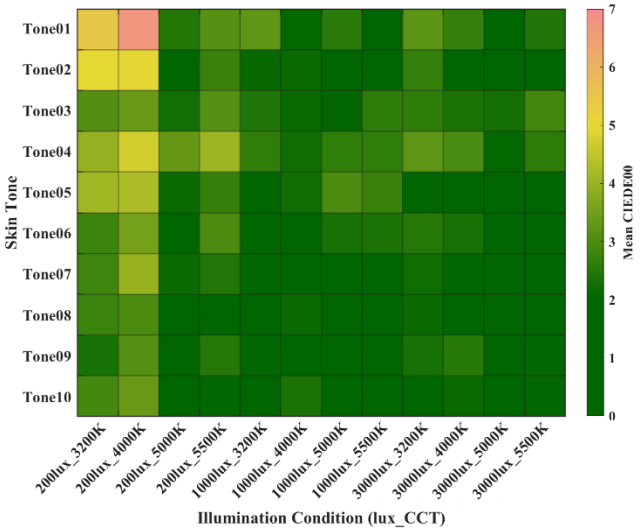


Figure 3. *Mean CIEDE2000 heatmap averaged across all exposures for RAW images captured with Panasonic Lumix DMC-FZ2500.*

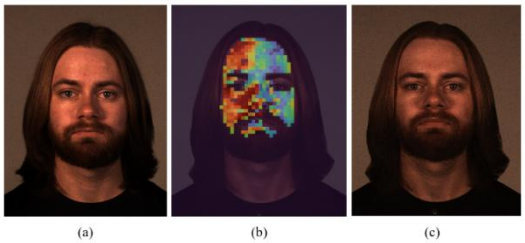


Figure 4: *Nominal exposure images for Tone 1 at 200 lux 4000 K CCT, including (a) original image, (b) CIEDE2000 heatmap of differences between original and print, and (c) printed chart.*

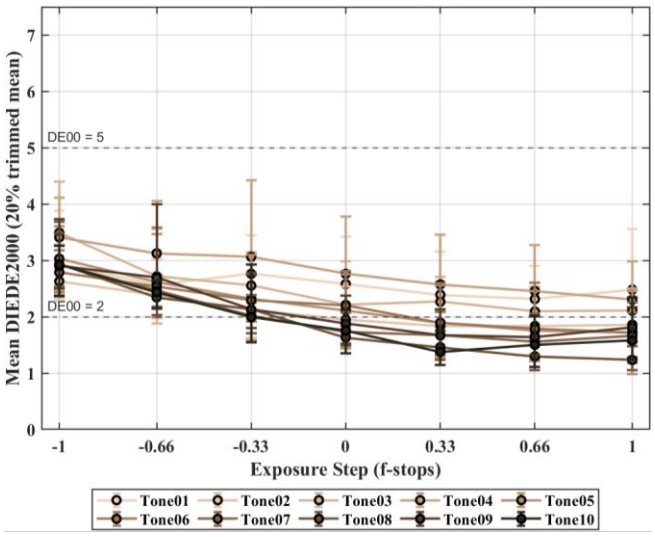


Figure 5: *Mean CIEDE2000 vs. exposure step for each tone from RAW images, aggregated across lighting conditions. 200 lux captures at 3200 K and 4000 K are excluded due to nonuniform illumination. Error bars represent the 10th-90th percentile range across lighting conditions (lux, CCT), and markers indicate the 20% trimmed mean.*

Figure 5 shows the mean ΔE_{00} for each tone across exposure bracket intervals, aggregated over all lighting conditions, excluding the 200 lux captures at 3200 K and 4000 K due to the nonuniform lighting described previously. For RAW images, error remains relatively stable across exposure levels, with a slight increase at lower exposures where reduced signal-to-noise ratio increases noise and channel imbalance in the RAW signal. After demosaicing, this manifests as increased hue error, contributing to higher ΔE_{00} and greater variability across lighting conditions, even when perceptual differences may be less pronounced at low luminance.

Performance Across Processed Images

While analysis in linear RAW space is useful for understanding sensor-level behavior, ΔE_{00} comparisons are more meaningful for processed images intended for viewing. The same analyses were therefore performed on the JPG images from the Panasonic camera.

The heatmap in Figure 6, averaged across all exposures, again shows higher ΔE_{00} error for the first two 200 lux lighting conditions, consistent with the nonuniform illumination described previously. The heatmap also reveals higher error for Tones 8-10 across most lighting conditions, which is a deviation from the RAW data.

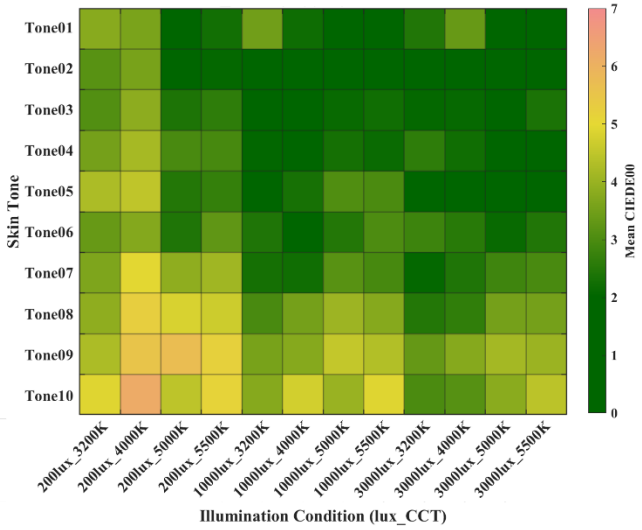


Figure 6. *Mean CIEDE2000 averaged across all exposures for JPG images captured with Panasonic Lumix DMC-FZ2500.*

Figure 7 shows the mean ΔE_{00} for each tone as a function of exposure change, aggregated across lighting conditions (excluding the affected 200 lux cases). A clear trend is observed for tones 7-10, which exhibit increasing error with increasing exposure. This indicates that printed charts diverge from real skin behavior in overexposed conditions, particularly for darker skin tones. Inspection of ROI heatmaps (see Figure 8) for Tone 9 reveals that error is not uniformly distributed across the face. Increased error is observed in shadowed regions and along facial edges at higher exposure levels. This behavior reflects another limitation of printed charts: they do not fully replicate the interaction between surface geometry, subsurface scattering, and directional illumination present in real skin.

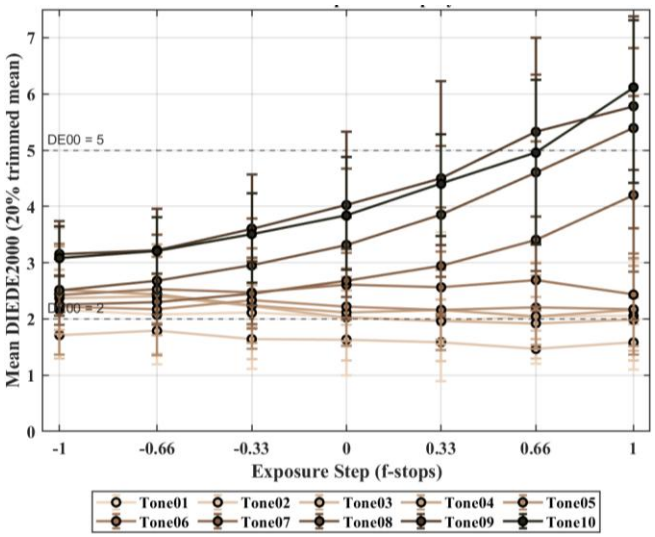


Figure 7. *Mean CIEDE2000 vs. exposure step for each tone from JPG images, aggregated across lighting conditions. 200 lux captures at 3200 K and 4000 K are excluded due to nonuniform illumination. Error bars represent the 10th-90th percentile range across lighting conditions (lux, CCT), and markers indicate the 20% trimmed mean.*

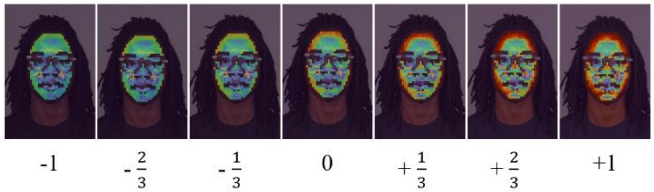


Figure 8. *ROI heatmaps for full exposure bracket set for Tone 9 at 1000 lux, 5000 K CCT. Error increases at the darker regions of the face (edges and shadows) as exposure increases.*

An example of this behavior is shown in Figure 9. The ROI pair has a high ΔE_{00} of 11.7. The ROI extracted from the printed chart appears significantly lighter and less saturated than the corresponding region from the real subject. This results in a visibly washed-out appearance relative to the original skin tone. This behavior can be attributed to two primary factors: (1) limitations in printer gamut at low L* values, and (2) differences in the physical properties of the printing substrate relative to real skin.

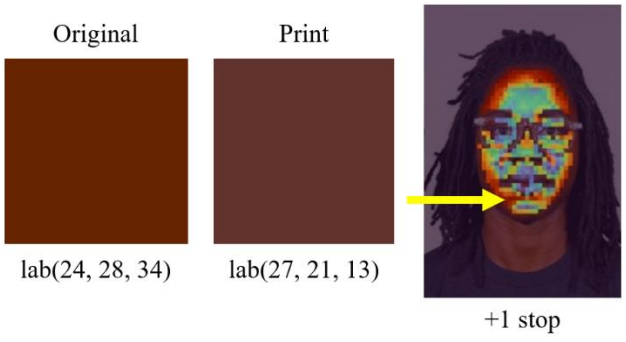


Figure 9. *Example ROI comparison from original vs printed images, extracted from the edge of the face from Tone 9 at the highest exposure bracket interval (+1 stop) captures at 1000 lux 5000 K.*

Although extended-gamut printers (CMYK+) provide wide color coverage, their ability to reproduce high saturation decreases at lower lightness levels, particularly on matte photo paper. As shown in Figure 10, some target colors fall outside the achievable gamut and therefore cannot be accurately reproduced.

In addition to gamut limitations, the matte paper substrate contributes to reduced contrast and chroma. Unlike glossy or luster papers, matte paper absorbs ink rather than allowing it to remain on the surface. While this reduces specular reflections and is desirable for imaging applications, it also causes increased light scattering, resulting in colors that appear lighter and less saturated than the corresponding real skin tones.

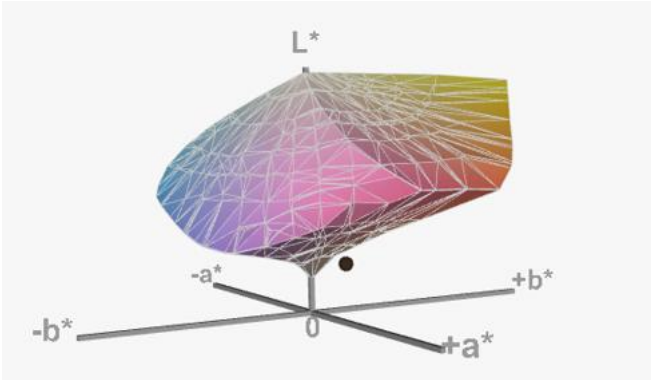


Figure 10. *Gamut volume for one of the printer/paper combinations used to print the experimental charts. The spot color is an example of a target color that falls outside of the gamut, and thus could not be accurately reproduced.*

Performance in Smartphone Images

CIEDE2000 heatmaps for the Google Pixel 9 Pro and the Samsung Galaxy S25+ are shown in Figs. 11 and 12, respectively. A single image was captured per lighting condition using automatic exposure. Across all conditions, ΔE_{00} values remained below 5, indicating close agreement between real and printed charts. Residual variation is primarily attributed to imperfect feature alignment due to differences in framing, subject motion, and the use of a Panasonic reference image for creating the chart.

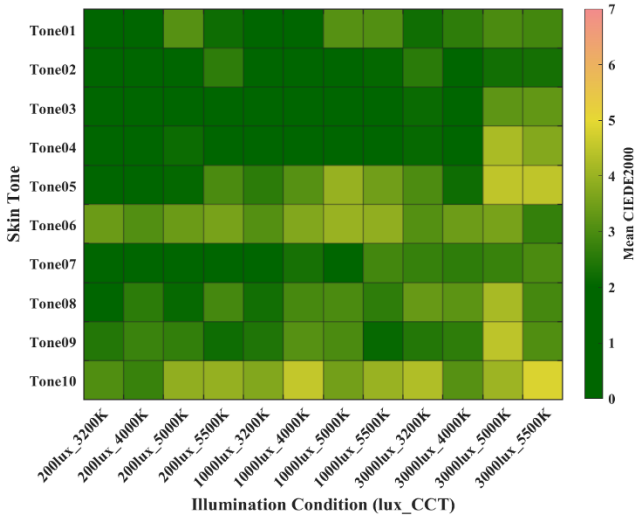


Figure 11. *Mean CIEDE2000 averaged across all exposures for JPG images captured with the Google Pixel 9Pro.*

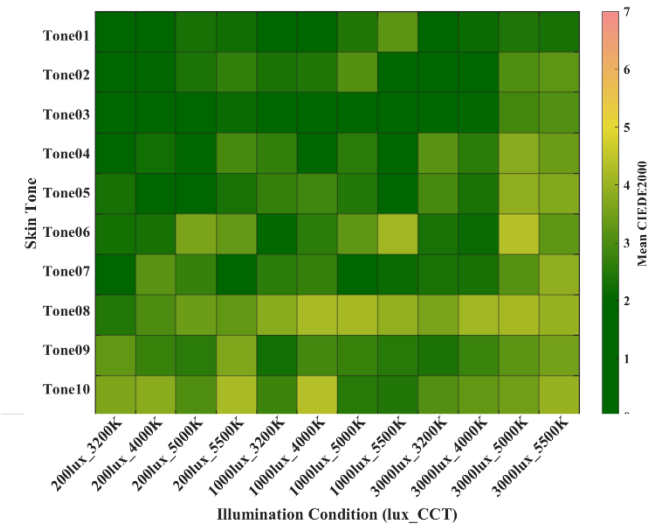


Figure 12. *Mean CIEDE2000 averaged across all exposures for JPG images captured with the Samsung Galaxy S25+.*

Discussion

Overall, the results demonstrate that printed face charts can serve as a useful and repeatable tool for evaluating color and exposure behavior in face-present scenes under controlled laboratory conditions. For light to medium skin tones, the printed charts closely track the behavior of real skin across a range of lighting conditions and exposure levels, with most ΔE_{00} falling within acceptable thresholds.

However, the validity of these charts is dependent on the testing conditions. The charts perform best under uniform illumination and moderate exposure levels, where the assumptions of flat, evenly illuminated targets are satisfied. Under these conditions, they provide a practical means of introducing diverse, detectable faces into image quality workflows.

Limitations arise in scenarios that deviate from these assumptions. Nonuniform illumination introduces discrepancies due to the inability of flat charts to replicate three-dimensional light interactions. Additionally, for darker skin tones, increasing error under higher exposure levels reflects limitations in printer gamut and substrate properties, which reduce saturation and dynamic range at low L* values. Exploration of alternative substrates and printing methods will be explored in future work.

These results indicate that printed face charts are not intended to replicate all aspects of real human appearance, but rather to provide a controlled and repeatable approximation that improves upon existing test methods. When used within their valid operating conditions—uniform illumination, moderate exposure, and controlled capture environments—they offer a meaningful approach for expanding skin tone diversity in lab-based image quality testing.

Conclusion

This work demonstrates that printed skin tone charts can effectively approximate the color behavior of real human faces across a range of lighting and exposure conditions, enabling repeatable and controlled evaluation of face-dependent camera systems. Across most conditions, printed charts produced color differences within acceptable thresholds ($\Delta E_{00} < 5$), and exhibited similar exposure-dependent trends to real human subjects, particularly in the RAW domain.

Several limitations were identified. Under nonuniform illumination, flat printed charts do not replicate the geometry of real faces, leading to localized color error. Discrepancies also increase in low-light conditions and at extreme exposure levels, where noise and chromatic instability become more significant. Differences between RAW and processed images further indicate that downstream image processing plays a critical role in tone-dependent color error, particularly for darker skin tones.

Despite these limitations, printed face charts provide a practical and scalable solution for controlled, repeatable testing of face-based imaging pipelines, offering a valuable complement to studies involving real human subjects.

References

[1] L. Roth, “Looking at Shirley, the Ultimate Norm: Color Balance, Image Technologies, and Cognitive Equity.” Canadian Journal of Communication, 2009.

[2] D. Pascale, “A Review of RGB Color Spaces,” *The BabelColor Company*, 2003.

[3] Google Research, “Towards more inclusive camera systems,” Google AI Blog, 2021.

[4] VCX Forum, *VCX : Valued Camera eXperience Benchmark Specification*, 2023.

[5] M. Borek, “Evaluating Camera Performance in Face-Present Scenes With Diverse Skin Tones” in *Electronic Imaging*, 2024.

[6] M. Borek, “Improving Image Equity: Representing Diverse Skin Tones in Photographic Test Charts for Digital Camera Characterization” in Electronic Imaging, 2025.

Author Biography

Megan Borek received her BS in Imaging Science from the Rochester Institute of Technology (2022). Since then, she has worked as an Imaging Scientist at Imatest LLC in Boulder, Colorado, where she works on the development of image quality testing software, contributes to image quality standards within ISO TC42, and leads research on skin tone reproduction in consumer cameras.

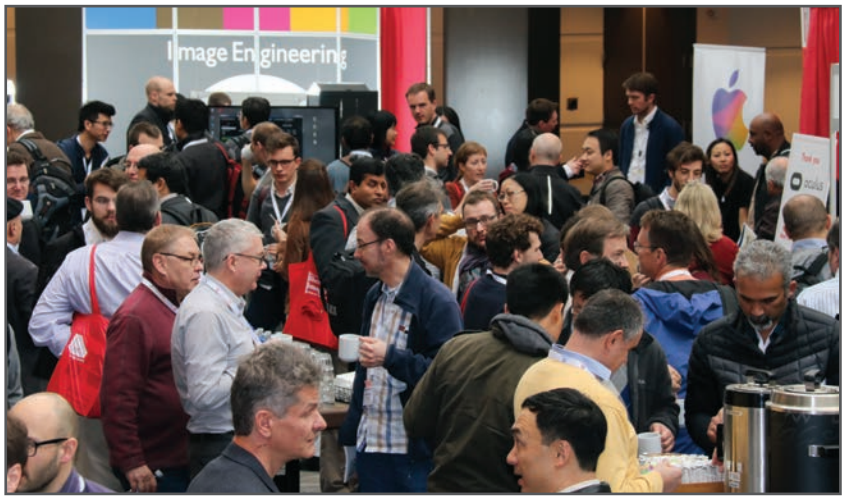
Amelia Limbocker received her BFA in Photography with a minor in English from Virginia Intermont College (2014), and a BS in Photographic and Imaging Science with a minor in Imaging Systems from Rochester Institute of Technology (2016). She currently works as an Imaging Science Engineer at Imatest, designing custom charts, providing chart recommendations, and ensuring chart quality.

Ellis Monk is Professor of Sociology at Harvard University. He earned his PhD in Sociology from the University of California, Berkeley and a BA in Sociology from the University of Michigan, Ann Arbor. He previously taught at the University of Chicago and Princeton University. His research focuses on the comparative examination of social inequality, especially with respect to race and ethnicity, in global perspective. By deeply engaging with issues of measurement and methodology, it examines the complex relationships between social categories and social inequality; and extends into topics such as social demography, health, aging, social psychology, sociology of the body, political sociology, and comparative/historical sociology.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org