SKIN TONE AND MEDICAL DEVICES:
WHY MEASUREMENT MATTERS

Dr. Ellis Monk
Professor
Department of Sociology
COLORISM

- **Census race/ethnicity and skin tone are NOT** the same characteristic.

- There is considerable heterogeneity in skin tone and other phenotypical features associated with race/ethnicity within and across Census race/ethnicity categories.

- Evidence shows, globally, that **skin tone is significantly associated with education, earnings, employment, health, and more in many countries around the world**. Research shows, including my own, discrimination is one factor that helps explain these inequalities.

- While **much attention is often given to race/ethnicity and racism, relatively little attention is paid to color and colorism**, even though it has **massive effects on inequality all around the world**.
“Race,” Color, and Education in the U.S.

Age: 25-44

BLACK-WHITE [NHIS]  10.2 months

LIGHT-DARK [NSAL]  15.4 months
COLOR-BLIND TECHNOLOGY
Autonomous Cars Can't Recognize Pedestrians With Darker Skin Tones

People with darker skin are more at risk of being hit by a self-driving vehicle.

By Jessica Miley

Apr 21, 2021 | Updated: Aug 09, 2021 09:56 EDT
PULSE OXIMETRY REGULATIONS

- 2 subjects or 15% of the pool are required to be darkly pigmented.

- Who is “darkly-pigmented”?
**Fitzpatrick Scale**

- Designed in 1975 by Dr. Thomas Fitzpatrick (Harvard Medical School).
- Intended to categorize how skin reacts to UV during phototherapy for skin conditions. Inconsistent even when used as intended.
Studies find, for example, that the Fitzpatrick Scale "excludes the majority of Blacks and yields data that overestimate Black population prevalence of type IV skin" (Pichon et al. 2010); and that the "[Fitzpatrick Scale] provides a restricted range of options for people with darker skin tones that do not capture variations in their skin color."

It is also worth noting that some studies find that the Fitzpatrick Scale performs poorly, even when used as intended (see above), especially on ethnoracial minorities (see Eilers et al. 2013; He et al. 2014).
Ironically, given its selection of skin tones, which lives in a very restricted ‘intermediate’ zone, it may be simultaneously **too dark** for many lighter-skinned people and **not dark enough** for many darker-skinned people.
Revisiting the Fitzpatrick Scale and Face Photo-based Estimates of Skin Phenotypes

October 29, 2020

John Howard, Yevgeniy Sirotin, & Jerry Tipton

The Maryland Test Facility

Arun Vemury

Director

Biometric and Identity Technology Center

Science and Technology Directorate
Rethinking Fitzpatrick

- FST is a questionnaire originally designed to determine the appropriate dose of oral methoxsalen for treating psoriasis using photochemotherapy in white individuals [1].

- FST is not skin color, in fact FST is known to be a generally unreliable estimator of skin pigmentation.

- The FST was developed explicitly because dosing based on observed phenotypes (hair and eye color) led to medical error.

- There is mounting evidence from the medical community that FST can be less reliable as an assessment for non-White individuals.

Conclusions

- The computer vision community recently began categorizing skin phenotypes in images using 6-point scales referred to as “Fitzpatrick Skin Types”.

- Calling these measures FST is problematic for the following reasons:
  - As originally developed, FST is assessed by a survey to measure sensitivity to UV light.
  - FST is measured by self-report or by a physician direct assessment.
  - FST as originally defined is not an appropriate measure of skin color.
  - FST has been shown in the medical literature to be an unreliable estimator of skin pigmentation.

- All existing work applying FST to computer vision has involved human raters judging the skin pigmentation of subjects in images.

- 6-point skin type classifications schemes have been conflated with FST.

- These measures likely do not reflect FST.

Table

<table>
<thead>
<tr>
<th>Study</th>
<th>Car Classification</th>
<th>Face skin phenotype measure</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender Shades: International Accuracy</td>
<td>2018 Gender classification</td>
<td>Fitzpatrick skin type (FST) assessed from analyzed sample</td>
<td>Images of women with FST IV-VI misclassified more than those with FST I-III.</td>
</tr>
<tr>
<td>Understanding Unequal Gender Accuracy from Face Images (Cook et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An Experimental Evaluation of Unconstrained Face Images (Lu et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Cards for Model Replication (Et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive Inequity in Object Recognition (Krishnapriya et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographic Effects in Face Recognition and Their Dependence on Image Quality (Cook et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An Evaluation of Eleven Color System (Crone et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Issues Related to Face Recognition: Varying Based on Race and Skin Tone (Krishnapriya et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Krishnapriya et al.

If we are reaching a consensus standard measure, is it the right one?

And are we measuring it the right way?

... the Fitzpatrick I–VI skin tone rating is the appropriate choice for this article due to its simplicity and widespread use, including prior use in the face recognition research community; e.g., metadata for face images in the IARPA IJB datasets [32], work by Buolamwini and Gebru [7], Lu et al. [30], and Muthukumar et al. [34].

- Krishnapriya et al.
MONK SKIN TONE SCALE

- Similar to Massey-Martin, it was explicitly designed to measure skin tone in diverse populations.

- Intended to be an easy to use (e.g., an optimal number of choices), reliable, and cost-effective means of measuring skin tone.

- The main way the Monk Scale mitigates biases relative to prior visual scales is by including a wider range of carefully selected skin tones to better represent the dynamic range of skin tones we see in the United States (and beyond).
MONK SKIN TONE SCALE

- Color selection based on extensive fieldwork in the U.S. & Brazil, computer software that creates facial stimuli for social psychological experiments (i.e., skin reflectance spectrum scores), maps of the distribution of UV exposure and human skin tone around the world.

- Validated through cognitive interviewing (NIA/NIH funded research with National Social Life Health and Aging Project) and nationally-representative surveys.

- Adopted for data collection in 2021 (W4 NSHAP).
Monk Scale

How to use: This scale is intended to be used to classify human skin color. Human skin color is immensely variable and complex, but scales, such as this one, are often used as a pragmatic and cost-effective means of approximately measuring common differences in skin color among human beings.

Print this page using color accurate printing. Once printed, consider using a hole punch to punch a hole in the center of each circle and the center of each rectangle. Hold the printed card over the skin at the site you wish to describe. Move the card to find the color that matched most closely and note the category letter (e.g., ‘E’). The circles or squares can be used and you should note in your methods which you have chosen. As you may notice the circles have more shades per category, which better resemble human skin and may aid in finding the closest match.
In partnership with Dr. Ellis Monk, we’re releasing a new skin tone scale designed to be more inclusive of the full spectrum of skin tones — the next step in our commitment to image equity and improving representation across our products. #GoogleIO
Improving skin tone representation in Google Photos

We’ll also be using the MST Scale to improve Google Photos. Last year, we introduced an improvement to our auto enhance feature in partnership with professional image makers. Now we’re launching a new set of Real Tone filters that are designed to work well across skin tones and evaluated using the MST Scale. We worked with a diverse range of renowned image makers, like Kennedi Carter and Joshua Kissi, who are celebrated for beautiful and accurate depictions of their subjects, to evaluate, test and build these filters. These new Real Tone filters allow you to choose from a wider assortment of looks and find one that reflects your style. Real Tone filters will be rolling out on Google Photos across Android, iOS and Web in the coming weeks.
U.S. SCALE VALIDATION RESEARCH
Which Skin Tone Measures are the Most Inclusive? An Investigation of Skin Tone Measures for Artificial Intelligence

COURTNEY M. HELDRETH 1*
Google Research, Seattle, cheldreth@google.com

ELLIS P. MONK
Harvard University, emonk@fas.harvard.edu

ALAN T. CLARK
The Value Engineers, alan.clark@thevalueengineer.com

SUSANNA RICCO
Google Research, Seattle, ricco@google.com

XANGO EYÉE
Google Research, Seattle, xango@google.com

Skin tone plays a critical role in artificial intelligence (AI). However, many algorithms have exhibited unfair bias against people with darker skin tones. One reason this occurs is a poor understanding of how well the scales we use to measure and account for skin tone in AI actually represent the variation of skin tones in people affected by these systems. To address this, we conducted a survey with 2,214 people in the United States to compare three skin tone scales: The Fitzpatrick 6-point scale, Rihanna’s Fenty™ Beauty 40-point skin tone palette, and a newly developed Monk 10-point scale from the social sciences. We find that the Fitzpatrick scale is perceived to be less inclusive than the Fenty and Monk skin tone scales, and this was especially true for people from historically marginalized communities (i.e., people with darker skin tones, BIPOCs, and women). We also find no statistically meaningful differences in perceived representation across the Monk skin tone scale and the Fenty Beauty palette. We discuss the ways in which our findings can advance the understanding of skin tone in both the social science and machine learning communities.
Consensus and Subjectivity of Skin Tone Annotation for ML Fairness

Candice Schumann  
Google  
United States  
cschumann@google.com

Ellis Monk, Jr.  
Harvard University  
United States  
emonk@fas.harvard.edu

Gbolahan O. Olanubi  
Google  
United States  
femio@google.com

Courtney Heldreth  
Google  
United States  
cheldreth@google.com

Auriel Wright  
Google  
United States  
aurielwright@google.com

Susanna Ricco  
Google  
United States  
rcco@google.com
Introducing Casual Conversations v2: A more inclusive dataset to measure fairness

March 9, 2023

Abstract

Developing robust and fair AI systems requires datasets with comprehensive set of labels that can help ensure the validity and legitimacy of relevant measurements. Recent efforts, therefore, focus on collecting person-related datasets that have carefully selected labels, including sensitive characteristics, and consent forms in place to use those attributes for model testing and development. Responsible data collection involves several stages, including but not limited to determining use-case scenarios, selecting categories (annotations) such that the data are fit for the purpose of measuring algorithmic bias for subgroups and most importantly ensure that the selected categories/subcategories are robust to regional diversities and inclusive of as many subgroups as possible.
CONCLUDING THOUGHTS

- Findings from our research show that the Monk Skin Tone Scale is as easy to use as the Fitzpatrick Scale, while being significantly more representative and inclusive; and that there are high levels of consensus, globally, using expert and crowdsourced (non-expert) annotators.

- ‘Subjective’ and ‘objective’ measures of skin tone are both important to collect.

- Different measures may help tap into different mechanisms through which skin tone may produce the inaccuracies researchers find with pulse oximeters (e.g., social determinants of health). There is still much to learn about the role of skin tone in pulse oximetry.
THANK YOU

Ellis Monk
Professor
Department of Sociology