Which Skin Tone Measures Are the Most Inclusive? An Investigation of Skin Tone Measures for Artificial Intelligence

COURTNEY M. HELDRETH, Google Research, USA
ELLIS P. MONK, Department of Sociology, Harvard University, USA
ALAN T. CLARK, The Value Engineers, USA
CANDICE SCHUMANN, XANGO EYEE, and SUSANNA RICCO, Google Research, USA

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.

CCS Concepts: • **Human-centered computing → User studies;** • **Social and professional topics → Cultural characteristics**;

Additional Key Words and Phrases: Responsible innovation, skin tone measurement, artificial intelligence, machine learning fairness, visual perception

ACM Reference format:
Courtney M. Heldreth, Ellis P. Monk, Alan T. Clark, Susanna Ricco, Candice Schumann, and Xango Eyee. 2024. Which Skin Tone Measures Are the Most Inclusive? An Investigation of Skin Tone Measures for Artificial Intelligence. *ACM J. Responsib. Comput.* 1, 1, Article 7 (March 2024), 21 pages.
https://doi.org/10.1145/3632120

1 INTRODUCTION

As **artificial intelligence (AI)** becomes more commonly applied to support decisions that affect the lives of many, understanding how to accurately capture demographic representation in

Courtney M. Heldreth and Ellis P. Monk contributed equally to this research.

Authors' addresses: C. M. Heldreth, S. Ricco, C. Schumann, and X. Eyee, Google Research, 1600 Amphitheatre Parkway, Mountain View, CA, USA, 94043; e-mails: {cheldreth, ricco, cschumann}@google.com, techwithx@gmail.com; E. P. Monk, Harvard University, 33 Kirkland Street, Cambridge Massachusetts, USA, 02138; e-mail: emonk@fas.harvard.edu; A. T. Clark, The Value Engineers, 440A 9th Street, San Francisco, CA, USA, 94103; e-mail: atclark@gmail.com.

2832-0565/2024/03-ART7
https://doi.org/10.1145/3632120

datasets, model training, and evaluation is gaining increasing attention to prevent the potential harms that result from biased algorithms [1]. This has become a critical problem, as algorithmic bias can place certain groups of people at a systematic disadvantage [2]. Indeed, as Fazelpour and Danks argue, there are many senses of the term bias, but "at its most neutral, algorithmic bias is… systematic deviation in algorithm output, performance, or impact, relative to some norm or standard" [3]. And as they also explain, "key social and personal decisions that impact our lives are increasingly guided by predictive algorithms. Medical diagnoses incorporate predictive models built using large datasets; loan approvals are informed by algorithmic judgments of credit worthiness; decisions to send social workers to investigate potential child abuse are guided by algorithm-based risk scores; and the examples multiply everyday. *At the same time, there is increasing awareness of the harmful impacts caused by biases in these algorithms: face recognition algorithms perform worse for people with feminine features or darker skin (and worse still for those with both) preventing people from accessing resources* [emphasis added]. That is, applications of facial recognition may result in disparate impact on darker-skinned individuals.

Given this, it should not be surprising that one key demographic characteristic used to evaluate fairness of datasets and algorithms is skin tone [4]. The use of skin tone to develop and test applications, such as facial detection and recognition, surveillance, and medical diagnosis, has grown significantly over the past few years. Therefore, it is important to utilize skin tone measures *that are representative of the people who will ultimately be affected by their applications*. This is critical to building machine learning systems that are trusted in their eventual domains of deployment. However, it is becoming apparent that existing measures do not achieve this goal.

The most commonly used skin tone measure in healthcare and computer science is the Fitzpatrick scale. The scale, originally used to plan treatments for psoriasis [5], is the industry standard used in dermatology to assess skin cancer risk or investigate variation in skin response to other types of injuries [6]. Despite its broad adoption, a growing number of studies highlight the limitations of the scale, including its inability to capture variation in global skin tones and underrepresentation of darker skin tones [6–9]. In response, researchers have proposed changes ranging from modifications to the questions used to define the types [9, 10], to explicit calls to replace the scale with a (yet to be determined) more equitable and inclusive alternative [8].

Recent work in computer vision adopts variants of the Fitzpatrick scale to measure skin tone for the purpose of fairness evaluation. Examples include datasets annotated with the six Fitzpatrick types [12, 13] or with binary darker / lighter classifications that group Fitzpatrick types I–II and IV–VI [14, 15]. Again, researchers have raised concerns about the lack of standardization in annotation practices and the need for more research into the validity of Fitzpatrick-based annotations compared to potential alternatives, especially given the limitations for people of color noted in the medical community [16]. We join these voices in calling for further research toward identifying more inclusive alternatives.

As a first step toward identifying a more inclusive skin tone scale, we ran a study to understand perceptual differences across three skin tone measurements among a nationally representative sample in the United States. We exposed participants to one of the three scales, had them rate how well they feel each scale represents them and explore the impact of demographic factors on these ratings to bring nuance to the question, "Which skin tone scale is perceived to be most inclusive?" In what follows, we provide background on skin tone and colorism, discuss algorithmic fairness and the measurement of skin tone in AI, then proceed to describe our data and methods before walking through our results. Embracing an interdisciplinary approach, we then conclude with recommendations for future researchers who wish to include skin tone measurement in their surveys, interviews, experimental research, or artificial intelligence applications. In particular, we urge researchers and practitioners to consider using a skin tone scale that is optimal: one that meets the threshold at which people feel enough sense of inclusivity—the key factors being the granularity and color selection on the scale—without being cognitively overwhelmed while still being efficient in ML models.

2 BACKGROUND AND RELATED WORK

2.1 Why Does Skin Tone Matter?

A vast and continually growing literature across the social sciences documents the many ways that race/ethnicity[1] is associated with life chances and outcomes. From health to wealth, the criminal justice system, wages, and much more, social scientists have amassed an impressive set of knowledge about how race/ethnicity is associated with inequality. Nevertheless, though relatively marginalized compared to the considerable literature on ethnoracial inequality, social scientists have also documented how inequality works in ways that go beyond mere membership in broad, aggregate ethnoracial categories (e.g., white, Black, Latinx, Asian, etc.) [17]. The major insight of this research is that gradational differences in skin tone within and across ethnoracial categories are associated with life chances and outcomes; and the magnitude of these inequalities within categories is often similar to or even exceeds what exists *between* ethnoracial categories [18]. This includes important outcomes such as educational attainment, income, and contact with and treatment within the criminal justice system [17, 19, 20]. In short, skin tone undeniably plays an important role in driving many social and economic outcomes for people of color, around the world. Given that the construction and definition of racial categories varies across societal context [21], research has shown skin tone to be a more stable signal to measure outcomes globally than racial categories [22, 23].One of the main mechanisms that links skin tone to inequality are the well-researched cognitive biases and stereotypes linked to light and dark skin tone and Afrocentric appearance, more broadly. For example, studies find that individuals with darker skin tones are more likely to be assigned to negative stereotypes [24], are perceived less positively by outgroup members [20], and experience lower incomes, worse jobs, and even poorer mental and physical health outcomes [19]. Therefore, skin tone plays an imperative role in our understanding of social stratification and inequality. This is because skin tone helps researchers capture how individuals are treated by others in their daily lives. People ascribed to and self-identifying with the same ethnoracial category may differ in skin tone and live vastly different lives given deeply ingrained skin tone biases by outgroups and even other members of their own ethnoracial category.

The conflation of race/ethnicity with skin tone and the lack of attention to intraracial heterogeneity in skin tone has, arguably, led to the relative marginalization of research on colorism compared to research that compares inequality between ethnoracial groups. Moreover, inattention to skin tone heterogeneity within and across ethnoracial groups is also related to the long history of skin tone-related biases in technology. On the one hand, then, evidence suggests darker-skinned individuals are targets for bias and discrimination across myriad social and economic domains, while, on the other hand, darker-skinned individuals are victims of bias by omission and elision when it comes to designing technology that works equally well for everyone. In terms of technology, many researchers point to the history of Kodak film printing, “which required that skin tones be matched to an image of a white model with the result that darker skin tones were oversaturated, or under-lit, so the only images that looked right were images of light-skinned people... [Importantly], light skin bias [was] not primarily a technical issue [because] film emulsions could have

[1] We use the terms race/ethnicity and ethnoracial to denote the considerable overlap between the concepts race and ethnicity in practice in everyday life and historically, which has led to them being used interchangeably. Some groups referred to as “ethnic” today were referred to as “racial” in the past and vice versa. In many ways, then, the separation of the terms is mostly artificial (see Wacquant 1997).

been designed that were more sensitive to a wider range of skin tones, but were not" [4]. In short, another way that skin tone stratification manifests itself is through technology—from medical devices to machine learning algorithms to facial recognition systems that are systematically biased against people of darker-skin. Until these systems are evaluated for fairness and representativeness and then made to be fairer and more inclusive these issues will remain.

2.2 Algorithmic Fairness and Skin Tone in Artificial Intelligence

To build technology, specifically, machine learning systems that are fair and trusted when they are deployed, it is important to recognize and mitigate bias throughout the AI development lifecycle [25]. Indeed, fairness and bias mitigation in machine learning is becoming common practice. In computer vision, particularly, it is common to release training and evaluation datasets, which include fairness attributes such as gender/gender presentation, age, and skin tone [26], to help facilitate disaggregated fairness analysis of machine learning models. In addition, research in recent years has produced algorithmic advances in machine learning bias mitigation techniques [25].

While concerns over and the academic literature regarding algorithmic fairness has exploded in recent years, it is worth noting that this area is rife with fierce debate over definitions. In fact, some have argued that some definitions and criteria for fairness are incommensurable, which has led to heated and rather abstract debate [3]. Nevertheless, what *is* clear from these debates and the policy discussions that these debates have spawned is that a fundamental aspect of algorithmic fairness and justice is ensuring that those who use and are affected by AI are empowered to participate in its regulation, from the very first stages [43]. Furthermore, algorithmic fairness and justice requires that algorithmic decision making is as transparent and traceable as possible [3]. Given this, it is somewhat striking that there have been so few studies on how diverse populations view the representativeness and inclusivity of skin tone scales that are central to data production and processing in computer vision (and more). This study seeks to make a key contribution to this burgeoning body of research.

Despite the fact that race/ethnicity and skin tone are distinct characteristics, foundational work in machine learning fairness has historically used racial categories to inform group fairness analyses [28]. Recent work in computer vision also adopts this practice, especially to measure group fairness in the context of facial analysis applications [4, 29]. However, as mentioned above, racial categories are an *unstable* measure for fairness, because racial categories are ill-defined, unstable temporally and geographically, and are culture specific [16, 30]. Given these issues, research suggests that phenotypes, which are observable characteristics, could help identify performance differences when conducting fairness analysis [14]. Multiple studies across applications have adopted this practice, demonstrating observed disparities in the datasets used to develop models [27], and actual model performance for people with darker skin tones [14, 15]. In addition, there is evidence that surfacing these disparities can spur algorithmic improvements [31].

When viewed in the context of the growing recognition from the social sciences of the importance of skin tone as one of the key phenotypic markers correlating with an individual's lived experience, the argument for considering skin tone for algorithmic fairness becomes even stronger. Common definitions of group fairness operationalize the intuitive goal that a deployed AI system should work well for all people who are ultimately affected by the system [28, 32]. However, these metrics rely on accurate and inclusive fairness attributes. Indeed, research has shown that noisy measurements may underestimate biases in machine learning models thus leading to false certainty of model performance [16, 33]. As such, it is imperative not just that fairness analyses of the development pipeline include skin tone as a dimension, but that the tools we use to define skin tone groups are inclusive, ensuring that everyone sees themselves represented and deserving of consideration.

There has been limited research showing that the general population struggles to understand basic fairness metrics—and when a person does not understand a metric fully, they tend to think the metric is valid [34]. Outside of machine learning, studies have shown that demographic characteristics of raters such as gender, race, and their amount of contact with diverse groups can influence their evaluations of the darkness or lightness of *others'* skin, regardless of which skin tone measure is used [7]. However, little is known about how these demographic factors may influence perceptions of the validity of the measures themselves. Therefore, in this study, we investigate the association between skin tone measure perceived representativeness and respondents' self-rated skin tone, race, and other demographic and personal characteristics. Understanding these relationships will not only help us understand whether existing measures used for skin tone annotation are perceived as representative of people more broadly, but it will also help us understand the ways our identities play a role in whether or not one feels represented by the skin tone measures used during annotation and social science interviews.

2.3 Skin Tone Measurements in AI

There are lots of ways to describe skin tone that have been used in social sciences, dermatology, and computer vision. In this section, we focus on the most common definitions used in computer vision fairness research. One common objective way to measure skin tone in computer vision is the **individual typology angle (ITA)**, which is a metric based on L* (lightness) and B* (yellow/blue) components of the CIE L* a* b* color space. The ITA score of an albedo map is considered to be the average of all pixel-wise ITA values within a skin region area. Skin color can then be classified using the ITA according to six categories, ranging from very light (category I) to dark (category VI) [35]. Given that the ITA can be easily computed from images, is seen as an objective metric, and significantly correlated with skin pigmentation, researchers have leveraged the ITA to determine the skin color measurement of a subject from camera images [36]. However, there are several limitations for using the ITA to estimate skin color. ITA values computed from images captured in uncontrolled environments can vary due to influence of scene lighting or other imaging artifacts, making it a less reliable scale to use in annotation for in-the-wild datasets.

The **Fitzpatrick Scale (FST)** is considered the "gold standard" of subjective skin tone measurement in dermatology [6, 13]. The FST was originally designed to assess UV-sensitivity of individuals "with white skin" for determining doses of ultraviolet A phototherapy. The instrument was released in 1975 and included four skin types (I–IV). In 1988, it was updated with an additional two skin types to account for individuals with darker skin (V–VI). The FST is a text-based assessment that determines an individual's skin type based on their responses to interview questions concerning the degree to which they burn or tan following typical sun exposure [5]. We refer to these values as self-reported FST.

In addition to dermatology, the FST has been used as a proxy for darker/lighter-skin to evaluate the fairness of algorithms, including face recognition and detection, and pedestrian detection for self-driving cars [12–15, 37]. In addition, crowdsourcing FST annotations is very common in computer vision research [26, 38]. Annotations are used to perform disaggregated fairness analysis on computer vision models [14], and for bias mitigation strategies when training machine learning models [39, 40]

A common criticism of the Fitzpatrick scale is that its types skew toward lighter complexions and have poor dynamic range for darker complexions. For example, in a survey with over 2,000 Black adults, 59% were unable to identify their skin tone using the Fitzpatrick scale [9]. Another study found that 114 of 270 participants from an ethnically diverse population gave responses to interview questions that did not directly map to Fitzpatrick types [11]. Depending on the exact protocol used, researchers warn that reliance on the Fitzpatrick scale risks overestimating the prevalence of type IV skin [9] or type VI skin [37] among Black individuals. Finally, a study that compared self-reported FST values with skin color as measured by a reflectance spectrophotometer found that just 5% of the variance in FST values for Black / Black Hispanic individuals was explained by differences in the L* skin color values [6]. Sommers et al. summarize the line of work by noting that "[FST] provides a restricted range of options for people with darker skin tones that do not capture variations in their skin color." This suggests that FST values do not represent the full spectrum of skin tones and in fact, may have a weak correlation to one's actual skin tone. Therefore, the FST may be limited in its ability to faithfully represent people, making it more difficult for systems built upon human annotation to be fairly evaluated across skin types. Due to these limitations, researchers have requested alternative, finer grain skin tone measurements not only in dermatology but also when measuring fairness in computer vision [8, 41]. The inclusion of finer grained scales that reflect a broad diversity of skin tones allows for researchers to better support design choices for disaggregate evaluations [42] and plays a critical role in addressing performance disparities, especially those that often impact communities that have faced injustices and have been historically underrepresented [14].

Given the evidence that alternative measures should be considered, the primary goal of this survey was to understand which skin tone measures are perceived as the most representative of potential end-users' own skin tones. We test this question through an important dimension of inclusiveness: How well does this skin tone measure represent the participant's perception of their own skin tone? We ask this question because when evaluating the fairness of an algorithm, an efficient and scalable skin tone scale allows ML researchers to support design choices for disaggregate evaluations, which plays a critical role in addressing performance disparities. If the skin tone measure used during the annotation task does not represent the skin tones of people depicted in the images, then the model may fail to "see" those skin tones, especially those from communities that have been historically underrepresented [43]. With the increase of focus on building inclusive datasets [44], and more specifically, because these datasets include images of people from a wide range of skin tones, people should feel that their skin tone (and by proxy, the measure they use to evaluate skin tone) is represented in the test set and by proxy, see themselves represented in the training set. In this way, technologists can ensure that performance is fair across the spectrum of skin tones.

The U.S. has grown increasingly racially and ethnically diverse in recent decades [45], and because of this diversity and the persistent, well-documented social and economic inequalities that can be observed in the United States based on gender and ethnicity [45, 46], we not only examine skin tone inclusiveness of skin tone measures across the entire population but also compare results between male and female, White and non-White, and lighter and darker skin tones. There is also research that suggests that our backgrounds may influence how we perceive our own skin tones and the skin tones of others [7], which helps us better understand how our identities play a role in whether or not one feels represented by skin tone measures used during annotation tasks. Therefore, our primary research questions are:

RQ1: How well do existing skin tone measures fare with respect to people's perceived representativeness of their skin tones?

RQ2: What is the association between a skin tone measure's perceived representativeness and respondents' demographic and personal characteristics?

To address these research questions, we conducted a survey to explore which skin tone measures are perceived to be representative of potential end-users' skin tones. We address concerns raised by prior literature that suggests that existing skin tone measures are insufficient in representing the broad spectrum of skin tone diversity by exploring whether novel skin tone measures are perceived to be more representative. Through this work, we argue that a consideration of skin tone measures beyond the FST that better align with human perceptions could greatly benefit fairness evaluations in machine learning.

3 SURVEY DESIGN

To measure the perceived representativeness of different skin tone measures, we launched an original data collection effort April–May 2021 in the United States through Qualtrics, a professional survey platform. Participants completed an online survey that assessed how represented they felt by three skin tone measures. First, participants responded to a series of demographic questions to better understand how their social environments influence their perceptions of skin tone scale inclusion. These factors included each participant's self-identified race, where participants were invited to choose all that apply from the following categories: white or Caucasian, Black or African American, LatinX, American Indian or Alaska Native, Asian or Pacific Islander, or other race (with a box to type in a specific answer). While these categories are, of course, imperfect, they do correspond both to how individuals classify themselves and others in everyday life and how the United States Census categorizes its population into ethnoracial groups for the purposes of tracking demographics and monitoring civil rights issues. Participants were randomly sampled on entry and then screened after answering to align with quotas based on population proportions from the US Census Bureau. Participants had to affirmatively answer that they agree to answer questions collecting potentially sensitive information, and to provide their best answers, and to be age 18 or older. The random sample was not nationally representative on entry, but data was weighted to be nationally representative on ethnoracial group and gender. We also asked raters to think about the skin tones of their closest friends as this could serve as proxies for exposure to individuals with diverse skin tones. The rationale being that individuals exposed to more diverse skin tones in their local environments may see their own skin tones differently than individuals with homogeneous friendship groups. How we understand our own skin tone, as exemplified by self-ratings, is fundamentally relational and profoundly affected by the skin tones we tend to encounter in our local environments [19, 20]. In short, "seeing" diversity in their social networks may influence their perceptions of their own skin tones and, ultimately, how inclusive and representative they found the various scales to be. Similarly, we included a measure of experience with cosmetics to control for the potential influence individuals' experiences with thinking about skin tone (i.e., domain expertise), such as engaging with makeup-related online content, may have on their feelings about being represented by a skin tone scale. Those individuals with more expertise in this domain may see scales differently than those who lack such expertise. In particular, those with more expertise may be systematically more meticulous and discerning about their perceptions of scale inclusivity. Finally, we asked participants their gender, highest level of education, and their income (for the full procedure, see Appendix A.1).

Next, respondents were randomly assigned to view one of the three skin tone scales (see Appendix A.2). These scales exemplify different contexts of creation and use-cases: one scale is dominant in the machine-learning community, one scale is from social scientific research on skin tone and inequality, and another scale is adapted from the cosmetic industry. We tested (1) a skin tone scale that is most commonly used in machine learning: the 6-point graphic-based Fitzpatrick scale [5], (2) a newly developed 10-point graphic-based skin tone scale[2] designed to capture ethnoracial

[2] The social scientific study of skin tone and colorism has mostly relied upon word-based scales or palettes that are used by field interviewers or survey respondents themselves to judge skin tone. Arguably, the most commonly used scale is the Massey-Martin scale (10 points), which was developed for the New Immigrant Study nearly two decades ago. This scale, however, has been criticized for its relative lack of color differentiation at the darker end of the skin tone scale and

diversity [41], and (3) a 40-option skin tone palette based on Rihanna's Fenty Beauty[3] makeup line. The palette was designed to capture a range of color undertones and a broad spectrum of color, created for people with darker skin tones in mind. By contrast, it is worth noting that the FST was *not* explicitly designed for this purpose. Instead, it was designed to capture how skin tone changes among Whites, specifically, with respect to phototherapy [8]. We included Rihanna's Fenty palette to better capture real skin colors and to explore how participants would respond to a broader range of skin tones. We then asked participants to describe their own relative skin tone based on the scale they were assigned to. This question asked them, "Which of the following is closest to your skin tone?" Based on this question, participants were categorized into light (Fitzpatrick I and II, Monk points 1–3, Fenty 100–190), medium (Fitzpatrick III and IV, Monk points 4–6, Fenty 200–390) and dark (Fitzpatrick V and VI, Monk points 7–10, Fenty 400–490) skin tone groups. To improve the reliability of this question, we asked participants to hold the back of their hand up to the screen[4] until they found the color on the scale that most closely matched their skin tone.[5]

To assess whether or not participants felt represented by each skin tone measure, participants were asked the extent to which they agreed their own skin tone was represented by the scale. For the sake of clarity, individuals' perceptions of the extent to which they feel their skin tones are represented may be distinct from the "objective" representativeness of a skin tone scale with respect to human skin tones. As is the case with any scale, it must necessarily reduce (though, in the best cases, optimally) the full range of human skin tones into a cognitively amenable range of options. This is why color selection and gradations are critically important in constructing any skin tone scale. In this study, we are concerned with individuals' *perceptions* of the extent to which they felt these scales were inclusive or representative of their understanding of their own skin tones. This is an important matter to assess for the sake of algorithmic justice, among other matters (see above).

The final analytic sample contained 2,214 respondents. As shown in Table 1, the sample generally conformed to the population characteristics of the target population. To ensure adequate statistical power and representation of darker-skinned participants who represent a smaller part of the United States population, we oversampled respondents who self-reported having medium and dark skin tones.

troubles with reliability (Hannon and DeFina 2016). In the case of Latin America, some studies have used the PERLA scale (11 points), for which it was specifically-designed (see Telles 2014). This study uses the Monk Scale (10 points), which was designed, through its color selection and gradations, to optimally capture ethnoracial diversity in skin tone across the Americas (e.g., North, South). This includes capturing skin tone variation within and across racial/ethnic categories in the United States (and potentially beyond).

[3] The Fenty Beauty palette is an inclusive range of 50 foundation shades created for Rihanna's foundation products. It was created after seeing a void in the industry for products that performed across all skin types and tones. Given the goal of this makeup line was designed "so that people everywhere would be included," we felt it was an important palette to test in this research. At the time the research was conducted, the Fenty Beauty palette contained 40 foundation shades. These shades were posted on Rihanna's Fenty Beauty website, and we worked with a designer to re-create the 40-skin tone shade palette that was used in the research.

[4] While not all devices render colors in exactly the same way, we included use of desktop vs. mobile devices to answer the survey as a control in statistical modeling and found no significant effects. Given our findings, we do not believe that slight color variations across displays would bias the data in any meaningful way.

[5] For some time there has been debate around the potential use of spectrophotometers to measure skin tone "objectively" in social scientific research (and other research) on skin tone. It is worth noting, however, that there are important questions of commensurability between spectrophotometer reflectance scores and the human perception of skin tone, even if they correlate to some extent. When it comes to the *social* significance of skin tone (e.g., its consequences for inequality, the representativeness of skin tone scales, etc.) it makes sense to put a premium on capturing how human beings perceive skin tone. Tellingly, one study finds that while spectrophotometer reflectance scores were not associated with poorer health, socially perceived skin tone, as a marker of exposure to discrimination and disadvantage, was associated with poorer health (Gravlee et al. 2005).

Table 1. Demographic Characteristics of Sample and Weighting To Be Nationally Representative of Race and Gender

Variable	*Levels*	*Sample*			
Variable	*Levels*	*N*	*Unweighted % of Sample*	*Weighted % of Sample*	*2019 / 2020 ACS 5 Year Estimates*
Used for weighting					
Sex	Male	1,039	47.3%	48.6%	49.3%
	Non-male	1,175	52.7%	51.4%	50.7%
Race/Ethnicity	White	804	36.3%	61.8%	61.5%
	Black	725	31.1%	13.4%	12.7%
	Latino/Latina	343	16.1%	11.3%	11.7%
	Asian	109	5.0%	5.3%	5,4%
	Native American	51	2.8%	0.7%	0.8%
	Other (incl. Hawaiian or Pacific Islander)	39	1.9%	4.6%	4.8%
	Mixed race	134	6.8%	3.0%	3.1%
Not used for weighting					
Age	18 to 29	869	38.7%	30.9%	
	30 to 39	408	19.0%	16.8%	
	40 to 49	296	13.0%	14.7%	
	50 to 64	324	14.7%	17.1%	
	65+	317	14.6%	20.5%	
Education	Less than high school	64	3.1%	2.4%	
	High school or GED	532	24.7%	21.8%	
	Some college	555	24.1%	23.8%	
	Bachelors or Associates	738	33.3%	34.1%	
	Graduate or professional degree	325	14.8%	18.0%	
Self-reported skin tone	Light	994	44.9%	61.8%	*no data available*
	Medium	582	27.1%	23.2%	*no data available*
	Dark	638	28.0%	15.0%	*no data available*

3.1 Data and Analytic Strategy

To answer the question of which skin tone scales that participants perceived best represented their skin tones, we tested for statistically significant differences between groups of respondents using pairwise comparisons (e.g., between respondents in the Fenty scale condition and respondents in the Fitzpatrick scale condition). We used Tukey's HSD test [48] for multiple comparisons of group means, with $\alpha = 0.05$. To account for differences between the sample of survey respondents and the United States population, group means were weighted to be nationally representative for race and gender.[6] We chose to run separate statistical analyses of self-reported skin tone and ethnoracial group. These are, of course, not fully independent of one another (e.g., respondents who self-reported light skin tones were 82% white after weighting). However, given small sample sizes of

[6] Using a raking technique to weight representative population proportions for gender and race (including mixed race) in the U.S., per the 2019 American Communities Survey's 5-year demographics estimates.

Table 2. Ordered Logistic Regression Model Results

Covariates	*Coefficient (log odds)*	*Std. Error*	*t value*	*P value*	*Coefficient (odds ratio)*
Not Male (vs. Male)	−0.17	0.10	−1.74	0.08	0.84
White (vs. Non-white)	0.10	0.12	0.84	0.40	1.10
Dark skin tone (vs. light skin tone)	0.02	0.13	0.14	0.89	1.02
Medium skin tone (vs. light skin tone)	−0.12	0.12	−1.01	0.31	0.89
Fenty scale (vs. Fitzpatrick scale)	0.30	0.11	2.89	**<0.01**	1.36
Monk scale (vs. Fitzpatrick scale)	0.22	0.10	2.10	**0.04**	1.24
Uses foundation (vs. does not)	0.13	0.11	1.122	0.26	1.14
Supports Democratic Party (vs. does not)	0.30	0.09	3.40	<**0.01**	1.35
Controls	*Coefficient (log odds)*	*Std. Error*	*t value*	*P value*	*Coefficient (odds ratio)*
Homogenous friends (vs. heterogeneous friends)	0.59	0.09	6.53	**<0.01**	1.81
Watches makeup content online (vs. does not)	0.40	0.11	3.59	**0.01**	1.50
Lives in urban cluster (vs. rural)	−0.53	0.35	−1.54	0.12	0.59
Lives in urbanized area (vs. rural)	−0.72	0.35	−2.05	**0.04**	0.49
Lives in minority white county (vs. does not)	0.02	0.09	0.19	0.85	1.02
Has college education (vs. does not)	−0.04	0.09	−0.47	0.64	0.96
Household income $75k+ annually (vs. <$75k annually)	0.22	0.10	2.19	**0.03**	1.24
Took survey on mobile (vs. took on desktop)	0.04	0.11	0.36	0.72	1.04
Age	−0.00	0.00	−1.08	0.28	1.00

Note that a number of variations on this model including interaction terms, alternate specificationsfˆrecordings of covariates such as race and income as well as adding interaction terms, did not produce meaningful differences in significance of coefficients or goodness-of-fit of the model. Therefore, we opted for a parsimonious model without interaction terms. Representativeness (5-point Likert scale, recoded from -2 (strongly disagree) to +2 (strongly agree)) was regressed on the following covariates. Covariates that were recoded binary outcomes have the reference level indicated in parentheses, e.g., Y (vs. reference level X).

certain subgroups (e.g., $n = 47$ respondents who reported being Black and having lighter skin tone), we thought it important to focus on consistency of effects across skin tone and ethnoracial groups, rather than directly comparing combinations of skin tone and ethnoracial group.

We also ran regression analyses to understand the relationship between participants' identity and personal experience and their perception of skin tone measures as representative. Specifically, we regressed perceived representativeness of skin tone scale (with -2 as "strongly disagree" and 2 as "strongly agree") on gender, whiteness, self-reported skin tone, skin tone scale, with controls for political affiliation, urbanicity (defined as rural, urban cluster, or urbanized per the U.S. Census Bureau)[7] , use of cosmetics, education level, and household income, survey mode, age, homogeneity of skin tones among close friend group, and consumption of makeup-related content. We found the same set of covariates to be significant when modeling strong agreement with a binary logistic regression. The full list of model covariates and control variables are available in Table 2.

[7] https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

4 RESULTS

In the main text, we report key findings of interest. Full statistics of all variables and group comparisons, as well as regression models, are included in Tables 1 and 2, respectively. We have also included group mean comparisons in Appendix A.3 for easier consumption.

Overall, participants generally agreed that at least one skin tone scale they saw represented their own skin tone (specified with "strongly disagree" as -2; and "strongly agree" as 2; median = 2; mean 1.39, SD 0.85, SE 0.02; see Figure 1). When comparing scale means via the Tukey test ($F_{2,2211} = 5.97$, $P < 0.01$),the Fitzpatrick scale (mean 1.32, SE 0.03) was perceived as significantly less representative of their own skin tone than both the Monk scale (mean 1.41, SE 0.03); Fitzpatrick vs. Monk, $P = 0.01$) and the Fenty scale (mean 0.44, SE 0.03); Fitzpatrick vs. Fenty, P <0.01). There was no statistically significant difference between the Fenty and Monk scales on this measure ($P = 0.87$).

We also see significant differences in perceived representativeness of these skin tone scales when comparing across and within demographic groups.

We found differences in perceived representativeness of skin tone scales across ethnoracial groups (Figure 2). When comparing means via the Tukey test for Black participants ($F_{2,722} = 4.71$, $P = 0.01$), the Fitzpatrick scale (mean 1.28, SE 0.09) is perceived as less representative than the Monk scale (mean 1.49, SE 0.09); Fitzpatrick vs Monk, $P < 0.01$) and the Fenty scale (mean 1.48, SE 0.08), Fitzpatrick vs. Fenty, $P = 0.03$). When comparing means via the Tukey test for white participants ($F_{2,801} = 0.20$, $P = 0.81$), there were no significant differences between the perceived representativeness of the scales (Fitzpatrick vs. Fenty, $P = 0.55$; Fitzpatrick vs. Monk, $P = 0.82$; Fenty vs. Monk, $P = 0.91$). Other ethnoracial groups did not exhibit statistically significant differences between perceived scale representativeness using the more conservative Tukey range test. However, when comparing scale means via the Tukey test for Asian participants ($F_{2,106} = 1.61$, $P = 0.20$), we found the Fitzpatrick scale (mean 1.05, SE 0.14) significantly less inclusive than the Fenty scale (mean 1.37, SE 0.12) using a one-tailed Student's T test (Fitzpatrick vs. Fenty, $P = 0.05$)—we think this result should be treated as marginally significant given the smaller sample of $n = 42$ Asian participants in the Fitzpatrick condition and $n = 35$ in the Fenty condition. In sum, then, there is evidence to suggest that the FST is viewed as less inclusive than the other scales we tested among non-Whites, specifically. Whites, however, seemed to view all scales as equally inclusive, which highlights how the use of the FST may be viewed as fair from the vantage point of dominant members of society, while being unfair to historically marginalized groups who may view it as non-representative.

We found significant differences in perceived representativeness of the skin tone scales based on participants' own skin tone (Figure 3). When comparing scale means via the Tukey test ($F_{2,635} = 6.29$, $P < 0.01$), participants with dark skin tones rated the Fitzpatrick scale (mean 1.12, SE 0.09) as significantly less representative of their own skin tone than the Monk scale (mean 1.47, SE 0.06; Fitzpatrick vs. Monk, $P < 0.01$) and the Fenty scale (mean 1.49, SE 0.01, Fitzpatrick vs. Fenty, $P = 0.01$). There were no significant differences between Fenty and Monk scales for participants with dark skin tones (Fenty vs. Monk, $P = 0.99$).

When comparing scale means via the Tukey test ($F_{2,579} = 1.65$, $P = 0.19$), there were also no significant differences among participants that self-reported having medium skin tones (Fitzpatrick, mean 1.21, SE 0.08; Fenty, mean 1.41, SE 0.05; Monk, mean 1.30, SE 0.07; Fitzpatrick vs Fenty, $P = 0.17$; Fitzpatrick vs Monk, $P = 0.65$; Fenty vs. Monk, $P = 0.58$). Similarly, when comparing scale means via the Tukey test ($F_{2,991} = 0.63$, $P = 0.53$), we saw no significant differences among respondents with light skin tones[8] (Fitzpatrick, mean 1.41, SE 0.04; Fenty, mean 1.44, SE 0.04; Monk, mean

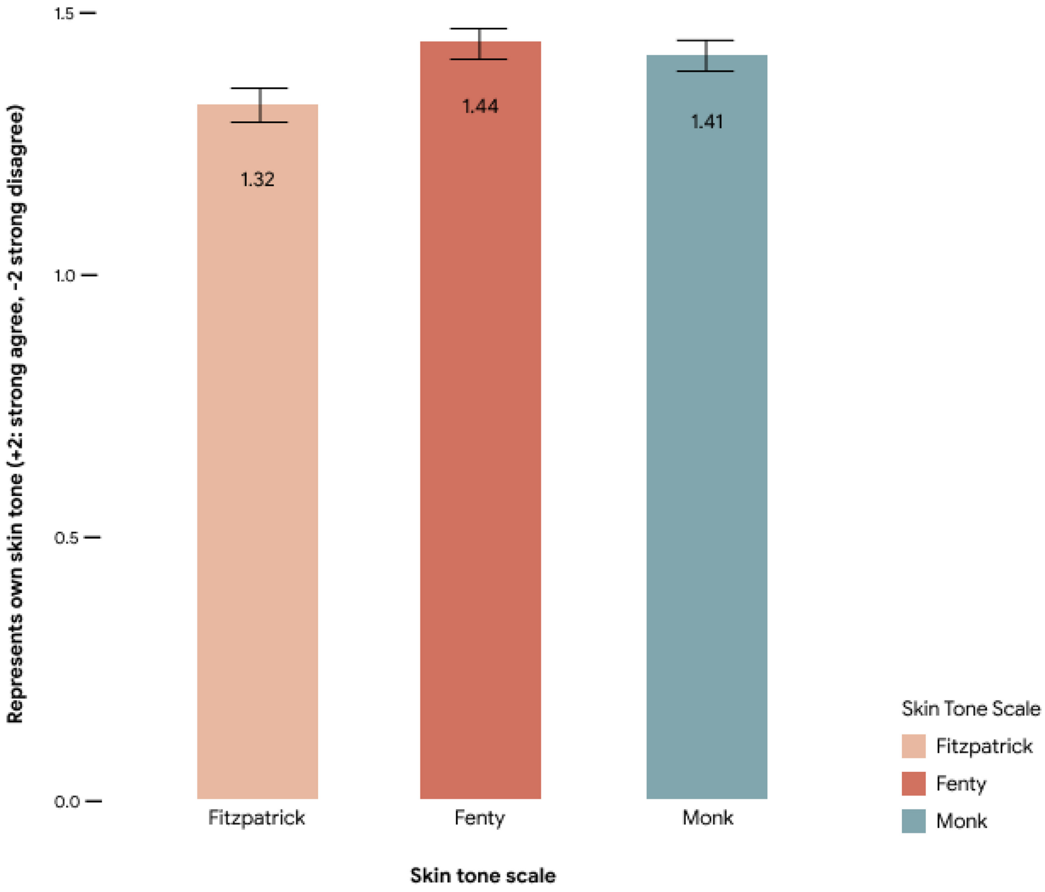


Fig. 1. Perceived representativeness of skin tone scales among overall sample.

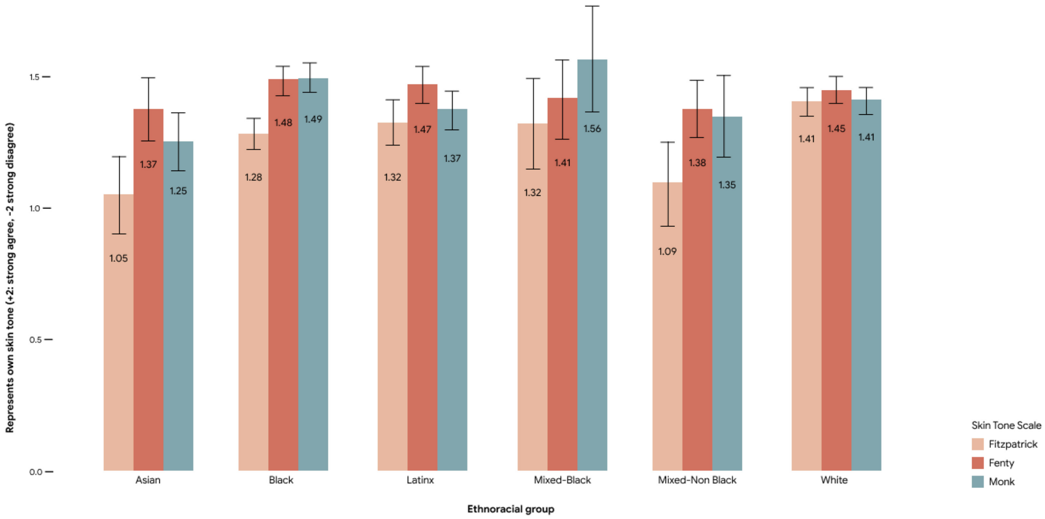


Fig. 2. Perceived representativeness of skin tone scales by ethnoracial group.

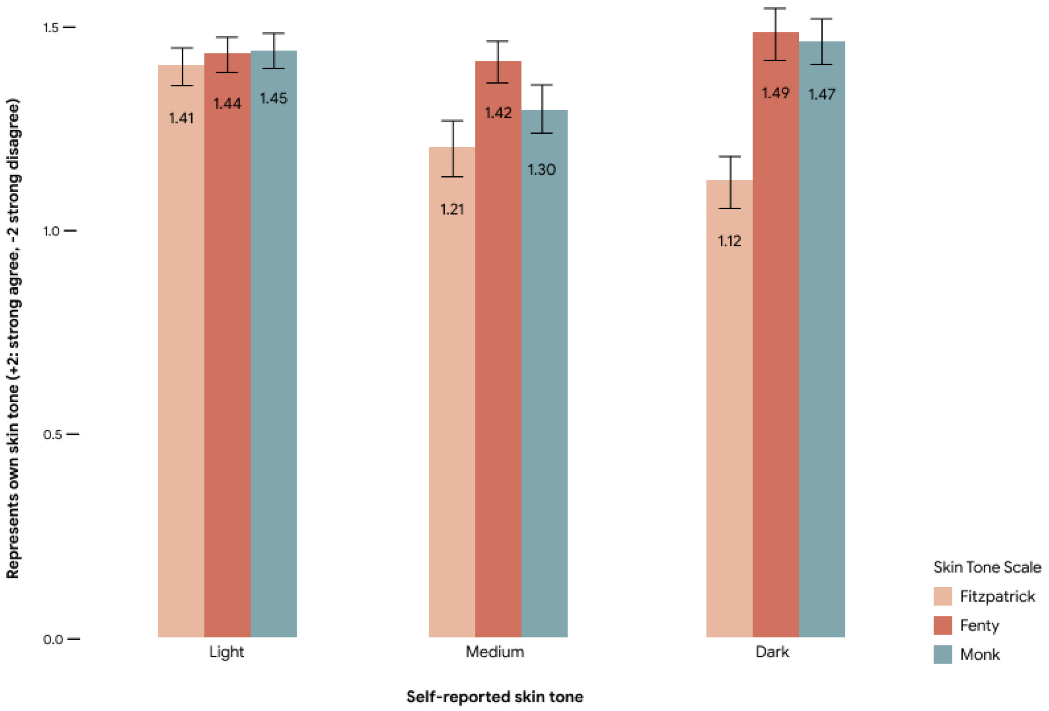


Fig. 3. Perceived representativeness of skin tone scales by self-reported skin tone.

1.45, SE 0.04; Fitzpatrick vs Fenty, $P = 0.57$; Fitzpatrick vs Monk, $P = 0.65$; Fenty vs. Monk, $P = 0.99$). Similar to the results above, the scales were viewed as roughly equivalent in terms of their fairness *except* for those with darker skin. In short, the FST may be viewed as fair by *some*, but not *all* members of U.S. society. It is also worth pointing out that given skin tone heterogeneity both within and across ethnoracial categories, this means that the FST may not be viewed as representative among darker-skinned Asian and Latinx individuals as well (i.e., not just Blacks).

When looking at the association between self-reported skin tone and representativeness, we found that there are gender differences in perceived representativeness of skin tone scales. We found that when comparing scale means for female and non-binary participants ($F_{2,1172} = 7.27$, $P < 0.01$), the Fitzpatrick scale (mean 1.28, SE 0.05) was perceived as less representative than the Monk (mean 1.46, SE 0.04; Fitzpatrick vs. Monk, $P = 0.02$) and Fenty scales (mean 1.44, SE 0.04; Fitzpatrick vs. Fenty, $P < 0.01$).

For male participants ($\mathrm{F}_{2,1036} = 0.67$, $P = 0.51$), there were no significant differences in perceived representativeness between the skin tone scales (Fitzpatrick, mean 1.36, SE 0.04; Fenty, mean 1.44, SE 0.04; Monk, mean 1.37, SE 0.04; Fitzpatrick vs Fenty, $P = 0.49$; Fitzpatrick vs Monk, $P = 0.92$; Fenty vs. Monk, $P = 0.74$; see Figure 4).

We note that intersectionality played a role in how people perceived the different skin tone measures. For example, we found that when comparing scale means for male participants with darker skin tones ($F_{2,294} = 3.23$, $P = 0.04$), we found that the Monk scale (mean 1.58, SE 0.09) was

[8] Skin tone, unsurprisingly, can be strongly connected with ethnoracial groups, but it is not a good proxy. 65.4% of light skin tone participants were white; 81.5% of dark participants were Black; medium skin tone participants were predominantly Black (28.9%), Latinx (27.0%), or white (20.6%)

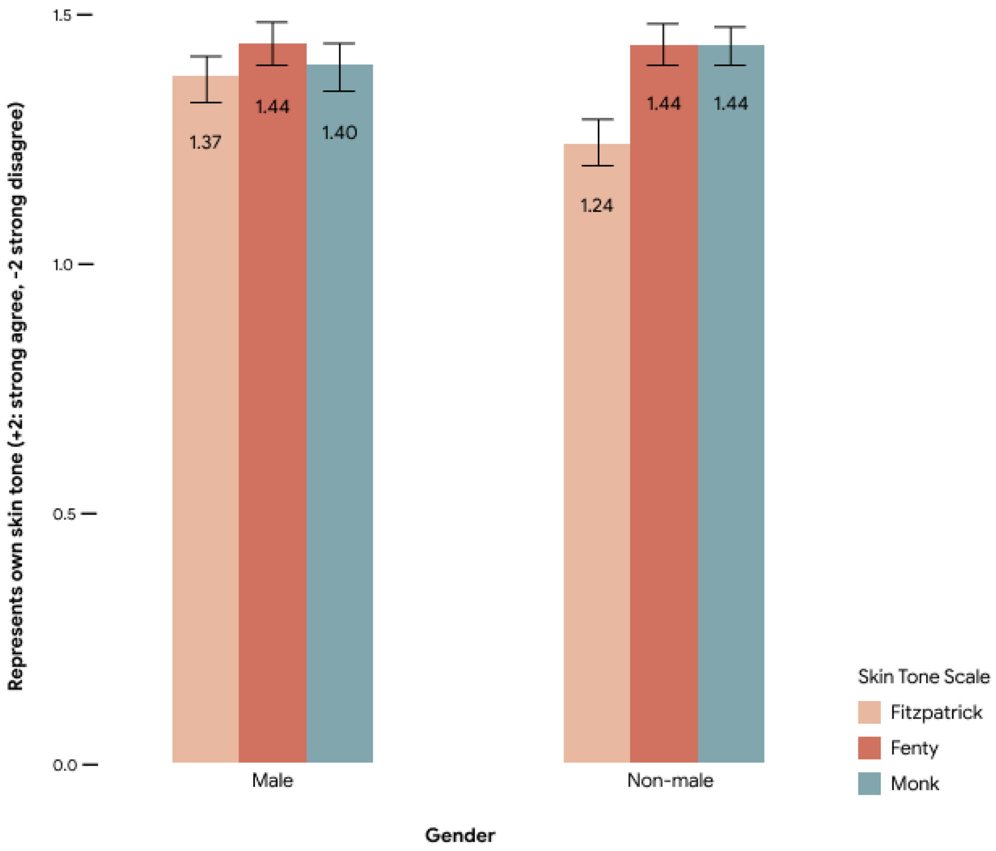


Fig. 4. Perceived representativeness of skin tone scales by gender.

perceived as significantly more representative than the Fitzpatrick scale (mean 1.20, SE 0.11; Fitzpatrick vs. Monk, $P = 0.03$), but Fenty was not (mean 1.47, SE 0.13; Fitzpatrick vs. Fenty, $P = 0.34$).

We also note the stark difference between answers from white male participants compared to all other participants. We found that when comparing scale means for white male participants ($F_{2,436} = 0.75$, $P = 0.47$), there was no statistically significant difference across the scales (Fitzpatrick, mean 1.45, SE 0.06; Fenty, mean 1.45, SE 0.05; Monk, mean 1.34, SE 0.06; Fitzpatrick vs. Monk, $P = 0.53$; Fitzpatrick vs. Fenty, $P = 0.99$, Fenty vs Monk, $P = 0.54$). When comparing scale means for everyone other than white male participants ($F_{2,1772} = 8.66$, $P < 0.01$), the Fenty (mean 1.43, SE 0.04, Fitzpatrick vs. Fenty, $P < 0.01$) and Monk (mean 1.45, SE 0.04; Fitzpatrick vs. Monk, $P < 0.01$) scales were perceived as significantly more representative than the Fitzpatrick scale (mean 1.26, SE 0.04).

To understand whether any of these demographic or personal factors were significantly associated with the rating of skin tone scale representativeness, we ran an ordered logistic regression. We regressed participants' score on the scale representativeness Likert scale question on a series of individual demographic variables (e.g., race, gender, skin tone), personal experience control variables (e.g., friend group heterogeneity, consumption of online makeup content), and geographic control variables[9] (e.g., urbanicity). Full specification details of this model and its results are shared in Table 2.

[9] Using the US Census' API, with participants' self-reported ZIP code mapped to county-level statistics.

We found that the following predictors had statistically significant ($\alpha = 0.05$, i.e., the 95% confidence interval for the coefficients do not cross 0) associations with participants ratings of a scale representing their own skin tone while holding all other variables constant: *which skin tone scale they viewed* and *household income.*

For participants who saw the Fenty Beauty palette, the odds of finding the scale representative of their own skin tone were 1.36 times that of participants who saw the Fitzpatrick scale, holding all other variables constant ($P < 0.01$). For participants who saw the Monk scale, the odds were 1.24 times the odds of participants who saw the Fitzpatrick scale ($P = 0.04$).

For participants with a household income of $75,000 or higher per annum, the odds of finding a scale representative of their own skin tone were 1.24 higher than those of participants from households with less than $75,000 per annum ($P = 0.03$).

We did not find significant main effects for gender ($P = 0.08$), self-reported dark skin tone ($P = 0.89$) or medium skin tone ($P = 0.31$), or ethnoracial group (recoded as White vs. non-White for model parsimony, $P = 0.40$) in this specific model. However, we note that the lived experience of people fitting certain demographic subgroups is likely better reflected in the group means used in the pairwise comparisons above, rather than in a predictive model like this one.

We ran goodness-of-fit tests recommended for ordered logistic regression [49] and failed to find evidence for lack-of-fit for the model (Hosmer-Lemeshow ordinal test, $\chi^2_{15} = 14.1$, $P = 0.52$; Lipsitz test, $\chi^2_9 = 2.8$, $P = 0.97$).

5 DISCUSSION

The encroachment of machine learning and artificial intelligence into nearly every realm of society has not gone unnoticed and each day there are more and more calls to ensure these systems and the algorithms underlying them are as unbiased and fair as possible. One aspect of algorithmic fairness and inclusion is transparency for and feelings of representativeness and fairness among those who are affected by these systems. Notably, many forms of computer vision technology rely upon skin tone scales, yet research examining the extent to which people feel measures of skin tone used in many algorithms are fair and inclusive is in its infancy. Presently, most research has focused on the reliability of various skin tone measures [7] and not perceptions of inclusivity and/or representativeness of skin tone measures. To our knowledge, this is the first study to date that has systematically explored a simple yet important question: *How represented do people feel by skin tone measures?* Our results provide new, survey-based evidence that the current industry standard, the Fitzpatrick scale, is perceived to be less inclusive than other skin tone measures that were intentionally designed to capture individuals with darker skin tones—the Monk scale and the Fenty Beauty palette. The differences in perceived representation between the scales were small, yet powerful and consistent given that participants were only assigned to evaluate one skin tone measure and thus not asked to compare across the skin tone measures.

Important to note were the differences in perceived representation across historically marginalized groups. Women, people with darker skin tones, and some ethnoracial minorities consistently felt less represented by the industry standard, the Fitzpatrick scale, compared to the Monk scale and the Fenty Beauty palette. This suggests that the Fitzpatrick scale is perceived as less representative of groups that do not fit the privileged majority of White males, further highlighting the need to shift away from the Fitzpatrick scale as it was not intended to capture a broader range of dark skin tone diversity [9].

Despite a few differences between the Fenty Beauty palette and the Monk scale, it is important to note that the Fenty Beauty palette contains four times the number of skin tone shades than the Monk scale. This is encouraging as it suggests that when designed to focus on darker skin tones, a simpler 10-point skin tone measure can be perceived to be as inclusive and robust as a 40-point skin tone measure. Our findings suggest that in the U.S. context, it is not simply the *number* of options that matters for perceptions of inclusivity, but the *quality* of options (i.e., the fit between options and the preferences of a given market). While it may plausibly be the case that more options will correlate with heightened perceptions of inclusivity, it is unclear how effectively a 40-point scale can be applied in a machine learning context. In terms of efficiency, a 40-point skin tone scale may be challenging to implement compared to utilizing a simpler 10-point scale that does not compromise in terms of representation. Asking raters to repeatedly evaluate images using a larger 40-point scale could introduce considerable cognitive load, potentially reducing interrater agreement. Gains in perceptions of inclusivity and representativeness may come at the cost of usability, reliability, and accuracy. Understanding how to effectively implement larger and more inclusive scales is an important area for future research and exploratory work in applied machine learning systems. Nevertheless, it is worth noting that our findings suggest that color selection—not just the expansion of a scale's number of swatches/options—is a promising avenue to address perceived representativeness and inclusivity.

While our findings paint a detailed picture of how people feel represented across three skin tone measures, it's important to note that our survey is U.S. focused. Therefore, it is unclear whether or not these findings will generalize outside of the United States. Still, it is worth noting that the Monk scale was designed specifically to cover skin tone distributions in the United States and Brazil and was informed by global research on the relationship between skin tone and UV radiation around the world, while the Fenty Beauty palette was also designed to be inclusive of all skin types. Given similarities in the range of human skin tones across the globe (see Reference [50]), there is strong reason to believe these results may hold globally. Nevertheless, further research is needed to validate this work in a global context, especially since cultural factors may influence how people both perceive their own skin tone and feel included (or excluded) by skin tone measures.

Importantly, these results demonstrate that participant's social, personal, and demographic characteristics influenced how included they felt by various skin tone measures—a novelty in the current literature on skin tone in machine learning and AI. In other words, how we think about our skin tone is shaped by our local environments [19]; and this extends to how we ultimately may feel about how representative measures of skin tone may be. This is reflected in our results, which demonstrate that while there are no differences among whites in terms of how representative they feel across the three skin tone measures, non-whites are more selective about the measures of skin tone they find more representative. This makes sense given the heightened salience of skin tone among non-white communities (e.g., familial socialization, perceiving discrimination, and, of course, displaying a wider range of different skin tones). Important to note is that as the demographics of the United States becomes increasingly non-white [51] and darker-skinned, not implementing representative measures of skin tones means that AI systems that do not use representative measures of skin tone may not work optimally for more than half of the United States population. In addition, globally speaking, non-whites make up a majority of the world population. This strongly suggests the need for more representative, fine-grained measures of skin tone for equitable products for people all over the world.

Ultimately, then, what we see as a North Star for inclusive and ethical AI/ML is adopting a measure of a social signal (skin tone) that meets the threshold (optimal cutoff point) of adequate representativeness while being practically useful for AI/ML annotation by human beings (i.e., not cognitively overwhelming). Future research, using experimental and/or interview-based approaches, should dig deeper by examining this potential trade-off by measuring perceptions of scales' ease of use and perceptions of the adequacy of the number of options (e.g., perceived as having too few or too many options), in addition to the more standard analyses of inter-rater reliability and consensus. This future research, then, using multiple methods, would further deepen the insight around a trade-off between inclusivity (ethics), cognitive load/practical use, and efficiency for adoption in ML. Solving bias in AI/ML will necessitate this ethos—interdisciplinary testing to ensure representation and inclusivity, which is rooted in recognizing the need for compromise to ensure representative and inclusive AI/ML remains practically useful and efficient enough to be implemented by practitioners across myriad use-cases (which themselves should be ethical!).

6 CONCLUSION

This study shows the role social science can play in informing machine learning best practices. Given the widespread use of skin tone measures in AI, the present work highlights the role colorism research plays in informing skin tone measurement and how people feel represented. The fact that the Fitzpatrick scale is only perceived to be inclusive among the non-marginalized, while a scale used in the social sciences is perceived to be inclusive by non-marginalized and ethnoracial minorities alike, confirms that an interdisciplinary approach can be helpful in identifying and mitigating harms that can be produced by AI systems that rely on skin tone. After all, using a scale that is only perceived as inclusive by those who are not ethnoracial minorities may potentially exacerbate algorithmic bias and be a signal of algorithmic injustice. As such, ensuring that people feel represented by skin tone measures used to develop AI systems is an inherently social and ethical question that requires rigorous, socio-technical approaches to build products that are inclusive. Given the ubiquity of skin tone in computer vision, machine learning, and AI, our results have wide-ranging implications for the deployment of algorithms in facial recognition, search, medical device testing, and much more.

A APPENDIX

A.1 Procedure

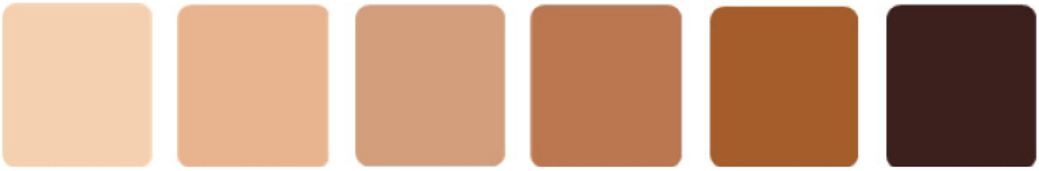
After consenting, participants were asked to provide their age, gender, and race. Participants were then randomly assigned to view one of three skin tone measures. To understand the composition of their friendship networks, we asked participants, “Which of the following statements best describes the skin tones of your closest friends?,” with options “Most of them have a similar skin tone to mine,” “Most of them have a lighter skin tone than mine,” “Most of them have skin tones darker than mine,” “My friends have a mix of skin tones.”

Next, participants were asked to identify the skin tone that best matched their own skin tone, “Which of the following is closest to your skin tone? For the most accurate results, hold the back of your hand up to the screen until you find the color that most closely matches.” If participants felt the skin tone measure did not contain a skin tone that matched them, then they could select “none of the above.” This option was infrequently selected across all three skin tone measures (0.38% Fitzpatrick, 0.09% Fenty, 0.28% Monk). After participants selected their skin tone, we asked how well the scale they were assigned to represented them, “Please look closely at this range of skin tones. To what extent do you agree or disagree with the following statement: My own skin tone is represented in this scale,” on a 5-point Likert scale from “Strongly disagree” to “Strongly agree.”

In separate questions, we asked participants how well the skin tone scales represented the skin tones of their community; the country; a broad range of skin tones; a diverse range of skin tones; lighter skin tones; medium skin tones; and darker skin tones. However, we focus here on participants’ feelings of how well the scales represent their own skin tones, as the best indicator of how representative the scales feel.

In open-ended text responses in the survey, we saw no mention of the Fenty Beauty scale, so we do not have any indication that the scale being perceived as more representative is due to recognition of the specific palette or association with Rihanna.

A.2 Skin Tone Measures

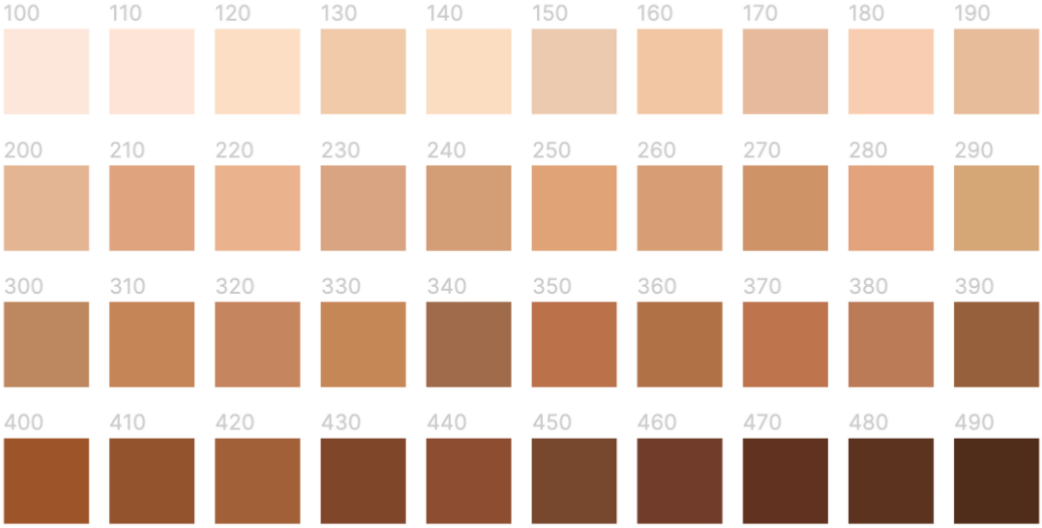


6-point Fitzpatrick scale

The Fitzpatrick scale (4) asked participants to select the option that was closest to their skin tone using this graphic as a guide:

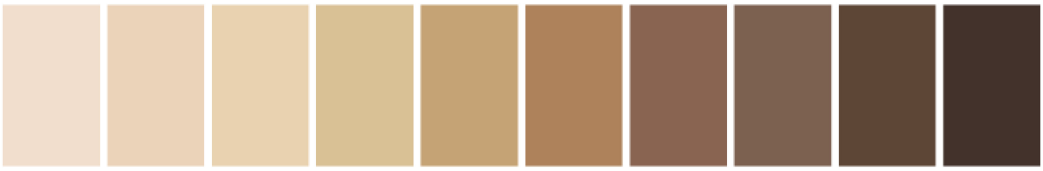
40-point Rihanna Fenty Beauty Palette

This scale asked participants to select the option that was closest to their skin tone based on a palette of skin tones. This measure was adapted from Rihanna's 40 Fenty Beauty Shades. The numbers correspond to the shade associated with her beauty line. To avoid confusion, these numbers were removed from the survey so participants just saw the skin tone options.

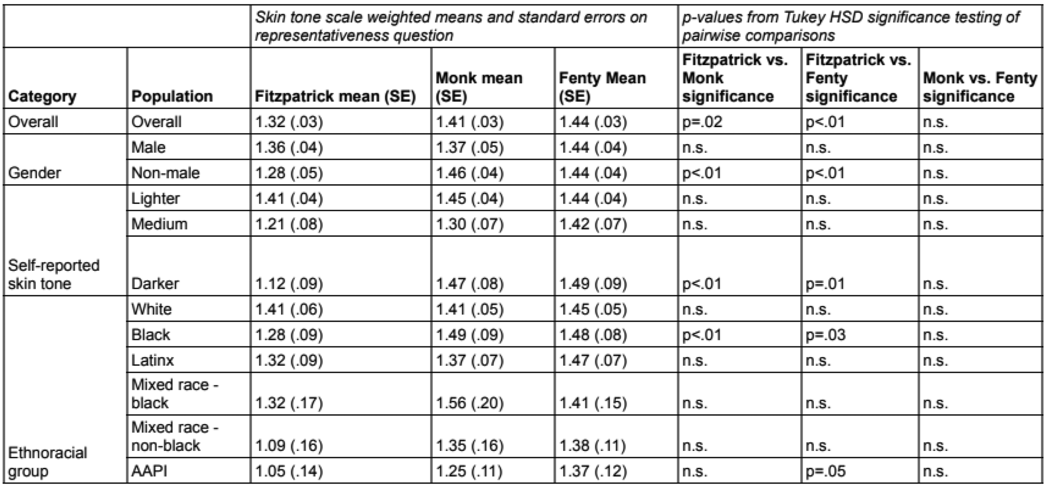


10-point Monk scale

This scale asked participants to select the option that was closest to their skin tone based on 10 options.



A.3 Group Means Comparisons



		Skin tone scale weighted means and standard errors on representativeness question			*p-values from Tukey HSD significance testing of pairwise comparisons*		
Category	**Population**	**Fitzpatrick mean (SE)**	**Monk mean (SE)**	**Fenty Mean (SE)**	**Fitzpatrick vs. Monk significance**	**Fitzpatrick vs. Fenty significance**	**Monk vs. Fenty significance**
Overall	Overall	1.32 (.03)	1.41 (.03)	1.44 (.03)	p=.02	p<.01	n.s.
Gender	Male	1.36 (.04)	1.37 (.05)	1.44 (.04)	n.s.	n.s.	n.s.
	Non-male	1.28 (.05)	1.46 (.04)	1.44 (.04)	p<.01	p<.01	n.s.
Self-reported skin tone	Lighter	1.41 (.04)	1.45 (.04)	1.44 (.04)	n.s.	n.s.	n.s.
	Medium	1.21 (.08)	1.30 (.07)	1.42 (.07)	n.s.	n.s.	n.s.
	Darker	1.12 (.09)	1.47 (.08)	1.49 (.09)	p<.01	p=.01	n.s.
Ethnoracial group	White	1.41 (.06)	1.41 (.05)	1.45 (.05)	n.s.	n.s.	n.s.
	Black	1.28 (.09)	1.49 (.09)	1.48 (.08)	p<.01	p=.03	n.s.
	Latinx	1.32 (.09)	1.37 (.07)	1.47 (.07)	n.s.	n.s.	n.s.
	Mixed race - black	1.32 (.17)	1.56 (.20)	1.41 (.15)	n.s.	n.s.	n.s.
	Mixed race - non-black	1.09 (.16)	1.35 (.16)	1.38 (.11)	n.s.	n.s.	n.s.
	AAPI	1.05 (.14)	1.25 (.11)	1.37 (.12)	n.s.	p=.05	n.s.

ACKNOWLEDGMENTS

We thank the Google Skin Tone team for their contributions to this work.

REFERENCES

[1] V. Rotemberg, A. Halpern, S. W. Dusza, and N. C. F. Codella. 2019. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. *Semin. Cutan. Med. Surg.* 38, 1 (2019), E38–E42.

[2] R. Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*, 1st ed. Polity, Cambridge, UK.

[3] S. Fazelpour and D. Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philos. Compass* 16, 8 (2021), e12760.

[4] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. In *Proceedings of the ACM on Human-Computer Interaction.* 1–35.

[5] T. B. Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Arch. Dermatol.* 124, 6 (1988), 869–871.

[6] M. S. Sommers, J. D. Fargo, Y. Regueira, K. M. Brown, B. L. Beacham, A. R. Perfetti, J. S. Everett, and D. J. Margolis. 2019. Are the Fitzpatrick skin phototypes valid for cancer risk assessment in a racially and ethnically diverse sample of women? *Ethnic. Dis.* 29, 3 (2019), 505–512. https://doi.org/10.18865/ed.29.3.505

[7] M. E. Campbell, V. M. Keith, V. Gonlin, and A. R. Carter-Sowell. 2020. Is a picture worth a thousand words? An experiment comparing observer-based skin tone measures. *Race Soc. Problems* 12, 3 (2020), 266–278. https://doi.org/10.1007/s12552-020-09294-0

[8] U. K. Okoji, S. C. Taylor, and J. B. Lipoff. 2021. Equity in skin typing: Why it is time to replace the Fitzpatrick scale. *Brit. J. Dermatol.* 185, 1 (2021), 198–199. https://doi.org/10.1111/bjd.19932

[9] L. C. Pichon, H. Landrine, I. Corral, Y. Hao, J. A. Mayer, and K. D. Hoerster. 2010. Measuring skin cancer risk in African Americans: Is the Fitzpatrick skin type classification scale culturally sensitive? *Ethnic. Dis.* 20, 2 (2010), 174–179.

[10] O. R. Ware, J. E. Dawson, M. M. Shinohara, and S. C. Taylor. 2020. Racial limitations of Fitzpatrick skin type. *Cutis* 105, 2 (2020), 77–80.

[11] S. Eilers, D. Q. Bach, R. Gaber, H. Blatt, Y. Guevara, K. Nitsche, and J. K. Robinson. 2013. Accuracy of self-report in assessing Fitzpatrick skin phototypes I through VI. *JAMA Dermatol.* 149, 11 (2013), 1289–1294.

[12] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer. 2020. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Trans. Technol. Soc.* 1, 1 (2020), 8–20.

[13] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer. 2021. Casual conversations: A dataset for measuring fairness in ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2289–2293.

[14] J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency.* 77–91. PMLR.

[15] B. Wilson, J. Hoffman, and J. Morgenstern. 2019. Predictive inequity in object detection. Retrieved from https://arXiv:1902.11097

[16] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. 2020. Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms. Retrieved from https://arXiv:2010.07979.

[17] E. P. Monk Jr. 2014. Skin tone stratification among black Americans, 2001–2003. *Soc. Forces* 92, 4 (2014), 1313–1337.

[18] E. P. Monk Jr. 2021. The unceasing significance of colorism: Skin tone stratification in the United States. *Daedalus* 150, 2 (2021), 76–90.

[19] E. P. Monk Jr. 2015. The cost of color: Skin color, discrimination, and health among African-Americans. *Amer. J. Sociol.* 121, 2 (2015), 396–444.

[20] E. P. Monk. 2019. The color of punishment: African Americans, skin tone, and the criminal justice system. *Ethnic Racial Studies* 42, 10 (2019), 1593–1612.

[21] S. Benthall and B. D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 289–298.

[22] S. R. Bailey, F. M. Fialho, and A. M. Penner. 2016. Interrogating race: Color, racial categories, and class across the Americas. *Amer. Behav. Sci.* 60, 4 (2016), 538–555.

[23] E. P. Monk Jr. 2016. The consequences of "race and color" in Brazil. *Soc. Problems*, 63, 3 (2016), 413–430.

[24] K. B. Maddox and S. A. Gray. 2002. Cognitive representations of black Americans: Reexploring the role of skin tone. *Personal. Soc. Psychol. Bull.* 28, 2 (2002), 250–259.

[25] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8919–8928.

[26] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, and Y. Saraf. 2022. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'22)*, 6162–6166. IEEE.

[27] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, and T. Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229.

[28] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* (2016), 29.

[29] P. Grother, P. Grother, M. Ngan, and K. Hanaoka. 2019. *Face Recognition Vendor Test (FRVT) Part 2: Identification.* U.S. Department of Commerce, National Institute of Standards and Technology.

[30] Z. Khan and Y. Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.* 587–597.

[31] I. D. Raji and J. Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 429–435.

[32] P. Garg, J. Villasenor, and V. Foggo. 2020. Fairness metrics: A comparative analysis. In *Proceedings of the IEEE International Conference on Big Data (Big Data'20).* IEEE, 3662–3666.

[33] J. Cho, A. Zala, and M. Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. Retrieved from https://arXiv:2202.04053

[34] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, and M. Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the International Conference on Machine Learning. PMLR,* 8377–8387.

[35] A. Chardon, I. Cretois, and C. Hourseau. 1991. Skin colour typology and suntanning pathways. *Int. J. Cosmetic Sci.* 13, 4 (1991), 191–208

[36] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, and O. Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1820–1828

[37] M. Wilkes, C. Y. Wright, J. L. du Plessis, and A. Reeder. 2015. Fitzpatrick skin type, individual typology angle, and melanin index in an African population: Steps toward universally applicable skin photosensitivity assessments. *JAMA Dermatol.* 151, 8 (2015), 902–903.

[38] D. Zhao, A. Wang, and O. Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 14830–14840.

[39] D. Madras, E. Creager, T. Pitassi, and R. Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning.* PMLR, 3384–3393.

[40] Y. Yang, A. Gupta, J. Feng, P. Singhal, V. Yadav, Y. Wu, and J. Joo. 2022. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 813–822

[41] Monk. 2019. Monk skin tone scale. Retrieved from https://skintone.google

[42] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, and H. Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 368–378.

[43] A. Chouldechova and A. Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[44] V. V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram, and O. Russakovsky. 2023. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. Retrieved from https://arXiv:2301.02560

[45] B. A. Lee, J. Iceland, and G. Sharp. 2012. *Racial and Ethnic Diversity Goes Local: Charting Change in American Communities over Three Decades.* Russell Sage Foundation, New York, NY.

[46] L. J. Zigerell. 2018. Black and white discrimination in the United States: Evidence from an archive of survey experiment studies. *Res. Politics* 5, 1 (2018), 2053168017753862.

[47] G. K. SteelFisher, M. G. Findling, S. N. Bleich, L. S. Casey, R. J. Blendon, J. M. Benson, and C. Miller. 2019. Gender discrimination in the United States: Experiences of women. *Health Services Res.* 54 (2019), 1442–1453.

[48] J. Jaccard, M. A. Becker, and G. Wood. 1984. Pairwise multiple comparison procedures: A review. *Psychol. Bull.* 96, 3 (1984), 589.

[49] M. Fagerland and D. Hosmer. 2017. How to test for goodness of fit in ordinal logistic regression models. *Stata J.* 17, 3 (2017), 668–686.

[50] Nina G. Jablonski. 2004. The evolution of human skin and skin color. *Annu. Rev. Anthropol* 33 (2004), 585–623.

[51] H. R. Outten, M. T. Schmitt, D. A. Miller, and A. L. Garcia. 2012. Feeling threatened about the future: Whites' emotional reactions to anticipated ethnic demographic changes. *Personal. Soc. Psychol. Bull.* 38, 1 (2012), 14–25.

Received 3 October 2022; revised 17 October 2023; accepted 18 October 2023