Measuring Skin Color:
Consistency, Comparability, and Meaningfulness of Rating Scale Scores and Handheld Device Readings

Rachel A. Gordon

University of Louisiana at Lafayette (Lafayette, LA, USA)

Amelia R. Branigan

University of Maryland (College Park, MD, USA)

Mariya Khan and Johanna G. Nunez

University of Illinois at Chicago (Chicago, IL, USA)

Rachel A. Gordon is Director of the Cecil J. Picard Center for Child Development and Lifelong Learning, University of Louisiana at Lafayette, Lafayette, LA, 70506 (email: rachel.gordon@louisiana.edu); Amelia R. Branigan is Assistant Professor, Department of Sociology and Population Research Center, University of Maryland, College Park, MD 20742 (e-mail: branigan@umd.edu); Mariya Khan is a Graduate Student and Project Coordinator, Department of Sociology, University of Illinois at Chicago, Chicago, IL 60607 (e-mail: mkhan252@uic.edu); and, Johanna G. Nunez is an Assistant Project Coordinator, Institute for Health Research and Policy, University of Illinois at Chicago (email: jnunez26@uic.edu). This material is based upon work supported by the National Science Foundation under Grant No. 1921526. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Dr. Gordon was Professor, Department of Sociology, and Associate Director,

Institute for Health Research and Policy, University of Illinois at Chicago, when this project was completed. The authors also thank the UIC Department of Sociology and Summer Research Opportunities Program for seed funding, six undergraduate students for participating in a supervised research course to help plan the study (Ruby Garcia, Nithya Karpagavinayagam, Sarah Main, Stephanie Molina, Johanna G. Nunez, and Kobie Price), three students for assisting in data collection (Sarah Main, Stephanie Molina, and Johanna G. Nunez), Anshuman Jyothi Das for consultation in the use of the Labby device, and Rosemary Seelaus for helpful comments. This study design and analysis was not preregistered. **Word Count (main text): 6,777.**

Abstract

As U.S. society continues to diversify and calls for better measurements of racialized appearance increase, survey researchers need guidance about effective strategies for assessing skin color in field research. This study examined the consistency, comparability, and meaningfulness of the two most widely used skin tone rating scales (Massey-Martin and PERLA) and two portable and inexpensive handheld devices for skin color measurement (Nix colorimeter and Labby spectrophotometer). We collected data in person using these four instruments from 46 college students selected to reflect a wide range of skin tones across four racial-ethnic groups (Asian, Black, Latinx, White). These college students, five study staff, and 459 adults from an online sample also rated 40 stock photos, again selected for skin tone diversity. Our results—based on data collected under controlled conditions—demonstrate high consistency across raters and readings. The Massey-Martin and PERLA scale scores were highly linearly related to each other, although PERLA better differentiated among people with the lightest skin tones. The Nix and Labby darkness-to lightness (L^*) readings were likewise linearly related to each other and to the Massey-Martin and PERLA scores, in addition to showing expected variation within- and between-race-ethnicities. Additionally, darker Massey-Martin and PERLA ratings correlated with online raters' expectations that a photographed person experienced greater discrimination. In contrast, the redness (a^*) and yellowness (b^*) undertones were highest in the mid-range of the rating scale scores and demonstrated greater overlap across race-ethnicities. Overall, each instrument showed sufficient consistency, comparability, and meaningfulness for use in field surveys when implemented soundly (e.g., not requiring memorization). However, PERLA might be preferred to Massey-Martin in studies representing individuals with the lightest skin tones, and handheld devices may be preferred to rating scales in

order to reduce measurement error when studies could gather only a single rating.

Significance Statement: As U.S. society continues to diversify and calls for better measurements of racialized appearance increase, survey researchers need guidance about effective strategies for assessing skin color in field research. Using controlled conditions, this study examined the consistency, comparability, and meaningfulness of the two most widely used skin tone rating scales and two portable and inexpensive handheld devices for skin color measurement.

1. INTRODUCTION

As U.S. society continues to diversify, survey researchers are challenged with quantifying race and ethnicity in ways that are socially and demographically meaningful. Survey questions about race-ethnicity have changed over time—such as by allowing respondents to choose more than one race—but many surveys, including the census, still fail to capture important aspects of racialized appearance like skin tone (Roth 2016; Telles 2018). This omission limits scholarly and societal understanding of the ways in which life experiences and opportunities are shaped by the color of a person's skin (Ennis et al. 2011; Qian and Lichter 2011). Telles (2018), for example, argued that skin color is such a central element of racial identity among people with Latin American heritage as to merit separate measurement in future census design. To this end, the present study contributes new evidence of the consistency, comparability, and meaningfulness of scores from four instruments that can be used to assess skin color in field research.

1.1 Literature Review

Demonstrated disparities by skin color on key outcomes of interest in the social sciences have made the need to assess skin color in survey contexts increasingly clear (Dixon and Telles 2017). An association between darker skin and lower educational attainment within race has been replicated across multiple samples (Branigan et al. 2013; Keith and Herring 1991), for example, and similar findings have been established for outcomes including wages (Goldsmith et al. 2006 2007), physical health (Sweet et al. 2007), partner selection (Udry et al. 1971), and political attitudes (Yadon and Ostfeld 2020). Survey researchers thus need guidance about which skin tone measures to include in surveys in order to effectively track across time and context the size and stability of these socially meaningful disparities among individuals, including those of the same race-ethnicity (Adams et al. 2016; Dixon and Telles 2017).

Even as the need to better measure skin color in social surveys is increasingly recognized, evidence about how to best do so is limited. The majority of large social surveys have measured skin color using respondent-coded or interviewer-coded categorical rating scales (Dixon and Telles 2017; Hannon and DeFina 2016 2020; Telles et al. 2015), but the reliability and validity of these scales remains an open question. One set of studies has demonstrated categorical skin color ratings of a single survey respondent to be inconsistent between survey waves (Hannon and DeFina 2016 2020), but it is unclear whether such variation reflects reliability and validity concerns inherent to the scales themselves versus variation in environmental or contextual factors in a field setting, such as ambient light, memorization of scales, and social interactions between interviewer and respondent. Another study attributed about one-fifth of the scale score variance to interviewers, although participants were nested within interviewers preventing examination of how different interviewers rated the same target (Cernat et al. 2019). To our knowledge, no prior study has assessed the comparability between in-person ratings of the two most common skin color rating scales used in US-based social surveys, nor between respondent and interviewer ratings on these scales, in a controlled setting.

In addition to categorical rating scales, more biologically-grounded fields such as physical anthropology and medicine have more commonly assessed skin color by measuring how light reflects off the skin using a colorimeter or spectrophotometer (Jablonski 2004; Jablonski and Chaplin 2000; Pershing et al. 2008; Wallace et al. 2000). Although earlier studies have examined the internal consistency of these device readings, information is lacking about newly available devices whose smaller size and lower cost make them feasible for large scale survey research (Das et al. 2016). It is also the case that scholars focused on social outcomes may assume human perceptions of skin tone are substantially different from these instruments'

readings, with the former expected to be more conceptually relevant to their research (Villarreal 2010). Yet, comparisons between categorical rating scales commonly used in surveys and handheld devices are lacking.

1.2 Studied Instruments

We specifically compare four different strategies for assessing skin color: two rating scales and two handheld devices. For the rating scales, we include the two most commonly used measures from recent social survey data collection efforts: the Massey-Martin scale (Massey and Martin 2003) and the PERLA scale (Telles 2014). The Massey-Martin scale was originally created for use in the New Immigrant Survey [NIS] but has since been implemented in numerous other nationally representative surveys (e.g., National Longitudinal Survey of Youth 1997 [NLSY97], General Social Survey [GSS], Fragile Families and Child Wellbeing Study). The scale ranges ten color categories from light to dark, with limited apparent variation in undertones (redness, yellowness). We reproduce the scale as used in the current study in the right panel of Figure 1. Like other recent studies such as the NLSY97, we omit the shape of a hand and shirt sleeve that was present in the earliest uses of the scale, given critiques that those features signal gender and socio-economic status; we also follow prior studies by omitting an albinism level zero. The PERLA scale, which was developed for the Project on Ethnicity and Race in Latin America, attempted to better capture undertones of redness and yellowness in addition to variation from light to dark in a set of eleven consecutively presented categories (Telles 2014). The PERLA Color Palette has since been adopted in other studies, including the biennial AmericasBarometer study (LAPOP 2018). The left panel of Figure 1 provides the scale as used in the current study.

For physical measurement of skin color, we utilize two recently available handheld

devices: Labby and Nix Mini (see Table 1). As a spectrophotometer, Labby captures light reflected from the skin across the full spectrum of humanly visible wavelengths. The palm-sized Labby was developed at the MIT Media Lab as a cost-effective means of measuring skin color, including to capture jaundice in resourced-limited settings. Labby interfaced with a smartphone app and cost just under $1000 at the time of our study (Das et al. 2016). Colorimeters, like the Nix Mini, focus on certain wavelengths, rather than the full light spectrum, making them smaller and less expensive than spectrophotometers. The Nix Mini had industry origins for measuring paint, also interfaced with smartphones, was about the size of a golf ball, and cost under $100 at the time of our study. Both devices provide readings scaled in the commonly used CIE *L*a*b** color space, providing coordinates on three axes: *L** capturing darkness-to-lightness, *a** quantifying greenness-to-redness, and *b** measuring blueness-to-yellowness. Whereas rating scales have rarely attempted to separate variation in undertones by disaggregating lightness, redness, and yellowness, the devices allow these distinct aspects of skin color to be studied.

1.3 Research Questions

We focus on three research questions, each addressing a central issue in survey measurement. First, we ask: how consistent is each measure of skin color? For the handheld devices, we answer this question by examining whether the instruments produce the same (or highly similar) values across multiple readings. For the rating scales, this entails whether different people rate a given target in the same way on the same scale (and, for in-person ratings, at the same time and in the same setting). Our approach importantly adds to limited evidence on this question. As noted, Cernat et al. (2019) attributed about 20% of the variance in PERLA ratings to interviewers in the AmericasBarometers survey, but their design did not have multiple interviewers rate the same target thus precluding their disentangling rater variance from true

score variance. Hannon and DeFina (2016) examined Massey-Martin scale ratings of participants between waves of the GSS and the American National Election Studies (ANES) but could not separate variation in rater perceptions from variation in rating conditions, such as context, clothing, lighting, and aging across the waves. The authors also emphasize the limitations of the approach used in GSS and ANES which required interviewers to memorize the scales and rate participants' skin tones after leaving the interview. These studies also could not fully consider the extent to which ratings vary across the race-ethnicity pairings of targets and raters, an important topic given widespread recognition of race of interviewer effects (Hill 2002; West and Blom 2017). We were able to examine this question with an online sample across raters who identified as Asian, Black, Latinx, and White. For the handheld devices, we build on prior studies (e.g., Shriver and Parra 2000) by considering repeated measures of the same location of the body, producing test-retest reliability data for the studied smaller and less expensive contemporary devices.

Second, we ask: how comparable are our four measures of skin color? The answer to this question is important, given the debate among social scientists, noted above, regarding the extent to which human perceptions differ from device readings (Villarreal 2010). The answer is also important given that, even as handheld devices become increasingly affordable, they can only be used in person and they are more expensive than rating scales both in direct cost and in terms of time needed for training and data collection. Comparing each of the $L^*a^*b^*$ coordinates to rating scale scores is also informative regarding the extent to which the rating scales primarily pick up the darkness-to-lightness (L^*) continuum or capture axes of redness (a^*) and yellowness (b^*) undertones. Reliable assessment of undertones can better recognize the wide variety of skin tone shades in the United States, including shades reflecting continued racial and ethnic

diversification. Undertones have been obscured in prior studies by early measures' focus on a single white-to-black continuum. Knowing how directly-assessed redness and yellowness relate to PERLA and Massey-Martin scores is also needed, given that the PERLA specifically aimed to better capture such undertones among persons in Latin American countries.

Our third research question asked: are the measurements socially meaningful? This question speaks in part to concerns in prior literature that interviewer-coded scales do not capture ample variation in skin color across all racial-ethnic groups (Branigan et al. 2013; Telles 2014). For example, self-identified White Americans had approximately half the variance in light reflectance as did self-identified Black Americans, as measured by colorimeter readings in the CARDIA study in the early 1990s (Branigan et al. 2013). As noted, PERLA attempted to better reflect socially relevant variation in skin undertone among Latinx populations (Telles 2014), yet, to our knowledge, prior studies have not used both Massey-Martin and PERLA nor used both colorimetry and spectrophotometry (including the three *L*a*b** dimensions) to compare variation in skin color across people identifying as White, Black, Latinx, and Asian in the United States. We fill this gap, allowing the first test of the potential greater variation captured by the PERLA versus Massey-Martin scale in U.S. populations and by the redness (*a**) and yellowness (*b**) device readings. We also use an online sample to examine how rater perceptions of the likely social experiences (including discrimination) of photographed individuals correlate with ratings of their skin color, again looking across target-rater race-ethnicity pairings.

Altogether, our study fills an important need by offering guidance to survey researchers regarding the leading options for capturing skin color in social surveys.

2. METHODS

2.1 Procedures

2.1.1 In-Person Sample. Data collection for the in-person sample occurred during 2018 at the University of Illinois at Chicago (UIC), a location well suited to our study given its racial-ethnic diversity with no single group in the majority (U.S. News & World Report n.d.). Our study protocol was informed by six undergraduate students—three of whom identified as Latina, one Black, one South Asian, and one White—who participated in a supervised research project in spring 2018 (supplemental online Appendix A provides the protocol which was approved by the UIC Institutional Review Board). Three of these undergraduate students (two Latina, one White) collected data in summer and fall 2018, along with the third author.

To recruit participants, study staff visited undergraduate classrooms to gather initial screener questionnaires. The screener asked potential participants to report their: (a) gender, (b) race-ethnicity, and (c) skin tone. The first and third authors reviewed the screener questionnaires on a rolling basis each week and used a stratified selection process to select 50 participants from a range of genders, race-ethnicities, and skin tones. For selection, we collapsed data to three categories of gender (male, female, other), four of race-ethnicity (Asian, Black, Latinx, White), and ten of skin tone. The ten skin tone regions were based on 66 color swatches developed by L'Oreal to identify the just-distinguishable gradations of skin color among diverse populations worldwide; these swatches varied in both darkness-to-lightness and in yellowness-to-redness (de Rigal et al. 2010). Altogether, we collected 230 screeners, sent 141 invitations to provide schedule availability, and received 87 responses (cooperation rate of 87/141 = 62%). We scheduled 82 visits to achieve the targeted 50 participants (cooperation rate of 50/82 = 61%; calculations based on COOP1; AAPOR 2016 p. 63). Data collected from the first four participants were omitted due to a problem with the Labby device software. This problem was fixed for the remaining 46 participants. Two sets of Labby readings were also missing due to

user error.

Table 2 summarizes the focal skin tone scores used in the study. At scheduled visits, data were collected in the following order by pairs of study staff: (a) a first self- and staff-rating of participant skin tone, (b) participant ratings of skin tone for 40 stock photos of strangers, (c) skin tone readings using the two handheld devices, (d) a second self- and staff-rating of participant skin tone; (e) a self- and staff-rating of participant race-ethnicity; and, (6) additional self-reported characteristics.

Participants and staff referred to the Massey-Martin and PERLA scales on a study-supplied tablet or laptop. The order of PERLA and Massey-Martin self-ratings was randomly counterbalanced across study participants. Order was also counterbalanced between the two study staff in rating participant skin tone with these scales. For the stock photos, each participant was randomly assigned to complete only one of the two rating scales, in order to reduce response burden. Five study staff also rated the stock photos with the Massey-Martin scale. Device readings were taken in quick succession from the same location on the participant's inner forearm, first three Labby readings and then three Nix readings. Study staff were trained in best practices (applying light pressure; avoiding veins, freckles, and blotches). Most data collection took place in a small windowless conference room with fluorescent lighting. Six visits took place in a similarly sized room with the same type of lighting and a small window. Participants dressed as they normally would for classes.

2.1.2 Online Sample. Additional data was collected in fall 2020 using the online participant recruitment platform Prolific with a protocol approved by the UIC Institutional Review Board (see supplemental Appendix B). Prolific caters to scientific researchers (Palan and Schitter 2018; Peer et al. 2017), and we used its pre-screening features to recruit U.S. residents

ages 20-39, who had 20/20 vision or were wearing corrective glasses/contacts, were not color blind, and had at least a 97% approval rate after completing at least five prior Prolific studies. Sixty eligible participants each were selected in eight categories that cross-classified Prolific's pre-screened binary sex (female, male) and four racial-ethnic (Non-Latinx Asian, Black, and White; Latinx) categories.

Prolific participants were randomly assigned either to rate the PERLA or the Massey-Martin scale. Participants first verified their age, vision, sex, and race-ethnicity; if any contradicted their Prolific screening profile then they exited the survey (n = 26). Qualtrics settings also screened out participants that Qualtrics' geographic (city) location indicator detected to be outside of the U.S. (n = 9). Participants using a mobile device also exited (n = 42). Eligible participants completed informed consent, provided additional demographics, practiced the skin tone scale, and then answered question sets (including skin tone and social experiences of persons depicted in stock photos), with set order randomized. The order of photos was also randomized within sets. Among the 459 participants who met eligibility and completed the survey, sample sizes ranged 55 to 59 in our 8 targeted gender by race-ethnicity cells (n = 25 to 32 for each scale, PERLA and Massey-Martin).

2.2 Materials and Measures

2.2.1 Skin Tone Scores.

In both studies, participants rated skin tone using the Massey-Martin and PERLA scales (Massey and Martin 2003; Telles 2014). As noted, Massey-Martin has 10 categories and PERLA 11 categories, with higher scores on each indicating darker skin tone (see again Figure 1). The Labby and Nix handheld devices took $L^*a^*b^*$ readings of in-person participants. The L^* readings could range from 0 to 100, with higher scores indicating lighter skin. For human skin tone, a^* and b^* values are positive with higher scores indicating darker

shades of redness (*a**) or yellowness (*b**).

2.2.2 Stock Photos. The stock photos were gathered in summer 2018 by one of the undergraduate students (who identified as Latina) with input from the first and third authors. Through internet searches, 40 stock photos were selected in categories defined by the 2 x 4 x 5 cross-classification of binary gender (male or female), four-categories of race-ethnicity (Asian, Black, Latinx, White), and five-levels of skin tone (from light to dark). Photos were selected to represent as much variation as possible in skin tone within each gender by race-ethnicity classification (see supplemental online Appendix C to access the photos).

2.2.3 Perceived Social Experiences. In the online-only (Prolific) study, the likely social experiences of photographed individuals were rated. Building on prior research (Dixon and Telles 2017; Adams, Kurtz-Costes, and Hoffman 2016), we asked raters to assess the chances that the photographed person would experience discrimination across social settings, such as in interactions with police, with healthcare, or when simply walking on the street. With these assessments, we contribute to the literature on skin color discrimination by quantifying social expectations of colorism—not just in individuals' reports of their own perceptions and preferences, but in people's extrapolations of how skin color discrimination is likely to operate for unknown individuals. Because the survey took place during the COVID-19 pandemic, raters were asked to think about the social world before the pandemic began, when people were not restricted by social distancing (i.e., they were asked to visualize where they lived during the last week of 2019 and to think about something they did that week). The stem reminded them of this reference period: "Around the last week of 2019, how likely is it that a person who looked like this would have." The six rated experiences were: (a) been discriminated against when trying to get a job? (b) been arrested after being stopped for speeding? (c) had a doctor take their

symptoms seriously? (d) had someone ask them for directions if lost? and (e) felt lonely? Response options were: 1 = Very Unlikely, 2 = Somewhat Unlikely, 3 = Somewhat Likely, 4 = Very Likely.

2.2.4 Demographics. During in-person data collection visits, college students reported detailed gender and racial-ethnic identities; they also reported personal characteristics, including their age, college major, class standing, parents' educational status, and country of origin. Online-only (Prolific) participants verified their binary sex (male, female) and four-category race-ethnicity (Asian, Black, Latinx, White) and reported their highest level of education as well as the urbanicity and region of their residence.

2.3 Sample Description

Appendix D (in online supplemental materials) describes both samples.

2.3.1 In-Person Sample. Regarding race-ethnicity, our selection process screened in nearly equal numbers of college students who identified as Asian, Latinx, and Black (n = 12, 13 and 13 respectively) and somewhat fewer who identified as White (8 participants). When allowed to elaborate on their race-ethnicity in person, four expanded their identifications (two screened as Latinx chose both Latinx and White in person; one screened as White chose both White and Latinx in person; one screened as Black wrote in "mixed black and white"). The majority of participants were females ages 18 to 22. Over half were first-generation U.S. citizens, and nearly two-fifths were first-generation college students. The students reported a wide range of majors and reflected all class standings, although the majority were seniors.

2.3.2 Online Sample. Reflecting our design, the online sample was uniformly distributed by gender and race-ethnicity (12 to 13% in each of the eight categories). Participants were also well distributed across their 20s and 30s and across regions of the U.S. Most lived in large cities

and had at least some college education, although some lived in smaller or rural areas and had a highest degree of high school.

3. ANALYSIS

3.1 Consistency

We examined consistency using the intraclass correlation (ICC) with the formula being: (a) two-way (targets and replications), (b) mixed effects (treating replications as fixed), and (c) absolute (to capture exact agreement, rather than simply consistency of ranked order; Shrout and Fleiss 1979). We reported both the individual ICC and average ICC values in order to demonstrate the degree to which multiple readings offered a gain in consistency over a single reading (formulas $\rho_{A,1}$ and $\rho_{A,k}$ in StataCorp 2019, p. 1059; see supplemental Appendix E for formulas and code). We calculated point estimates and 95% confidence intervals. Although strict thresholds for ICCs are debated, we used the reference points of .60 to .74 for good and .75 to 1.00 for excellent consistency (Cicchetti 1994; Lance et al. 2006). Our online-only (Prolific) design allowed us to estimate ICCs within target-rater race-ethnicity and gender pairings. We exceeded our goal of at least 15 completed ratings in each of the eight categories (two genders by four race-ethnicities), a goal set so as to achieve sufficient precision for 95% confidence intervals to be of width .2 or less (e.g., range from .7 to .9 if the true ICC was .8; Bonett 2002).

3.2 Comparability

3.2.1 Comparability of Readings between Handheld Devices.

Given high within-device consistency, we created three-reading averages for each device to use when examining cross-device comparability. The Pearson correlation assessed the degree of linear association between the two devices' three-reading averages. Because such correlations can be high even when one device has readings systematically higher or lower than the other, we also: (a) reported absolute

agreement by calculating ICCs that treated each device as a replication, and, (b) calculated the average of differences between scores. Scatterplots of the points visualize the associations underlying these values.

3.2.2 Comparability of Scores between Rating Scales. We similarly created averages of repeated rating scale scores of the same target. For in-person ratings, this average was based on three replications (one self-rating and two staff-ratings). For stock photos, this average was based on the 20 (PERLA) or 26 (Massey-Martin) college student ratings or the 222 (PERLA) or 216 (Massey-Martin) Prolific ratings. We then examined cross-scale comparability by creating scatterplots of the average scores on the two scales and calculating the Pearson correlation. Although the anchors of the two scales differ, we calculated the average difference in scores to help identify any consistent correspondence of categories from one scale to the other.

3.2.3 Comparability of Scores between Handheld Devices and Rating Scales. To examine the comparability between the device readings and rating scores, we graphed the average Massey-Martin and PERLA ratings against the average L^* (lightness), a^* (redness), and b^* (yellowness) values. In supplemental online materials (Appendix F), we also reported results of regressing each rating score average on the L^*, a^*, and b^* average readings, using quadratic terms to test for curvilinearity and interactions to test for different slopes by the four screener-classified racial-ethnic groups. For space constraints, we featured Labby results in the manuscript and provided Nix results in online supplemental Appendix F.

3.3 Meaningfulness

3.3.1 Overall Variation. To quantify the overall variation of scores, we reported the standard deviation (*SD*), mean, and their ratio (the coefficient of variation, *CV*), again using the averages across replications.

3.3.2 Within and Between Race-Ethnicity Variation. To visualize variation within and between race-ethnicity, we graphed the average PERLA and Massey-Martin ratings of college students' skin by their screener-reported race-ethnicity. We also graphed stock-photo average PERLA and Massey-Martin ratings by the photos' original classifications by study staff during photo selection. We used ANOVA, homogeneity of variance, and two-group exact randomization tests of whether means, variances, and coefficients of variation differed by race-ethnicity (Brown and Forsythe 1974; Kaiser and Lacy 2009; Rosner 2006).

3.3.3 Social Experiences. We correlated Prolific participants' average ratings of photographed individuals' likely social experiences with the photos' average skin tone ratings. We reduced shared method variance by using each set of ratings—skin color or social experience—only from raters who randomly saw that set first. This ensured that the averages were not affected by prior exposure, and, that different Prolific raters' scores contributed to averages of skin color and averages of social experiences. With these criteria, we retained 123 ratings of social experiences, 67 ratings of skin color based on PERLA, and 50 ratings of skin color based on Massey-Martin. The general pattern of results and conclusions are consistent when we used all Prolific participants (results available in supplementary online Appendix G). We also excluded from all analyses of social experiences one photo classified as the darkest-skinned Black female because her young perceived age made most of the social experiences' questions irrelevant.

4. RESULTS

4.1 Consistency

4.1.1 Consistency of Readings from Handheld Devices. Within-device consistency was excellent (top rows of Table 3). ICC point estimates were at least .94 for individual readings and

at least .98 for the three-reading averages.

4.1.2 Consistency of Scores from Rating Scales. Consistency of scores assigned by different raters to the stock photos were also excellent (bottom rows of Table 3). This was true for each type of rater and scale. PERLA point estimates were .87 for individual ICCs for college students and .85 for Prolific raters. Massey-Martin point estimates were .86 for college students, .91 for study staff, and .83 for Prolific raters.

In terms of in-person ratings of the college students' skin tone, consistency was also generally excellent. Individual ICCs had point estimates of .84 and .83 when comparing the two staff-ratings to each other on PERLA and Massey-Martin respectively. Consistency of participants' self-ratings with these staff-ratings was somewhat higher for PERLA (at .85 for each staff person) than Massey-Martin (at .71 to .77 for the two staff-ratings), a pattern we attribute to the greater variation of PERLA scores presented below.

4.1.3 Accounting for Target and Rater Race-Ethnicity and Gender. Point estimates of *ICCs* for Prolific ratings of stock photos within target-rater gender and race-ethnicity are graphed in Appendix H of online supplemental materials. Notably, with two exceptions, all *ICCs* were at or above .60 (good consistency). One exception was ratings of targets classified as White. Here, *ICCs* were lower across all of the rater race-ethnicities and genders and for both PERLA and Massey-Martin, suggesting a systematic source such as the constrained variation of skin tone among Whites shown below. The other exception was unique to Black males' ratings of Latinas with the Massey-Martin scale; given this result did not replicate with the PERLA scale nor across rater types it may reflect sampling error.

4.2 Comparability

Given the good-to-excellent consistency across repeated measures, we relied upon

averages when examining comparability of scores.

4.2.1 Comparability of Readings between Handheld Devices. The *L** (lightness) readings were highly comparable between the Nix and Labby three-reading average scores, with a linear correlation of .95 (see also scatterplot in top panel of Appendix I in online supplemental materials). The ICC was .75, reflecting that Nix scores tended to be systematically lower than Labby scores by an average of about 6 points. In contrast, between-device comparability was lower for the *a** (redness) and *b** (yellowness) readings (middle and bottom panels of Appendix I). Pearson correlations were moderate for *a** and *b** (below .60), and *ICCs* were also low (below .30), with mean differences of about 3 and 12 points respectively.

4.2.2 Comparability of Scores between Rating Scales. PERLA and Massey-Martin average scores were highly linearly related, both for college students' skin (r = .97) and for stock photos (r = .98 to 1.00; see supplementary Appendix J). PERLA scores were systematically about 1.0 to 1.5 points higher than Massey-Martin scores, consistent with PERLA's 11-point versus Massey-Martin's 10-point scale. In Figure 1, we had vertically aligned the swatches to reflect this correspondence. A floor effect was also evident for the Massey-Martin scale: stock photos that were rated a 1 on Massey-Martin were distributed among values from 1 to 3 on PERLA.

4.2.3 Comparability Between Handheld Devices and Rating Scales. A consistently linear association was visually evident between *L** scores and the PERLA and Massey-Martin scores (see Figure 2). Recall that the negative associations are expected, since higher *L** values reflect lightness whereas higher rating scale scores reflect darkness. In Appendix F of supplemental online materials, regression analyses showed that this negative association was most evident for participants who identified as Black, but also significantly negative for participants who

identified as Asian and Latinx, and least evident for participants who identified as White where variation was most constrained.

For *a** (redness) and *b** (yellowness) readings, in contrast, the color undertones had higher values (were most saturated) in the middle of the PERLA and Massey-Martin scales, and lower values (less saturated) at both the lower and the higher ends of the scales. In other words, a negative association was seen between *L** (lightness) and both *a** (redness) and *b** (yellowness) for participants who identified as Black—and had the darkest skin tones. These individuals were represented with black circles in Figure 2. Among those identified with race-ethnicities other than Black—and had lighter skin tones—the associations were positive. These individuals were represented with grey and white circles in Figure 2.

4.3 Meaningfulness

4.3.1 Overall Variation. As expected, due to PERLA's 11-point rather than 10-point scale, the PERLA scores had a larger range and *SD* and higher mean than the Massey-Martin scores (top panels of Table 4). But, the *CV*—which cancelled out the different metric of each scale—was higher for the Massey-Martin ratings. For the device readings (bottom panels of Table 4), the Labby scores had higher values on all statistics than Nix for *L** (lightness). In contrast, *a** (redness) and *b** (yellowness) showed higher *SDs* and *CVs* for Nix.

4.3.2 Within and Between Race-Ethnicity Variation. Several salient patterns emerged for the ratings and readings of college students' skin color (see graphs in supplemental Appendix K). Regarding means, a meaningful pattern was evident for the PERLA and Massey-Martin rating scales and for the *L** (lightness) readings, in that darkness was highest for Blacks and lowest for Whites with Latinx and Asians falling in between. PERLA and *L** picked up more variation and reflected greater overlap among race-ethnicities than did Massey-Martin, however. There were

fewer and less consistent differences in mean scores for *a** (redness) and *b** (yellowness) across race-ethnicity. Regarding variances, the most consistent pattern was some evidence of the largest standard deviations in PERLA, Massey-Martin and *L** (lightness) scores for Blacks, and the highest coefficients of variation for *L** for Blacks. Consistent differences were not evident by race-ethnicity for variation in *a** (redness) and *b** (yellowness) scores.

In relation to ratings of the stock photos, as expected, photos we had initially classified as darker skinned generally had higher average ratings (see Figure 3). The top scores for the White photos fell below the bottom scores for the Black photos. Greater variation was also evident in scores among Blacks than Whites, the former having about twice the range of the latter and especially on the Massey-Martin scale. The photos classified as Asian and Latinx fell in the middle, each with comparable or larger variation as those classified as Black and with scores overlapping the average values of those classified as White and as Black at each extreme.

4.3.3 Social Experiences. Online (Prolific) participants' ratings of the likely social experiences of photographed individuals confirmed that people perceived to have darker skin were also expected to be more likely to: (a) experience job discrimination, (b) be arrested, (c) be less likely to have others ask them for directions, and, (d) have doctors take their symptoms seriously (see Table 5). These correlations were generally large (above |.8|) in magnitude and were evident for both the Massey Martin and PERLA scales as well as across race-ethnicity and gender, with the exception of white females. The correlations were smaller for the social experience of loneliness.

5. DISCUSSION

5.1 Summary of Findings and Implications for Survey Researchers

Our results—based on data collected under controlled conditions—demonstrate that if a

study is focused on skin lightness-darkness (not redness or yellowness), then survey researchers could select either handheld device or either rating scale and obtain almost identical conclusions. Results were also consistent enough across three readings from each handheld device that a single reading would generally suffice from the perspective of measurement error. We did experience data loss due to user error in two cases for the Labby, which might warrant repeated measurement, although a recently released next generation Labby with an improved interface might reduce such user error. Data loss due to user error was not an issue for Nix, which had been commercially available for longer and had a more robust user interface. Survey researchers might thus choose the less expensive Nix over Labby if the L^* dimension is of primary interest and full spectral data is not needed. The Nix is also smaller, lighter weight, and designed to withstand rougher handling than the Labby version that we tested, which could be advantageous in a field survey setting.

Our results also demonstrated that, in a controlled environment, different raters provided consistent ratings of the same person with both the PERLA and the Massey-Martin scales. This was especially true for perceptions of stock photos which had similar *ICC*s across rating modes (rated during an in-person visit versus online only) and rater type (college student, study staff, other adults). Consistency was evident across the race-ethnicities of raters and of the target photos in our online study, with the primary exception of ratings of white target photos where skin tone range was most constrained. For in person ratings, consistency was also good-to-excellent, although participants ratings of their own skin tone differed more from staff on the Massey-Martin scale possibly because it demonstrated less variation in the lighter range of the scale than did PERLA. Although results suggest a single rating would typically be sufficient, survey researchers might consider strategies such as taking photos of study participants in the

field that could then be rated by multiple people (including using online panels, such as our Prolific surveys) in order to reduce measurement error by averaging scores and in order to avoid requiring interviewers to memorize the person's skin tone or the rating scale as have many prior surveys.

Compared to each other, the PERLA and Massey-Martin scores were more similar than might have been expected based on their historical origins. The linear correlations between scores exceeded .96 both for the stock photos and for the in-person ratings, suggesting that correlations with other variables would be highly similar regardless of which scale was used (as we saw in relation to online ratings of photographed individuals' likely social experiences). When we lined up the two scales' color swatches based on our results, this similarity was apparent, with a PERLA score corresponding to a Massey-Martin score one-to-two numbers lower (see again Figure 1). PERLA scores did better differentiate skin color among those with the lightest tones, however, which might lead survey researchers to choose it for samples that include lighter-skinned participants.

Whereas the darkness-to-lightness (L*) device readings correlated highly with the rating scale scores, consistency and comparability were lower for the redness (a*) and yellowness (*b**) readings. Potentially, the inconsistencies between Labby and Nix on these tones reflected their different technologies. Both Labby and Nix used an instrument geometry that captured how a beam of light bounced off the skin at a 45° angle, but the opening through which the light passed (aperture) was larger for Nix which can affect readings of colors with different wavelengths (longest for red, shortest for blue, with yellow and green in between). A newly available attachment for Nix also aims to improve its measurement of skin, including the redness and yellowness tones.

5.2 Discussion of Results in Relation to Prior Studies and Future Research

Our results add to previous research regarding survey-based assessment of skin color, emphasizing the need for collecting skin color data in social surveys. Extending prior research on self-reports and self-perceptions (Dixon and Telles 2017; Uzogara and Jackson 2016), we affirmed that raters perceive individuals with darker skin color as being more likely to experience discrimination across multiple contexts: with law enforcement, when accessing healthcare, in an employment setting, and when simply standing on the street. Our findings are particularly novel in that we were able to remove shared variance due to exposure or halo effects and due to common raters—i.e., we calculated skin tone averages based on one set of raters and we calculated averages of likely discriminatory experience using another set of raters, each set seeing the photos for the first time. Interestingly, respondents did not infer differential emotions (loneliness) to target photos on the basis of skin color; raters expected differences by skin color only in how an individual was treated by others.

Our findings that skin color rating scales commonly used in social surveys are highly comparable to the *L** (darkness-to-lightness) reading taken by the handheld devices might be surprising to social scientists who expect human perception of skin tone to reflect cues beyond "true" color, such as racialized facial features or other contextual cues. Yet, this result is unsurprising from a color science perspective, as *L*a*b** color measurement aims to approximate human color perception. The greater consistency among human raters in our in-person study than in prior studies may be due to our use of a controlled setting, with same-time and same-setting ratings. In contrast, prior field-based studies examined ratings from different times and settings (Cernat et al. 2019; Hannon and DeFina 2016 2020). We also presented color swatches below photographs during ratings, differing from other studies that, as noted above,

asked interviewers to memorize the scale (e.g., ANES, NIS, GSS). Prior studies also often asked interviewers to rate skin tone at the end of interviews, after learning about people's backgrounds and their interactional styles and with the potential for rating based on memory of the participant's skin color (e.g., NLSY97, AmericasBarometer). Handheld devices have the advantage of avoiding such bias due to context, memory, or interviews, when a study's focus is on correlates of the darkness-to-lightness of skin. When a study's interest is studying interviewer variance, survey researchers might take photos of the same people in different settings (e.g., home, workplace, street) and with different cues (e.g., clothing, hair style, background art/books, nighttime lighting), or use photo editing software to vary contextual cues and appearance (e.g., altering skin tones, facial features, hair texture/style, background, lighting). If paired with in-person device readings, such studies would provide valuable information on whether device readings differ more from human perceptions in certain circumstances, and on which specific social and physical factors other than skin color may influence interviewer-coded skin color measures collected in a field context.

Future research should also prioritize better capturing the undertones of redness and yellowness, which we found overlapped more across the racial-ethnic groups than did lightness-to-darkness but was least well measured. Capturing the full multidimensional complexity of human skin tone will best reflect the diverse shades of brown evident in contemporary U.S. society and around the world. To date, such variety of skin tones has been better recognized in the beauty industry than the social sciences. For instance, one skincare company used spectrophotometer readings from a global sample of individuals to identify 66 shades (de Rigal et al. 2010). These shades might be used to expand current rating scales, such as in the lighter tones where we found both PERLA and Massey Martin captured limited variation.

5.3 Limitations

The limitations of our study should be considered when interpreting our results. The results for handheld devices were based on a relatively small sample of students from one university. Although the participants were purposefully sampled to reflect diverse race-ethnicity and skin tone, a replication using the handheld devices with a larger and more representative sample would importantly extend evidence. Our online surveys achieved larger sample sizes, although future studies might go even further such as by having sufficient numbers for statistical power to examine additional characteristics such as the rater's own skin color. Our staff raters were also diverse, but not numerous enough to probe systematic rater variance for in-person ratings, although we did offer such evidence for our larger online rating sample. We likewise selected a diverse set of stock photos for participants and staff to rate, but were limited by available options (e.g., photos varied by pose, clothing, and age). Future studies might use photos taken under uniform conditions. In addition, because our focus was on portable and affordable options for measurement of skin color in survey settings, we used only handheld devices that were small, lightweight, and relatively inexpensive. The devices were calibrated to industry standards, but we did not directly compare our readings with larger and costlier spectrophotometers. We also examined additional aspects of validity, such as associations with social experiences, only for perceptions of stock photos in relation to scale ratings.

With the above limitations in mind, our study offered important information to survey researchers regarding widely used contemporary options for skin color assessment.

Table 1

Handheld Devices Used for In-Person (College Student) Study

Devices	Type		Cost	Size
Nix	Colorimeter (specific light wavelengths)		~ $100	Golf-ball-sized
Labby	Spectrophotometer (all light wavelengths)		~ $1000	Palm-sized
Device Readings				
*L**	Darkness-to-Lightness	Higher scores = lighter skin		
*a**	Greenness-to-Redness	Higher scores = redder skin		
*b**	Blueness-to-Yellowness	Higher scores = yellower skin		

Table 2

Listing of Instruments Used in the Current Study, and When They Were Employed by Whom to Assess What

Ratings of participants' skin (In-Person Study)

What (target)	What (instrument)	Who (rater)	When
Participant's skin	Massey-Martin	Participant Staff1 Staff2	Which instrument (Massey-Martin or PERLA) was completed first was randomly assigned.
Participant's skin	PERLA	Participant Staff1 Staff2	

Device readings of participants' skin (In-Person Study)

What (target)	What (instrument)	Who (operator)	When
Participant's skin	Labby Nix	Staff1 & Staff2	3 readings in succession for each device, with Labby before Nix

Ratings of stock photos (In-Person and Online-Only Studies)

What (target)	What (instrument)	Who (rater)	When
Stock photos	Massey-Martin	Five study staff 1/2 of UIC students 1/2 of Prolific sample	In-person participants rated photos after their 1st self-rating and before device readings; Staff rated photos before data collection began. Online participants rated photos counterbalanced randomly before or after rating the photographed individuals' likely social experiences.
	PERLA	1/2 of UIC students 1/2 of Prolific sample	

Note. The in-person study was conducted with UIC students in 2018. The online study was conducted in 2020 on the Prolific platform. For the in-person study, staff conducted data collection in pairs. The two staff are designated as Staff1 and Staff2. Assignment of 1/2 of the UIC students and 1/2 the Prolific sample to rate the stock photos with either Massey-Martin or PERLA was randomized.

Table 3

Intraclass Correlations (ICCs) for Device Readings and Rating Scales from In-Person (College Student) and Online-Only (Prolific) Samples

	n		Individual Raters/Readings			Across Rater/Reading Averages		
				95% *CI*			95% *CI*	
	targets	readings/ raters	point estimate	lower	upper	point estimate	lower	upper
Readings with Labby								
*L** (lightness)	45	3	0.99	0.98	0.99	1.00	0.99	1.00
*a** (redness)	44	3	0.95	0.93	0.97	0.98	0.97	0.99
*b** (yellowness)	44	3	0.94	0.90	0.96	0.98	0.96	0.99
Readings with Nix								
*L** (lightness)	46	3	0.99	0.99	1.00	1.00	1.00	1.00
*a** (redness)	46	3	0.94	0.90	0.96	0.98	0.97	0.99
*b** (yellowness)	46	3	0.98	0.97	0.99	0.99	0.99	1.00
Ratings of Stock Photos								
In-Person Study								
Participants, PERLA	39	20	0.87	0.82	0.92	0.99	0.99	1.00
Participants, Massey-Martin	38	26	0.86	0.80	0.91	0.99	0.99	1.00
Study Staff, Massey-Martin	40	5	0.91	0.86	0.95	0.98	0.97	0.99
Online Study								
Participants, PERLA	40	222	0.85	0.78	0.90	1.00	1.00	1.00
Participants, Massey-Martin	40	216	0.83	0.76	0.89	1.00	1.00	1.00
Ratings of College Students								
Staff1 & Staff2, PERLA	46	2	0.84	0.73	0.91	0.91	0.84	0.95
Staff1 & Staff2, Massey-Martin	46	2	0.83	0.71	0.90	0.91	0.83	0.95
Staff1 & Self, PERLA	46	2	0.85	0.74	0.91	0.92	0.85	0.95
Staff1 & Self, Massey-Martin	46	2	0.71	0.54	0.83	0.83	0.70	0.91
Staff2 & Self, PERLA	46	2	0.85	0.75	0.92	0.92	0.86	0.96
Staff2 & Self, Massey-Martin	46	2	0.77	0.62	0.87	0.87	0.76	0.93

Note. ICCs are based on a two-factor mixed model for absolute agreement using multiple readings/ratings per target. CIE 1976 L*a*b* readings measure continua of darkness-to-lightness (L*), greenness-to-redness (a*), and blueness-to-yellowness (b*). Higher values reflect the term on the right in each pairing. Two participants were missing Labby readings due to user error. PERLA and Massey-Martin are rating scales applied to stock photos and to in-person participants; higher scores reflect darker skin (and sometimes greater undertones of yellowness and redness). Staff1 was the study staff person who rated the in-person participant before using the handheld devices to measure skin tone. Staff2 rated the in-person participant after device measurement. Self was the in-person participant's self rating. One photo had missing data from in-person participants' PERLA scores as did two photos for in-person participants' Massey-Martin scores.

Table 4
Variation of Average Rating Scale Scores and Device Readings from the In-Person (College Student) and Online-Only (Prolific) Studies

	n	*Min*	*Max*	*Range*	*Standard Deviation*	*Mean*	*Coefficient of Variation*
Target: Photos							
In-Person Study							
PERLA	40	1.25	9.85	8.60	2.30	4.73	0.48
Massey-Martin	40	1.00	8.27	7.27	1.99	3.19	0.62
Online Study							
PERLA	40	1.30	10.21	8.91	2.49	4.55	0.55
Massey-Martin	40	1.07	8.61	7.54	2.16	3.43	0.63
Target: In-Person Participants							
Rating scales							
PERLA	46	1.67	10.67	9.00	2.00	4.54	0.44
Massey-Martin	46	1.00	8.00	7.00	1.83	2.88	0.63
Device readings							
Labby							
*L**	45	37.03	85.24	48.21	11.79	69.04	0.17
*a**	45	9.80	17.53	7.73	1.68	13.09	0.13
*b**	45	17.67	30.02	12.36	2.60	25.51	0.10
Nix							
*L**	46	39.00	74.67	35.67	8.61	62.83	0.14
*a**	46	3.67	19.00	15.33	3.62	10.28	0.35
*b**	46	7.33	19.00	11.67	3.53	13.88	0.25

Note. Values based on averages of repeated measures. For photos, averages are of in-person or online participants' ratings of each photo with each scale (20 and 222 ratings with PERLA scale respectively, 26 and 216 ratings with Massey-Martin scale). For rating scale scores of the in-person participants, averages are of the one self-rating and two study staff ratings. For device readings of in-person participants' skin, averages are of the three readings taken in succession with each device. There were 40 stock photos and 46 in-person participants. One in person participant was missing all three Labby readings due to user error.

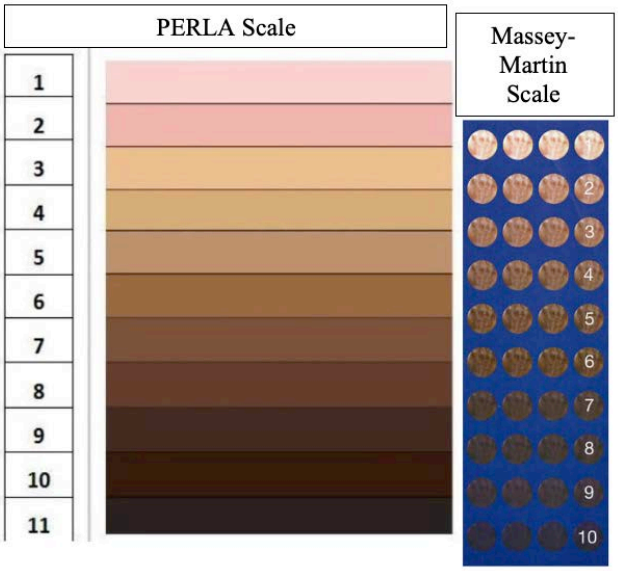
Table 5

Correlations of Skin Color Ratings with Likely Social Experiences, Online Participants' Ratings of Stock Photos

		Around the last week of 2019, how likely is it that a person who looked like this would have				
	n photos	Been discriminated against when trying to get a job?	Been arrested after being stopped for speeding?	Had a doctor take their symptoms seriously?	Had someone ask them for directions if lost?	Felt lonely?
PERLA ratings of:						
All Photos	39	**0.86**	**0.83**	**-0.84**	-0.73	0.31
Photos of:						
Asian Females	5	0.77	**0.91**	-0.77	0.19	0.49
Asian Males	5	**0.97**	**0.92**	**-0.91**	**-0.96**	**0.89**
Black Females	4	**0.92**	**0.96**	-0.80	**-0.91**	0.63
Black Males	5	**0.83**	0.76	-0.78	**-0.89**	0.50
Latinx Females	5	**0.94**	**0.81**	**-0.89**	**-0.92**	-0.41
Latinx Males	5	**0.89**	**0.92**	**-0.89**	**-0.87**	0.13
White Females	5	0.46	0.15	-0.40	-0.71	-0.19
White Males	5	**0.89**	0.54	**-0.84**	-0.63	0.12
Massey-Martin ratings of:						
All Photos	39	**0.85**	**0.84**	**-0.83**	-0.72	0.32
Photos of:						
Asian Females	5	0.77	**0.92**	-0.77	0.14	0.50
Asian Males	5	**0.97**	**0.92**	**-0.91**	**-0.96**	**0.90**
Black Females	4	**0.91**	**0.93**	-0.76	**-0.90**	0.65
Black Males	5	**0.87**	**0.81**	-0.80	**-0.89**	0.48
Latinx Females	5	**0.96**	**0.86**	**-0.88**	**-0.88**	-0.37
Latinx Males	5	**0.87**	**0.91**	**-0.87**	**-0.87**	0.13
White Females	5	0.29	0.33	-0.55	-0.62	-0.33
White Males	5	**0.84**	0.72	**-0.97**	**-0.83**	0.26

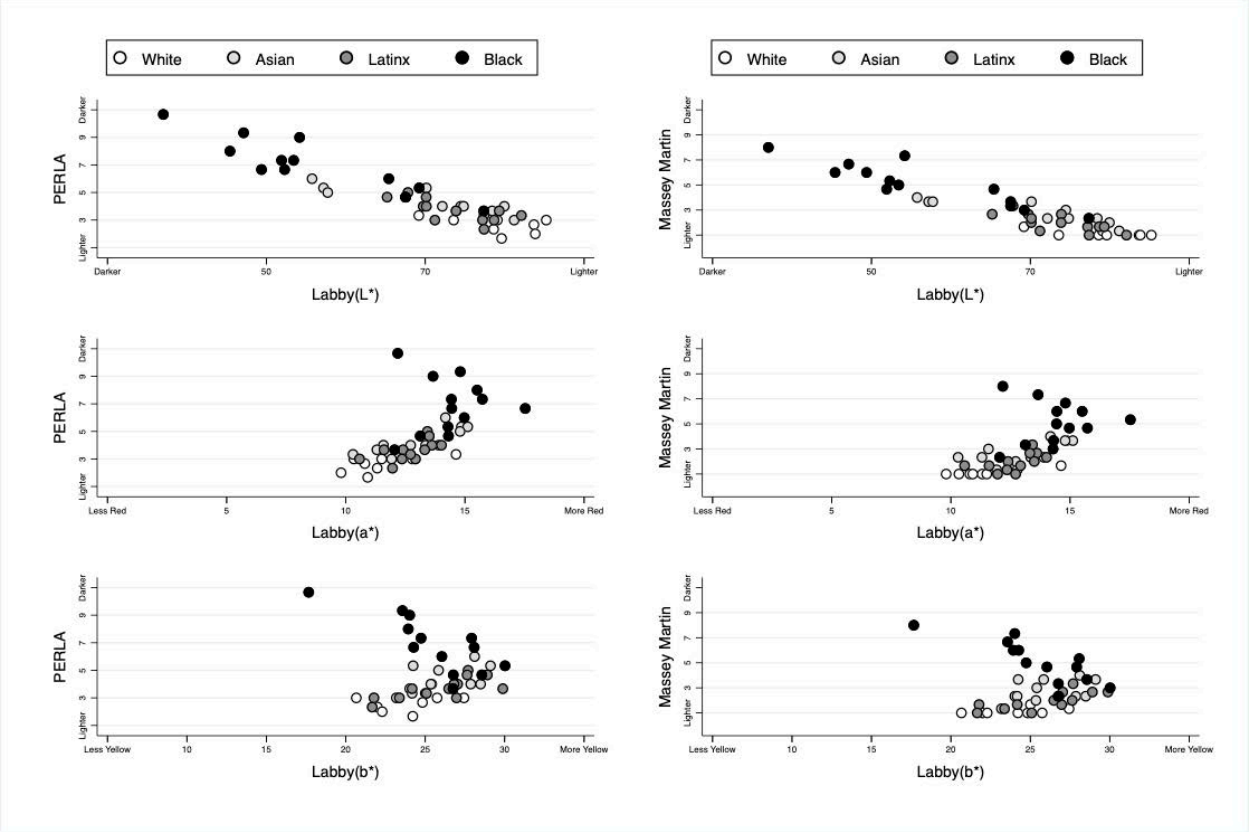
Note. Values are correlations between averages of skin tone and social experiences ratings of stock photos (those above |.8| are bolded). One photo excluded because its young perceived age made most of the social experiences' questions irrelevant. PERLA and Massey Martin are rating scales completed by Prolific raters; higher scores reflect darker skin (and sometimes greater undertones of yellowness and redness). For the likelihood of the five social experiences, response options were: *1 = Very Unlikely, 2 = Somewhat Unlikely, 3 = Somewhat Likely, 4 = Very Likely.* Shared method variance was reduced by selecting only ratings of each type (skin color, social experience) from Prolific participants who randomly saw that set first (n = 123 ratings of social experiences, 67 ratings of skin color based on PERLA, and 50 ratings of skin color based on Massey-Martin).

Figure 1
Rating Scales as Used in Current Studies



Source: https://nis.princeton.edu/downloads/nis-skin-color-scale
https://perla.soc.ucsb.edu/data/color-palette

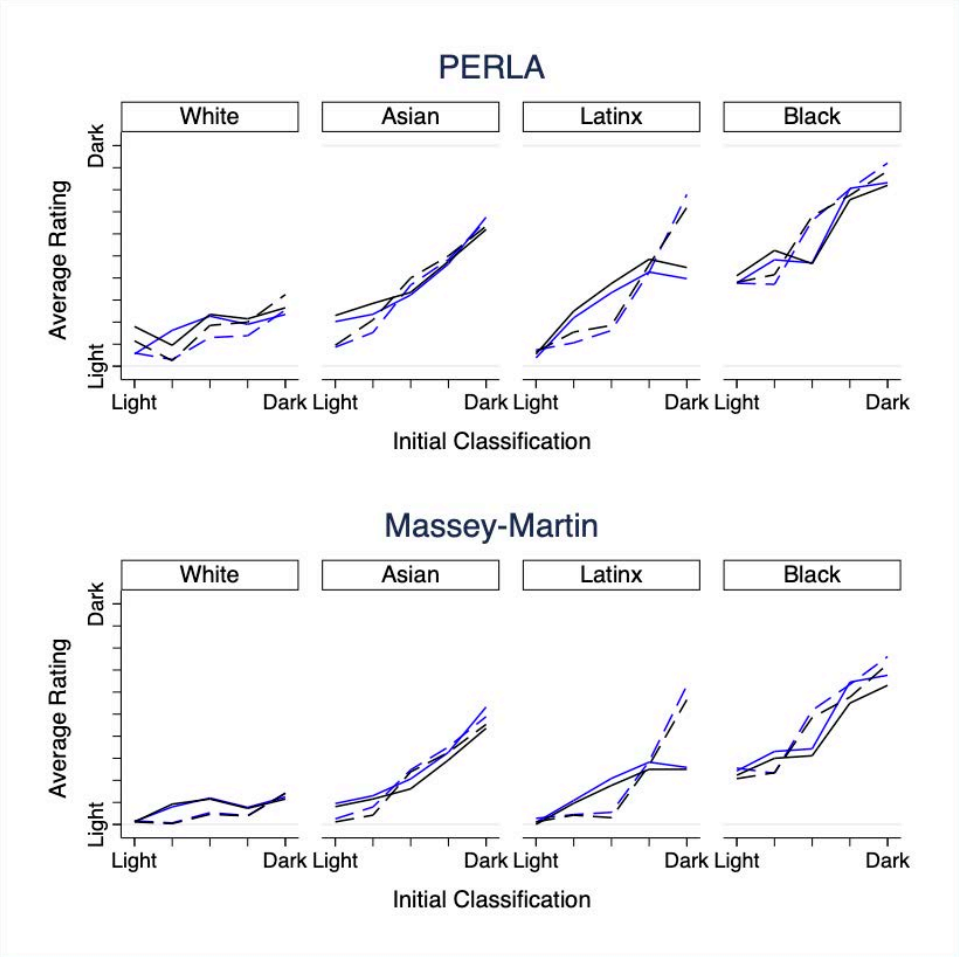
Figure 2
Associations of Average Rating Scale Scores with Average Device Readings in In-Person (College Student) Sample



Note. $n = 45$. Values based on averages of repeated measures. For rating scale scores, averages are of the one self-rating and two study staff ratings. For device readings, averages are of the three readings taken in succession with each device.

Figure 3

Ratings of Stock Photos from In-Person (College Student) and Online-Only (Prolific) Samples, by Study Staff Initial Gender, Racial-Ethnic, and Skin Tone Classifications of the Photos



Note. Values on the Y-axis are average ratings of 40 photos on PERLA (top panel) or Massey Martin (bottom panel). Values on the X-axis are initial classifications by study staff of each photo into 5 levels of darkness. Four racial-ethnic groups are based on study staff's initial classifications. Staff classification of photos as male and female are reflected by dashed and straight lines, respectively. Blue lines based on averages of Prolific participants' ratings (n = 222 PERLA; n = 216 Massey-Martin). Black lines based on averages of college students' ratings (n = 20 PERLA; n = 26 Massey-Martin).

REFERENCES

Adams, E. A., Kurtz-Costes, B. E. and Hoffman. A. J. (2016). "Skin Tone Bias among African Americans: Antecedents and Consequences across the Life Span," *Developmental Review*, 40, 93–116.

Bonett, D. G. (2002). "Sample Size Requirements for Estimating Intraclass Correlations with Desired Precision," *Statistics in Medicine,* 21, 1331-1335.

Branigan, A. R., Freese, J., Patir, A., McDade, T. W., Liu, K., and Kiefe, C. I. (2013), "Skin Color, Sex, and Educational Attainment in the Post-Civil Rights Era," *Social Science Research*, 42, 1659–1674.

Brown, M. B., and A. B. Forsythe. (1974), "Robust Tests for the Equality of Variances*," Journal of the American Statistical Association*, 69, 364–367.

Cernat, A., Sakshaug, J. W., and Castillo, J. (2019). "The Impact of Interviewer Effects on Skin Color Assessment in a Cross-National Context," *International Journal of Public Opinion Research*, 31, 779-793.

Cicchetti, D. V. (1994). "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment,* 6, 284-290.

Das, A. J., Wahi, A., Kothari, I., and Raskar, R. (2016), "Ultra-Portable, Wireless Smartphone Spectrometer for Rapid, Non-Destructive Testing of Fruit Ripeness," *Scientific Reports,* 6, 32504.

de Rigal, J., Des Mazis, I., Diridollou, S., Querleux, B., Yang, G., Leroy, F., and Barbosa, V. H. (2010), "The Effect of Age on Skin Color and Color Heterogeneity in Four Ethnic Groups," *Skin Research and Technology,* 16, 168–78.

Dixon, A. R., and Telles, E. E. (2017), "Skin Color and Colorism: Global Research, Concepts, and Measurement," *Annual Review of Sociology,* 43, 405–24.

Ennis, S. R., Ríos-Vargas, M., and Albert, N. G. (2011), "The Hispanic Population," *2010 Census Briefs*, C2010BR-04.

Goldsmith, A. H., Hamilton, D., and Darrity Jr., W. (2006). "Shades of Discrimination: Skin Tone and Wages." *The American Economic Review,* 96, 242–45.

Goldsmith, A. H., Hamilton, D., and Darity Jr., W. (2007). "From Dark to Light: Skin Color and Wages among African-Americans," *The Journal of Human Resources,* 42, 701–38.

Hannon, L., and DeFina, R. (2014), "Just Skin Deep? The Impact of Interviewer Race on the Assessment of African American Respondent Skin Tone," *Race and Social Problems*, 6, 356-364.

Hannon, L., and DeFina, R. (2016), "The Reliability of Same-Race and Cross-Race Skin Tone Judgments," *Public Opinion Quarterly,* 80, 534–41.

Hannon, L., and DeFina, R. (2020), "Reliability Concerns in Measuring Respondent Skin Tone by Interviewer Observation," *Race and Social Problems,* 12, 186–194.

Hill, M. E. (2002). "Race of the Interviewer and Perception of Skin Color: Evidence from the Multi-City Study of Urban Inequality," *American Sociological Review*, 67, 99-108.

Jablonski, N. G. (2004), "The Evolution of Human Skin and Skin Color," *Annual Review of Anthropology,* 33, 585–623.

Jablonski, N. G., and Chaplin, G. (2000), "The Evolution of Human Skin Coloration." *Journal of Human Evolution,* 39, 57–106.

Kaiser, J., and Lacy, M. G. (2009), "A General-Purpose Method for Two-Group Randomization Tests," *The Stata Journal*, 9, 70-85.

Keith, V. M., and Herring, C. (1991). "Skin Tone and Stratification in the Black Community," *The American Journal of Sociology*, 97, 760–78.

Lance, C. E., Butts, M. M., and Michels, L. C. (2006), "The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?" *Organizational Research Methods,* 9, 202-220.

LAPOP. (2018), "About AmericasBarometer," Available at https://www.vanderbilt.edu/lapop/about-americasbarometer.php.

Massey, D. S., and Martin, J. A. (2003), "The NIS Skin Color Scale." Available at https://nis.princeton.edu/downloads/nis-skin-color-scale

Palan, S., and Schitter, C. (2018), "Prolific.ac—A Subject Pool for Online Experiments," *Journal of Behavioral and Experimental Finance,* 17, 22-27.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017), "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research," *Journal of Experimental Social Psychology*, 70, 153-163.

Pershing, L. K., Tirumala, V. P., Nelson, J. L., Corlett, J. L., Lin, A. G., Meyer, L. J., and Leachman, S. A. (2008), "Reflectance Spectrophotometer: The Dermatologists' Sphygmomanometer for Skin Phototyping?" *Journal of Investigative Dermatology,* 128, 8.

Qian, Z. and Lichter, D. T. (2011), "Changing Patterns of Interracial Marriage in a Multiracial Society," *Journal of Marriage and Family*, 73, 1065-1084.

Rosner, B. (2006), *Fundamentals of Biostatistics,* Belmont, CA: Thomson.

Roth, W. D. (2016), "The Multiple Dimensions of Race," *Ethnic and Racial Studies,* 39,1310–1338.

Shriver, M. D., and Parra, E. J. (2000), "Comparison of Narrow-Band Reflectance Spectroscopy and Tristimulus Colorimetry for Measurements of Skin and Hair Color in Persons of Different Biological Ancestry," *American Journal of Physical Anthropology,* 112, 17–27.

Shrout, P. E., and Fleiss, J. L. (1979), "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin,* 86, 420–428.

StataCorp. (2019), *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

Sweet, E., McDade, T. W., Kiefe, C. I., and Liu, K.. (2007). "Relationships Between Skin Color, Income, and Blood Pressure Among African Americans in the CARDIA Study," *American Journal of Public Health*, 97, 2253–2259.

Telles, E. (2014), *Pigmentocracies: Ethnicity, Race, and Color in Latin America,* Chapel Hill, NC: The University of North Carolina Press.

Telles, E., Flores, R. D., and Urrea-Giraldo, F. (2015). "Pigmentocracies: Educational Inequality, Skin Color and Census Ethnoracial Identification in Eight Latin American Countries," *Research in Social Stratification and Mobility*, 40, 39-58.

Telles, E. (2018), "Latinos, Race, and the U.S. Census." ANNALS, AAPSS, 677, May 2018.

Udry, J. Richard., Bauman, Karl E., and Chase, Charles. (1971). "Skin Color, Status, and Mate Selection," *American Journal of Sociology*, 76, 722-733.

U.S. News & World Report. n.d. "Campus Ethnic Diversity," Available at https://www.usnews.com/best-colleges/rankings/national-universities/campus-ethnic-diversity

Uzogara, E. E., and Jackson, J. S. (2016). "Perceived Skin Tone Discrimination Across Contexts: African American Women's Reports," *Race and Social Problems*, 8, 147–59.

Villarreal, A. (2010), "Stratification by Skin Color in Contemporary Mexico." *American Sociological Review,* 75, 652-678.

Wallace, V. P., Crawford, D. C., Mortimer, P. S., Ott, R. J., and Bamber, J. C. (2000), "Spectrophotometric Assessment of Pigmented Skin Lesions: Methods and Feature Selection for Evaluation of Diagnostic Performance," *Physics in Medicine & Biology,* 45, 735-751.

West, B. T., and Blom, A. G. (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175-211.

Yadon, N., and Ostfeld, M. C. (2020). "Shades of Privilege: The Relationship Between Skin Color and Political Attitudes Among White Americans," *Political Behavior*, 42, 1369–1392.