Manuscript for the Special Issue of *Research in Human Development* on

New Directions for Research on Colorism Across the Lifespan

How Well Do Contemporary and Historical Skin Color Rating Scales Cover the Lightness-to-Darkness Continuum? Descriptive Results from Color Science and Diverse Rating Pools

Mariya Adnan Khan and Hai Nguyen

University of Illinois at Chicago

Amelia R. Branigan

University of Maryland College Park

Rachel A. Gordon

Northern Illinois University

* Corresponding author: Mariya Adnan Khan, mkhan252@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. The authors thank the Skin Tone Identities and Inequalities Project team, including co-project coordinators Johanna Nunez and Dahlya El-Adawe.

Word Count: 8616 words

Abstract

Colorism (priviliging of lighter over darker skin) affects development across the lifespan. Empirical evidence of these developmental implications is enhanced with better understanding of color science, including the strengths and limits of past skin color rating scales and directions for future measure development. The current manuscript demonstrates this potential using the L* (lightness) dimension of L*a*b* color space. The degree to which color-swatch based human rating scales approximate interval metrics is examined. Overlaps, reversals, and gaps were evident in two historical scales. More equidistant swatches were revealed in a scale more recently developed based on color science. Results across all scales suggested that Black-identified raters perceived more variation in the skin tones of Black-identified photos than didraters identifying with other race-ethnicities. Yet, this pattern was most consistent for the color-science-developed scale. The extent to which humans perceive the L* (lightness) dimension was validated by a high (0.97) correlation between averages of human ratings on the latter scale and color science calculation of L* for photographed individuals' skin color. Despite this high correlation, level differences were evident. Humans tended to choose darker swatches for darker-skinned and lighter swatches for lighter-skinned individuals than their measured values. Implications for future research are discussed.

Data availability statement: The L* values for skin color swatches are available within the article's Table 1. Although the individual participant data cannot be shared outside the research team, the code will be made available in an OpenICPSR archive.

How Well Do Contemporary and Historical Skin Color Rating Scales Cover the Lightness-to-Darkness Continuum? Descriptive Results from Color Science and Diverse Rating Pools

Recent decades have seen scholars across numerous disciplines better recognize race as a dynamic and multifaceted construct, considering aspects of self-definition and physical appearance in addition to socially constructed racial classification (Roth 2016). This discourse has included reconsiderations of how to best operationalize racialized aspects of physical appearance, including skin color, in social surveys (Campbell et al. 2020; Gordon et al. 2022; Hannon et al. 2021; Hannon and DeFina 2016; Telles 2018). Social and behavioral science research on the relationship between skin color and social outcomes has largely drawn on skin color data that is assessed using categorical scales, which may be rated by respondents, interviewers, or both. These scales have the advantage of being intuitive and inexpensive, making skin color data easy to collect, and thus increasing the number of large social surveys collecting skin color data since the early 2000s.

But the simplicity of skin color rating scales also makes it straightforward to generate new scales, resulting in an increasing number of scales being implemented across varying data collection efforts. Such scales typically range from three to ten categories, and are either labeled with text (e.g. “dark” vs “medium” vs. “light”) or linked to a visual reference image depicting a palette of color swatches (Campbell et al. 2020; Williams 1933). The most common measure of skin color in large U.S.-based social surveys has been the Massey-Martin scale (Massey and Martin 2003), which includes ten color swatches ranging from a very pale, near white skin color to a very dark, near black skin color (see Figure 1). Since it was developed in 2003 for the New Immigrant Study (NIS), the Massey-Martin scale has been fielded in numerous additional

surveys including the National Longitudinal Survey of Youth 1997 (NLSY97), the General Social Survey (GSS), and the Fragile Families and Child Wellbeing Study (FFCW). More recent swatch-based scales, such as that developed for the Project on Ethnicity and Race in Latin America (PERLA) surveys (Telles, Flores, and Urrea-Giraldo 2015), have attempted to better capture variation in skin undertone such as pinkness or yellowness, albeit still as a set of eleven categories ranging lighter to darker (see again Figure 1). Little information is available on the construction of the scales used in major social surveys, e.g., the Massey-Martin scale was reportedly "constructed with assistance from a graphic designer," while the PERLA "came from internet photographs," but the study documentation does not fully describe how the swatches were selected in either case (Massey and Martin 2003; PERLA 2012). Expanding on these earlier traditions, new skin color scales have proliferated across a range of disciplines in recent years (Dadzie et al. 2022; Ostfeld and Yadon 2022). For instance, the Google AI website provides a *Monk Skin Tone (MST) Scale* to support incorporation of skin color into artificial intelligence applications for more inclusive facial recognition (Google AI 2022).

< FIGURE 1 ABOUT HERE>

To better compare, assess, and potentially harmonize results from this increasing array of scales, scholars of human development will benefit from understanding fundamental concepts from color science and how these concepts are represented in existing scales (Branigan et al. 2019; Gullickson 2005; Udry, Bauman, and Chase 1971). Color science is a broad field, spanning disciplines from chemistry to physics and having applications in the visual arts, computer graphics, and manufacturing (Germer, Zwinkels, and Tsai 2014). "Color spaces" quantitatively map how humans perceive color by transforming the measured intensity of different light wavelengths visible to the average human eye into easy-to-understand scales,

supporting applications that readily identify and then reproduce the visual color spectrum as it appears under various conditions such as lighting, angle, and surface type. The first color space was developed by the International Commission on Illumination (CIE) in 1931, and the CIE L*a*b* color space remains a common system for color measurement in the sciences to the present day. The three axes of the L*a*b* color space capture darkness-to-lightness (L*), greenness-to-redness (a*), and blueness-to-yellowness (b*). Darker and lighter human skin colors are distributed across the L* continuum, with undertones of pinkness and yellowness represented across portions of the a* and b* continua. Scientists from L'Oreal used color science to create a Skin Color Chart for use in the cosmetics industry (see again Figure 1; De Rigal et al. 2007) that has since been utilized in other fields as well (e.g. Campbell et al. 2020; Garcia and Abascal 2016).

Our results are relevant to scholars from across disciplines who have studied the consequences of colorism for human development (Burton, Bonilla-Silva, Ray, Buckelew, and Freeman 2010; Crutchfield, Keyes, Williams, and Eugene 2022). Developmental and social psychologists, for instance, have examined when children first attach socially-constructed valences to light and dark skin and how their self-concept and racial identities are shaped by perceptions of their own skin (including using dolls and drawings depicting persons with varying skin tones; Mandalaywala, Ranger-Murdock, Amodio, and Rhodes 2019; Spencer 1984; Swanson, Cunningham, Youngblood, and Spencer 2009). Sociologists, economists, and developmentalists have also documented how darker-skinned adolescents and adults complete less schooling, are less likely to be hired and are paid less, and are more often suspended and incarcerated than their lighter-skinned counterparts (including with measures like those shown in Figure 1 as used in the National Longitudinal Survey of Youth; Abascal and Garcia, 2022;

Branigan, Freese, Sidney, and Kiefe, 2019; Devaraj and Patel 2017; Hannon, DeFina and Bruch 2013). Physical anthropologists and evolutionary biologists have further studied population-level variability in skin pigmentation, and medical professionals match skin color when conducting craniofacial surgeries and consider disease presentation in lighter-to-darker skin (including using color-science instruments; Seelaus, Arias, Morris, & Cohen, 2021; Venkatesh, Maymone, and Vashi 2019). Given that each field has tended to use certain types of methods, the methodological advances we discuss are relevant not only within these disciplines but also to support cross-field fertilization and the synthesis of findings through comparable measurements. These advances connect with overall efforts to attend to the multiple facets and intersections of racialized experience, of which racialized appearance and skin color spectrum is one important part.

In the first part of this manuscript, we demonstrate how various skin color rating scales can be compared, and potentially harmonized, using color science. We feature the L* continuum for simplicity, and return in the discussion to future directions for considering the a* and b* continua. This first section highlights important methodological questions largely untouched in previous literature regarding the construction and comparability of palette-based skin color rating scales and offers novel methods and insight into these questions. In the second part of the manuscript, we illustrate how well the scales capture the distributions of skin colors across diverse sets of photographed individuals based on their original categorical coding and when translated to the L* continuum. Here we illustrate how insights from color science can be informative when integrating disparate findings across studies using different scales. In the third section, we offer validating evidence regarding how averages of human ratings translated to the L* metric reflect color science measurement of L*. These results support the potential for

utilizing the L* metric from swatch-based scales, while also illuminating places in which future studies can consider tendencies for humans to choose swatches lighter or darker than color science assessments. We end by discussing how developmental scientists might leverage color science to synthesize historical studies that have used a range of scales, identify the limitations of such efforts, and point to future directions that can improve on existing scales.

Comparing Skin Color Rating Scales Using Color Science

We begin by comparing the color swatches using L* values extracted using a color science utility programmed into Adobe Photoshop 2020. Doing so allows us to place the swatches from different scales relative to one another. In other words, we can consider questions like: Does one scale start or end at a lighter or darker level than another scale? Doing so also allows us to place the swatches relative to each other within a scale. In other words, we can consider questions like: Are the color science L* values of a scale consistent with the order in which the scale developers placed the swatches?

We focused on the Massey-Martin, PERLA, and L'Oreal scales given their prominence, as noted above, and because their swatches were available in digital images: the digital Massey-Martin scale is part of the documentation for the New Immigrant Study (Massey and Martin 2003); a digital PERLA scale is available on a website hosted by Princeton University (PERLA 2012); and the digital L'Oreal Skin Color Chart is available on the L'Oreal (n.d.) website and documented in de Rigal et al. (2007). The L*a*b* values were read in with the Eyedropper tool in Adobe Photoshop 2020, which enables colors to be read from a single point or averaged over a region of specified dimensions. The PERLA and L'Oreal scales consist of solid rectangular swatches, with no visible color variation within each swatch. As a result, L*a*b* values were read as a 3-by-3 pixel average from the approximate center of the swatch. Supplemental readings

averaging over a larger area within each swatch in the L'Oreal and PERLA scales yielded no difference in the L*a*b* values. In contrast, the swatches of the Massey-Martin scale contain color variation across highlights and shadows, due to the depiction of each color using images of hands. We calculated L*a*b* values as a 51-by-51 pixel average spanning from just above the knuckles of each hand image to just above the wrist.

L* values for the swatches of all three scales are reported in Table 1. We also graphed the L* values for ready visual comparison, shown in Figure 2. We evaluate the relationship between the swatches *within* each scale, asking: a) whether the swatches were monotonically distributed (strictly decreasing in L*), and, b) whether each pair of adjacent swatches was equidistant (as would be expected in an interval metric). We then described the relationships *among* the three scales, including: a) whether the scales encompass similar ranges of darkness-lightness across their swatches, and, b) whether the scales capture similar distances between the swatches.

Figure 2 presents the graph. Each scale is shown in a separate row, labelled on the Y axis. The L'Oreal scale is shown in six rows because its 66 swatches are distributed across six tones of yellowness to pinkness, each of which ranges across 11 levels of darkness. In contrast, the other scales are each arrayed as if along a single continuum, with less systematic variation in pinkness and yellowness. The dots along each row represent the scale's swatches in the L^* metric, as labelled on the X axis. L^* values can range from 0 to 100, with higher values representing lighter shades. Across all three scales, the range of L* was relatively similar. In the L'Oreal scale, the range was 23 to 88; in the Massey-Martin scale, 21 to 86; and in the PERLA scale, 16 to 88.

<TABLE 1 AND FIGURE 2 ABOUT HERE>

As might be expected given the construction of the L'Oreal scale based on color science (de Rigal et al. 2007), the swatches in the L'Oreal scale were strictly monotonic in L* (see again

Table 1), with higher (darker) categories of swatches always having lower (darker) values of L*. The distances between swatches were also relatively constant across the six rows, i.e., the scale's six levels of pinkness-to-yellowness undertones. At the same time, the L'Oreal scale was not strictly interval: spacing (ΔL*, or difference in L*) between the swatches increased slightly from the lighter to darker regions across all rows (ΔL* = -5 or -6 between the first two swatches; ΔL* = -7 or -8 between the last two swatches; see again Table 1).

In contrast, the L* values of the remaining two scales' swatches were sometimes spaced closely together and sometimes far apart, with no consistent trend. In the Massey-Martin scale, ΔL* varied from 0 (between swatches 8 and 9) to -16 (between swatches 1 and 2). In the PERLA scale, ΔL* varied from 1 (between swatches 2 and 3) and 0 (between swatches 10 and 11) to -15 (between swatches 5 and 6). The zero or positive ΔL* between one or two pairs of adjacent swatches in each scale meant neither was strictly monotonic on the L* darkness-lightness continuum. The varying spacing between swatches, including some overlapping values and some large gaps, meant each had considerable departures from an interval scale.

Capturing Distributions of Skin Colors across Diverse Sets of Photographed Individuals

The overlaps/reversals and large gaps between adjacent swatches just noted have implications for how well each scale captures distributions of skin colors across diverse individuals, and thus for uses of the scales to examine substantive questions. Mean differences among groups may be less well detected when there are large gaps in certain regions of darkness-to-lightness. People located in those regions will be placed to the left or to the right of their actual skin shade. Variability may also be over- or under-stated in the presence of such gaps. Finer-grained distinctions cannot be made among those located within a large gap region, meaning variation would be understated especially when samples are concentrated on that

region. On the other hand, variation may be greater when samples spread beyond these gaps, given individuals located in a large gap will be assigned the swatch on one side or the other of the gap. It is also the case that statistics assuming interval metrics like the mean and standard deviation are better suited for the L* metric than the ordinal metric often linked to a scale's ten or eleven categories. Figure 2 showed that the Massey-Martin, and PERLA scales are better treated as ordinal, given the uneven spacing of their swatches, and even ordinality was violated in one or two cases for each scale. Although the L'Oreal was close to interval, the distances between swatches still varied to some degree across levels of the L* darkness-lightness continuum. The implications of such non-interval nature of each scale would be most revealed when a sample has numerous individuals located across the regions with larger and smaller gaps between swatches. Ultimately, the reduced accuracy in estimating means and standard deviations may affect substantive conclusions about group differences in skin color or about how skin color associates with other constructs such as health and well-being.

We explore these implications by considering the long-studied question of whether the identities of the persons doing the ratings and the persons being rated affect selection of skin color categories. Social psychologists and survey methodologists have long traditions of considering these kinds of systematic rater effects, including whether scores differ based on the race-ethnicity of the rater and ratee. One theme considered has been whether individuals see more variation among those who share their social identities than those who do not, including as out-group homogeneity bias. Although some studies have considered similar themes in relation to skin tone—e.g., do Black-identified raters perceive more variation in skin color of Black-identified ratees than do White-identified raters?—it is important to recognize recent critiques as these studies differ from traditional social psychological studies focused on facial recognition

(i.e., the latter considering whether individuals can better recognize the identity of a newly introduced person of one's own race-ethnicity than of other race-ethnicities; Hannon et al. 2021). We consider such themes by drawing on unique data in which we have scores on the original categorical and the L* scales of swatch choices for the three above-discussed rating scales—Massey-Martin, PERLA, and L'Oreal. Our data also has multiple ratings of the same photo sets balanced across four rater and ratee racial-ethnic identities. This design is an advance on prior studies which have more often had single ratings of each person, making it unclear whether rating differences reflect rater tendencies or ratee characteristics.

Specifically, in two substudies from the National Survey Rating Skin Tone (NSRST), U.S residents rated the skin color of individuals depicted in a series of photographs. The two substudies shared several design features, although they differed in which skin color scale was used and which photographs were rated. For both studies, we recruited respondents from the online survey platform Prolific, which caters to scientific researchers (Palan & Schitter 2018; Peer, Brandimarte, Samat, and Acquisti 2017). In each substudy, respondents were stratified with a goal of recruiting equal numbers by their self-identifications in relation to sex (male/female) and race (Latino/a; non-Latino/a Asian, Black, or White). Both NSRST substudies were approved by the Institutional Review Board of the University of Illinois at Chicago (Protocol # 2020-1326). All participants provided informed consent.

In terms of scales and photos, for the first NSRST substudy, 459 Prolific respondents were randomly assigned to use either the Massey-Martin or the PERLA to rate forty stock photos selected from online photograph databases. The photos were selected by an undergraduate research assistant who identified as Latina and searched online databases for headshot-style photos of individuals she believed would commonly be perceived as either male or female and as

either Asian, Black, White, or Latino/a of varying lightness to darkness of skin color. Another undergraduate assistant and two senior members of the research time offered feedback on photo selection until consensus was reached on these attributes.

In second NSRST substudy, another set of Prolific respondents (n = 384) used the L'Oreal scale to code the skin color of one of three randomly assigned sets of 40 photographs. Photographs for this phase were drawn from the Chicago Face Database (CFD: Ma, Correll, and Wittenbrink 2015; Ma, Kantner, and Wittenbrink 2021), a collection of high-resolution standardized headshots of individuals taken for use in scientific research. These photographed individuals were between the ages of 17 and 65, reported their sex (male/female) and race (Asian, Black, White, or Latino/a), and were outfitted in gray t-shirts and asked to display a neutral facial expression. We selected 120 of 597 CFD photos to reflect variations in skin color within sex and race. We first had two or three independent coders extract L*a*b* readings using Adobe Photoshop 2020 from each photo subject's forehead, cheek, and neck, and then averaged. The individual readings were highly correlated among locations (r = .91 to .94 across all 597 photographs). We next stratified all 597 photos by quintiles of skin lightness/darkness (assessed as L* values) and tertiles of skin undertones (pinker, yellower, or neutral, assessed as the ratio of a* to b* values) within sex and race. Where random selection did not yield a visually clear gradient of skin color across lightness-darkness quintiles, the survey team used consensus to replace photos from others within the strata; in total, 21 photos were replaced in this stage.

To illustrate the potential insights offered by considering swatch distributions with color science principles, we focused on the Black-identified photos (10 photos from the stock photo set for the Massey-Martin and PERLA scales; 30 photos from the CFD photo sets for the L'Oreal scale). These Black-identified photos were selected because of prior research themes just

discussed, as well as the CFD photos demonstrating the greatest variation in L* among Black-identified photos (from about L* of 20 to 60) consistent with prior research (Branigan et al., 2013). We further focused on the standard deviations of ratings using the original category values and the color science associated L* values, to consider the noted substantive theme of whether Black-identified raters perceive more variation in these Black-identified photos than do raters identifying with other race-ethnicities. For L'Oreal, rater sample sizes were: White (n = 94), Asian (n = 90), Latino/a (n = 93), and Black (n = 107). Each rater assessed 10 Black-identified CFD photos; these 10-photo-sets were randomly assigned from three tones of pinkness, yellowness, or neutral (30 Black-identified photos total). For PERLA, rater sample sizes were: White (n = 56), Asian (n = 56), Latino/a (n = 54), and Black (n = 64). For Massey-Martin, rater sample sizes were: White (n = 59), Asian (n = 61), Latino/a (n = 59), and Black (n = 50). PERLA and Massey-Martin raters assessed the same 10 stock photos reflecting Black-identified individuals.

Table 2 provides the standard deviations of ratings of the Black-identified photos within four rows representing rater race-ethnicity. The columns provide results for each of the three scales, each in its original 10-or 11-point metric (labelled LD) and the L* metric. Within each scale/metric, significant differences between standard deviations are designated by subscripts based on p-values from Levene's (1960) robust variance test centered on the median. Capital letter subscripts represent $p < .01$. Lowercase letter subscripts represent $p < .05$. When interpreting these findings, it is important to keep in mind that PERLA and Massey-Martin ratings were of one set of photos (Stock Photos) and L'Oreal ratings were of another set of photos (CFD Photos). We thus have confidence in comparisons of results in units of L* to the original metrics within each scale, and comparisons of results between the PERLA and Massey-

Martin scales for each metric. We are cautious regarding comparisons between the L'Oreal results and the PERLA and Massey-Martin results, given that the former and latter scales' ratings are based on different photo sets. There were also more raters using the L'Oreal than the PERLA and Massey-Martin, because raters were randomly assigned to one of the two latter scales in that substudy.

<TABLE 2 ABOUT HERE>

A consistent finding was that Black-identified raters identified more variation across the Black-identified photos than did other raters. This finding was most evident for the L'Oreal and PERLA scales, and clearer in the L* than categorical scale for PERLA and Massey-Martin. Specifically, for L'Oreal, the comparisons had small p-values on both the categorical and *L** metrics. The similar precision in both metrics for the near-interval spaced L'Oreal swatches is reflected in p-values rounding to nearly identical values (p = .003 in both metrics for the comparison of Black-identified and White-identified raters; p = .005 in both metrics for the comparison of Black-identified and Asian-identified raters; p = .002 in the categorical and p = .003 in the L* metric for the comparison of Black-identified and Latino/a-identified raters). These group comparisons also had small p-values for the PERLA scale, although somewhat larger on the categorical than L* scale (p = .008 in the categorical metric and p = .002 in the L* metric for the Black- vs. White-identified rater comparison; p = .012 and p = .001 for the Black- vs. Asian-identified rater comparison; p = .001 and p = .000 for the Black- vs. Latino/a -identified rater comparison). These results suggest some gain in precision of estimation by using the interval L* metric. The results differed for the Massey-Martin scale, where only the comparison of Black with Asian raters had a small p-value and only on the L* scale (p = .007); for the Massey-Martin scale, the comparison of Asian and White raters also had a small p-value,

although moreso on the *L** than categorical scale (p = .047 on the categorical metric and p = .009 on the L* metric). Again, these results suggest greater precision of estimation with the L* metric. The differences between the PERLA and Massey-Martin scale results may also reflect the fact that Massey-Martin's swatches are less evenly distributed in the region of skin color evident among Black-identified photos (L* about 20 to 60).

Returning to the limitation of the use of two different photo sets, we have confidence in the finding that Black-identified raters perceived more variation across the Black-identified photos than did other raters given this finding was robustly seen across two different scales that were used to rate photos in different photo sets (L'Oreal and PERLA). In addition, within the Stock Photo set, PERLA better detected the pattern than did Massey Martin, especially in the L* metric. On the other hand, that PERLA and Massey-Martin ratings were based on one photo set and L'Oreal another photo set limits our confidence in the stronger pattern in units of L* than in the original metric using the PERLA and the Massey-Martin scales versus the pattern being more equally well detected in both the L* and original metric for the L'Oreal scale.

Validation of Average Human Ratings in L* Metric Mirroring Photoshop L*

We end by offering insight into the validity of applying the L* conversion to visual scale ratings, specifically comparing the averaged L* values taken from multiple human ratings of the CFD Photos on the L'Oreal scale to the Photoshop extracted L* values for those same photographs. The use of L*-converted human ratings allows for each measure to be considered on the same interval scale. The results inform us regarding the extent to which, collectively, human raters are perceiving the same value that color science calculations produce. Of course, human ratings vary around this collective average, a point we return to in the discussion (e.g., in our data, the standard deviation of ratings of each photo in the L* metric ranges from about 3 to

about 10 across the 120 photos).

Figure 3 visualizes the association in a scatter plot with the Photoshop-extracted L* values on the y-axis and the average L* values across human ratings on the x-axis. We use marker colors to represent the photographed individuals' racial-ethnic identifications: Black-filled markers = Black-identified photos. Black-outlined markers = White-identified photos. Red-outlined markers = Latino/a-identified photos. Blue-outlined markers = Asian-identified photos. The red line reflects a best-fitting regression line. The black line is an identity line, reflecting the same values of L* on the Y and X axes.

<FIGURE 3 ABOUT HERE>

One general conclusion is that the resulting scatterplot visualizes a strong linear association. The markers are clustered around the red line, which has a standardized slope (correlation) of 0.97. The corresponding R^2 value of 0.94 means that the averaged human ratings and Photoshop readings of L* share 94% variation. At the same time, the unstandardized slope is not equal to one, but rather 0.61, reflecting systematic tendencies for averaged human ratings to fall above the black line at lower values of L* (darker skin color) and below the black line at higher values of L* (lighter skin color). This result suggests a tendency for darker-skinned photographed individuals to be rated darker in comparison to the values extracted directly from photographs using the color science formulas built into Photoshop, and the inverse to be evident for lighter-skinned photographed persons. The marker colors also demonstrate that it is largely persons who identify as Black whose skin color tends to be rated as darker than the Photoshop readings (black-filled markers), and those who identify as Asian, Latino/a, or White whose skin colors tend to be rated lighter than the Photoshop readings (blue-, red-, and black-outlined markers). One possibility this result suggests is that human raters' swatch choices may be

affected by additional aspects of physical appearance that signal race-ethnicity. Also evident in the graph is the greatest span of L* covered by the skin tones of persons identified as Black (about 35 points on the L* metric), the smallest span for those identified as White (about 10 points), and mid-range span for those identified as Asian and Latino/a (about 20 points). Circling back to the first set of findings above, the varying sizes of between-swatch gaps may have varying implications for those identifying with different race-ethnicities (e.g., the single point on the Massey-Martin scale above about 70 means the scale may not well differentiate among the skin tones of individuals identified as White).

DISCUSSION

Together, our three sets of results demonstrate the potential for developmental scientists from a range of disciplines to incorporate the principles of color science into their skin color research. By extracting color science L* (lightness) values from the swatches often used in human skin color rating scales, our first set of results demonstrated how we can relate the scales to one another in terms of their coverage of the darkness-lightness continuum and see how well each approximated a monotonic interval metric. Our second set of results then illustrated the implications of using the interval L* metric versus ordinal values attached to categories for identifying evidence that informs substantive questions. Our results reflected the ways in which precision can be increased when human rating scales approximate an interval scale. Finally, our third set of results validated that when we average across human ratings of photographs in the L* scale the values are highly linearly related to color science calculated L* values from those same photographs. At the same time, the results also suggested important directions for future research into the ways humans may systematically rate photographed skin darker or lighter than Photoshop readings and the ways scale swatches may better differentiate skin colors in some

regions of darkness-lightness than others.

There are several possible impacts of the gaps, reversals, and overlap in L* value identified in part one of this paper for past and future colorism research. Extracting L* values from existing scales' swatches can help scholars choose among existing scales. Doing so can also help scholars interrogate inconsistent findings across existing studies using various scales. Scholars can also leverage color science when considering arrays of possibilities for swatches to be included in new scales. Potentially, this might include tailoring sets of swatches to study samples and research objectives, such as using an approach akin to computerized adaptive testing by offering raters a series of finer grained choices after they make an initial selection.

In terms of the three scales considered here, our results highlighted the ways in which their ten or eleven swatches (within rows [undertones] for L'Oreal) represented to a greater and lesser degree different regions of the darkness-to-lightness continuum. The variations have implications to be considered when using and interpreting each scale. Where the scale had a large gap between adjacent swatches, as in the mid-to-upper regions of the Massey-Martin and PERLA scales, the skin lightness of people located in these gaps would have been less well differentiated than by a scale like L'Oreal, which had swatches more evenly distributed. Where a scale had swatches overlapping or reversed, the scale is inefficient in its coverage of the darkness-to-lightness continuum—moving one of these swatches to a shade that falls within a larger gap would better cover the L* continuum. It may also be possible that, when confronted with a color palette with poor coverage of skin color, gaps between swatches (in the metric of L*), or swatch ordering that feels counterintuitive, interviewers and raters may discount the actual color of swatches and interpret the scale meaning more subjectively, potentially complicating the utility of recoding swatches based on *L** values. In either case, these situations

put raters into a mentally-taxing position.

The implications of the non-interval nature of existing human rating scales was illustrated in the second part of this paper. We demonstrated the ways in which Black-identified raters' perception of greater variation in Black-identified photos was better captured in scales with better coverage of their skin lightness-darkness. Similar implications may be evident in other studies using human rating scales to answer substantive questions about skin color—in other words, associations may be amplified or muted in prior studies that used the rating scales just demonstrated to have varying inter-swatch distances. Whereas we focused our illustration on standard deviations, it is also the case that means may be under- or over-stated due to gaps, reversals, or overlaps, as discussed above. Altogether, the non-interval nature of many existing human skin color rating scales means correlations with other variables and mean group differences may be mis-estimated in studies using these scales.

Finally, the third part of our study emphasized the potential of using color science in future multidisciplinary developmental studies of the ways in which skin color affects health and well-being. On the one hand, averaging across human ratings of photographs translated to the L* (lightness) continuum produced scores that were highly correlated with color science calculated L* values of those same photographs. On the other hand, human rating averages appeared to systematically overstate the darkness of skin color for Black-identified individuals and to systematically overstate the lightness of skin color for those identified with Asian, Latino/a, and White race-ethnicities. The first result suggests either approach could be used if correlations with other variables of interest, although relying on L* would avoid the need to average across many human raters. The second result suggests the approach would produce different findings if the focus was on the actual skin color, as measured or perceived. Future studies can leverage color

science to probe such results. For instance, a set of photos in the range where we see the greatest overlap across racial-ethnic identities (around L* of 60), might be systematically varied across a series of equivalent shades ranging from darker or lighter. Using mixed methods might allow raters to talk through their choices of swatches for these photos. When choosing among scales in studies, scholars may also want to consider how their substantive question may be informed by: a) averages of human rated values based on photographs of study participants (which could better reflect how others perceive an individual), and, b) color science calculated value of the same photos, and c) the difference between the two (which could reflect an extent of social bias).

There are limitations to the current study, and ample avenues for future research. We focused on the L* (darkness-lightness) continuum. Skin undertones of pinkness (a*) and yellowness (b*) are important avenues of future research, especially given the limited prior research investigating how skin lightness/darkness and undertones independently relate to social outcomes has affirmed the salience of skin lightness/darkness for social experience, while the social meaningfulness of skin undertones remains less clear (Gordon et al. 2022). Additionally, our requirement that scales have a set digital palette graphic precluded inclusion of other common scales such as the Fitzpatrick scale, used in medical research. Although this scale is often depicted visually, it lacks a standardized set of reference images, and the colors of swatches in its graphic depictions can differ widely. We also used different photo sets with the L'Oreal versus the Massey-Martin and PERLA scales, and future studies that probe the implications of swatch overlaps/reversals and gaps might use the same photo set across all scales. Collection of diverse photo sets worldwide might also inform the extent to which scales are covering the full range of human skin color (or plausibly going beyond the range of realistic human skin color). Finally, we would re-emphasize that human ratings vary, and studies that rely

on a single rating would not be expected to achieve as high a correlation with Photoshop L* as seen in our third set of results which relied on averages of human ratings. We also asked raters to assess photographs onscreen, whereas other studies collect ratings in the field under varying lighting conditions.

With these limits in mind, we encourage future studies to use color science and conversion of swatches into L* values to build on our study, including to identify gaps, reversals, and overlaps in color scales, and to consider whether harmonization on the L* scale reveals meaningful insights into group mean differences or associations with health and well-being. Cross-study comparisons might be facilitated by conversion of swatches into L* prior to research, and color science could be used to create new scales. This avenue of research can also be leveraged to explore the extent to which raters notice reversals, gaps, or overlap, as well as how raters navigate such situations and the impact, if any, this has on outcomes being studied.

For scholars of human development, for instance, these methods might be applied across the lifespan. Color-science readings of L* (from photographs, or taken in person with small, unobtrusive handheld devices; Gordon et al., 2022) might be ideal for capturing longitudinal changes in skin coloration and its consequences, given the high accuracy of single readings. Taking readings in multiple locations can capture variations across sun-exposed and sun-protected regions (including across seasons) and across self-presentation of visible skin (including use of tanning or lightening cosmetics). For instance, existing dermatologic studies suggest that skin darkens, reddens, and yellows with age, although these findings are from comparisons across age cohorts in convenience samples (e.g., Dobos et al., 2015; Kelly et al., 1995; Trojahn et al., 2015). Human ratings may be added when the focus is understanding self- and other-perceptions of skin color, and their relationships with the development or perception of

social identities such as race and gender as related to the lightness-to-darkness shades of skin (and their yellowness-to-pinkness tones). We encourage scholars to take up the exciting potential of bridging fields, including contemporary studies building on the historical studies reviewed above on early to middle childhood that tended to rely upon visual representations (dolls, drawings) and the studies of adolescence and adulthood that tended to rely upon swatch presentations (as in the scales shown in Figure 1). Cross-fertilization is already evident as some recent developmental research has relied upon historical swatch-based scales (e.g., Massey-Martin, Kim and Calzada 2018) and photo-manipulation and swatches informed by color-science (Dunham, Dotsch, Clark, and Stepanova, 2016), whereas others have relied upon words to describe skin tone (very light to very dark; e.g., Adams, Kurtz-Costes, Hoffman, Volpe, and Rowley, 2020; Blake, Keith, Luo, Le, and Salter 2017; Kiang, Espino-Pérez, and Stein, 2020; Landor and Halpern 2016; Uzogara, Lee, Abdou, and Jackson, 2013). Mixed methods studies that draw upon color-science based instruments, visual representations, and swatch presentations might combine quantitative with qualitative insights into stabilities and fluidities of skin color and associated identities and inequalities.

REFERENCES

Abascal, Maria and Denia Garcia. 2022. "Pathways of Skin Color Stratification: The Role of Inherited (Dis)Advantage and Skin Color Discrimination in Labor Markets." *Sociological Science* 9: 346-373. DOI: 10.15195/v9.a14

Adams, Elizabeth A., Beth Kurtz-Costes, Adam J. Hoffman, Vanessa V. Volpe, and Stephanie J. Rowley. 2020. "Longitudinal Relations Between Skin Tone and Self-Esteem in African American Girls." *Developmental Psychology* 56: 2322-2330. doi: 10.1037/dev0001123

Blake, Jamilia J., Verna M. Keith, Wen Luo, Huong Le, and Phia Salter. 2017. "The Role of

Colorism in Explaining African American Female's Suspension Risk." *School Psychology Quarterly* 32: 118-130. doi: 10.1037/spq0000173

Branigan, Amelia R., Jeremy Freese, Assaf Patir, Thomas W. McDade, Kiang Liu, and Catarina I. Kiefe. 2013. "Skin Color, Sex, and Educational Attainment in the Post-Civil Rights Era." *Social Science Research* 42(6):1659–74.

Branigan, Amelia R., Jeremy Freese, Stephen Sidney, and Catarina I. Kiefe. 2019. "The Shifting Salience of Skin Color for Educational Attainment." *Socius* 5:2378023119889829. doi: 10.1177/2378023119889829.

Burton, Linda M., Eduardo Bonilla-Silva, Victor Ray, Rose Buckelew, and Elizabeth Hordge Freeman. 2010. "Race Theories, Colorism, and the Decade's Research on Families of Color." *Journal of Marriage and Family* 72: 440-459. doi:10.1111/j.1741-3737.2010.00712.x

Campbell, Mary E., Verna M. Keith, Vanessa Gonlin, and Adrienne R. Carter-Sowell. 2020. "Is a Picture Worth A Thousand Words? An Experiment Comparing Observer-Based Skin Tone Measures." *Race and Social Problems* 12(3):266–78. doi: 10.1007/s12552-020-09294-0.

Crutchfield, Jandel, Latocia Keyes, Maya Williams, and Danielle R. Eugene. 2022. "A Scoping Review of Colorism in Schools: Academic, Social, and Emotional Experiences of Students of Color." *Social Sciences* 11:15. doi: 10.3390/socsci11010015

Dadzie, Ophelia E., Rick A. Sturm, Damilola Fajuyigbe, Antoine Petit, and Nina G. Jablonski. 2022. "The Eumelanin Human Skin Colour Scale: A Proof-of-concept Study." *British Journal of Dermatology* 187(1):99–104. doi: 10.1111/bjd.21277.

Darity Jr., William A., Jason Dietrich, and Darrick Hamilton. 2005. "Bleach in the Rainbow:

Latin Ethnicity and Preference for Whiteness." *Transforming Anthropology* 13(2):103–9.

de Rigal, Jean, Marie-Laurence Abella, Franck Giron, Laurence Caisey, and Marc André Lefebvre. 2007. "Development and Validation of a New Skin Color Chart." *Skin Research and Technology: Official Journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)* 13(1):101–9. doi: 10.1111/j.1600-0846.2007.00223.x.

de Rigal, Jean, Marie-Laurence Abella, Franck Giron, Laurence Caisey, and Marc André Lefebvre. 2007. "Development and Validation of a New Skin Color Chart®." *Skin Research and Technology* 13(1):101–9. doi: 10.1111/j.1600-0846.2007.00223.x.

Devaraj, Srikant and Pankaj C. Patel. 2017. "Skin Tone and Self-Employment: Is there an Intra-Group Variation among Blacks?" *The Review of Black Political Economy* 44: 137-166. doi: 10.1007/s12114-017-9249-x

Dixon, Angela R., and Edward E. Telles. 2017. "Skin Color and Colorism: Global Research, Concepts, and Measurement." *Annual Review of Sociology* 43(1):405–24. doi: 10.1146/annurev-soc-060116-053315.

Dobos, G., C. Trojahn, A. Lichterfeld, B. D'Alessandro, S. V. Patwardhan, D. Canfeld, U. Blume-Peytavi, and J. Kottner. 2015. "Quantifying Dyspigmentation in Facial Skin Ageing: An Explorative Study." *International Journal of Cosmetic Science* 37: 542-549. doi: 10.1111/ics.12233

Dunham, Yarrow, Ron Dotsch, Amelia R. Clark, and Elena V. Stepanova. 2016. "The Development of White-Asian Categorization: Contributions from Skin Color and Other Physiognomic Cues." *PLoS ONE* 11: e0158211. doi:10.1371/journal.pone.0158211

Garcia, Denia, and Maria Abascal. 2016. “Colored Perceptions: Racially Distinctive Names and Assessments of Skin Color.” *American Behavioral Scientist* 60(4):420–41. doi: 10.1177/0002764215613395.

Germer, Thomas A., Joanne C. Zwinkels, and Benjamin K. Tsai. 2014. *Spectrophotometry: Accurate Measurement of Optical Properties of Materials*. Amsterdam: Elsevier Science.

Google AI. 2022. “Developing the Monk Skin Tone Scale.” Retrieved October 20, 2022 (https://skintone.google/the-scale).

Gordon, Rachel A., Amelia R. Branigan, Mariya Adnan Khan, and Johanna G. Nunez. 2022. “Measuring Skin Color: Consistency, Comparability, and Meaningfulness of Rating Scale Scores and Handheld Device Readings.” *Journal of Survey Statistics and Methodology* smab046. doi: 10.1093/jssam/smab046.

Gullickson, Aaron. 2005. “The Significance of Color Declines: A Re-Analysis of Skin Tone Differentials in Post-Civil Rights America.” *Social Forces* 84(1):157–80.

Hannon, Lance, and Robert DeFina. 2016. “Reliability Concerns in Measuring Respondent Skin Tone by Interviewer Observation.” *Public Opinion Quarterly* 80(2):534–41. doi: 10.1093/poq/nfw015.

Hannon, Lance, Robert DeFina, and Sarah Bruch. 2013. "The Relationship Between Skin Tone and School Suspension for African Americans." *Race and Social Problems* 5: 281-295. doi: 10.1007/s12552-013-9104-z

Hannon, Lance, Verna M. Keith, Robert DeFina, and Mary E. Campbell. 2021. “Do White People See Variation in Black Skin Tones? Reexamining a Purported Outgroup Homogeneity Effect.” *Social Psychology Quarterly* 84(1):95–106. doi: 10.1177/0190272520961408.

Kelly, Robert I., Rupert Pearse, Richard H. Bull, Jean-Luc Leveque, Jean de Rigal, and Peter S. Mortimer. 1995. "The Effects of Aging on the Cutaneous Microvasculature." *Journal of the American Academy of Dermatology* 33: 749-756. doi: 10.1016/0190-9622(95)91812-4

Kiang, Lisa, Kathy Espino-Pérez, and Gabriela L. Stein. 2020. "Discrimination, Skin Color Satisfaction, and Adjustment among Latinx American Youth." *Journal of Youth and Adolescence* 49: 2047-2059. doi: 10.1007/s10964-020-01244-8

Kim, Yeonwoo and Esther J. Calzada. 2018. "Skin Color and Academic Achievement in Young, Latino Children: Impacts Across Gender and Ethnic Group." *Cultural Diversity and Ethnic Minority Psychology* 25: 220-231.

Landor, Antoinette M. and Carolyn Tucker Halpern. 2016. "The Enduring Significance of Skin Tone: Linking Skin Tone, Attitudes Toward Marriage and Cohabitation, and Sexual Behavior." *Journal of Youth and Adolescence* 45: 986-1002. doi: 10.1007/s10964-016-0456-8

Levene, H. 1960. "Robust tests for equality of variances." In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ingram Olkin, Sudhist G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, (pages 278-292). Menlo Park, CA: Stanford University Press.

L'Oréal Groupe. n.d. "Expert in Skin and Hair Types Around the World." https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

Ma, Debbie S., Joshua Correll, and Bernd Wittenbrink. 2015. “The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data.” *Behavior Research Methods* 47(4):1122–35. doi: 10.3758/s13428-014-0532-5.

Ma, Debbie S., Justin Kantner, and Bernd Wittenbrink. 2021. “Chicago Face Database: Multiracial Expansion.” *Behavior Research Methods* 53(3):1289–1300. doi: 10.3758/s13428-020-01482-5.

Mandalaywala, Tara M., Gabrielle Ranger-Murdock, David M. Amodlo, and Marjorie Rhodes. 2019. "The Nature and Consequences of Essentialist Beliefs about Race in Early Childhood." *Child Development* 90: e437-e453. https://doi.org/10.1111/cdev.13008

Massey, Douglas S., and Jennifer A. Martin. 2003. *The NIS Skin Color Scale*. https://nis.princeton.edu/downloads/nis-skin-color-scale.pdf

Ostfeld, Mara C., and Nicole D. Yadon. 2022. “¿Mejorando La Raza?: The Political Undertones of Latinos’ Skin Color in the United States.” *Social Forces* 100(4):1806–32. doi: 10.1093/sf/soab060.

Palan, Stefan, and Christian Schitter. 2018. “Prolific.ac—A Subject Pool for Online Experiments,” *Journal of Behavioral and Experimental Finance* 17: 22–27.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. “Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research,” *Journal of Experimental Social Psychology* 70: 153–163.

PERLA. 2012. “PERLA Color Palette.” *PERLA Color Palette*. Retrieved October 27, 2022 (https://perla.soc.ucsb.edu/data/color-palette).

Roth, Wendy D. 2016. “The Multiple Dimensions of Race.” *Ethnic and Racial Studies* 39(8):1310–38. doi: 10.1080/01419870.2016.1140793.

Seelaus, Rosemary, Eduardo Arias, David Morris, and Mimis Cohen. 2021. "State of the Art in Computer-assisted Facial Prosthetic Rehabilitation." *The Journal of Craniofacial Surgery* 32: 1255-1263. doi: 10.1097/SCS.0000000000007530

Spencer, Margaret Beale. 1984. "Black Children's Race Awareness, Racial Attitudes, and Self-Concept: A Reinterpretation." *Journal of Child Psychology and Psychiatry* 25: 433-441. doi: 10.1111/j.1469-7610.1984.tb00162.x

Swanson, Dena Phillips, Michael Cunningham, Joseph Youngblood, and Margaret Beale Spencer. 2009. "Racial Identity Development During Childhood." In Neville, Helen A., Brendesha M. Tynes, and Shawn O. Utsey (Eds). Handbook of African American Psychology (Chapter 20, pp. 269-281). Sage Publications.

Telles, Edward. 2018. "Latinos, Race, and the U.S. Census." *The ANNALS of the American Academy of Political and Social Science* 677(1):153–64. doi: 10.1177/0002716218766463.

Telles, Edward, René D. Flores, and Fernando Urrea-Giraldo. 2015. "Pigmentocracies: Educational Inequality, Skin Color and Census Ethnoracial Identification in Eight Latin American Countries." *Research in Social Stratification and Mobility* 40:39–58. doi: 10.1016/j.rssm.2015.02.002.

Trojahn, Carina, Gabor Dobos, Andrea Lichterfeld, Ulrike Blume-Peytavi, and Jan Kottner. 2015. "Characterizing Facial Skin Ageing in Humans: Disentangling Extrinsic from Intrinsic Biological Phenomena." *BioMed Research International* 318586. http://dx.doi.org/10.1155/2015/318586

Udry, J. Richard, Karl E. Bauman, and Charles Chase. 1971. "Skin Color, Status, and Mate Selection." *The American Journal of Sociology* 76(4):722–33.

Uzogara, Ekeoma E., Hedwig Lee, Cleopatra M. Abdou, and James S. Jackson. 2013. "Comparison of Skin Tone Discrimination among African American Men: 1995 to 2003." *Psychology of Men and Masculinity* 15: 201-212. doi: 10.1037/a0033479

Venkatesh, Samantha, Mayra B. C. Maymone, and Neelam A. Vashi. 2019. "Aging in Skin of Color." *Clinics in Dermatology* 37: 351-357. https://doi.org/10.1016/j.clindermatol.2019.04.010

Williams, George Dee. 1933. “The Measurement of Skin Color.” *Science* 78(2018):192–93.

Table 1

Photoshop L Values by Skin Color Rating Scale*

	L'Oreal row							
Swatch	pinkest	pinker	pink	yellow	yellower	yellowest	Massey-Martin	PERLA
1 (lightest)	87	87	88	88	88	87	86	88
2	82	82	82	83	83	82	70	**80**
3	76	76	76	77	77	76	62	**81**
4	70	70	70	70	70	70	54	74
5	64	64	64	64	64	65	47	65
6	58	58	58	58	58	58	42	50
7	51	51	51	51	51	51	33	41
8	45	45	45	44	44	44	**30**	32
9	37	37	38	38	38	38	**30**	23
10	31	31	31	30	30	31	21	**16**
11 (darkest)	23	23	23	23	23	23		**16**

Note. Higher L* values reflect lighter colors. Bolded values are increasing (darker) scale categories but have equal or increasing (lighter) L* values. L readings were based on Adobe Photoshop 2020; repeated readings were stable to minor variation in sample center within +/-1 unit.

Table 2

Standard Deviations of Swatch Choices for Black-Identified Photos in LD (Categorical) and L (Interval) Metrics, by Rater Race-Ethnicity*

	L'Oreal (CFD Photos)		PERLA (Stock Photos)		Massey-Martin (Stock Photos)	
	Standard Deviation		Standard Deviation		Standard Deviation	
	LD	L*	LD	L*	LD	L*
Rater Race-Ethnicity						
White	1.71_{A}	11.64_{A}	2.36_{A}	21.73_{A}	2.37_{a}	14.37_{A}
Asian	1.71_{B}	11.68_{B}	2.29_{b}	21.11_{B}	2.24_{a}	13.37_{AB}
Latino/a	1.67_{C}	11.42_{C}	2.24_{C}	20.57_{C}	2.28	14.33
Black	1.95_{ABC}	13.18_{ABC}	2.53_{AbC}	23.19_{ABC}	2.32	15.01_{B}

Note. Values are standard deviations of ratings based on the categorical swatch choices (LD) and the swatches' associated L* values. Subscripts indicate values that differ at either $p < .01$ (upper-case) or $p < .05$ (lower-case) based on Levene's (1960) robust test of variances. For L'Oreal, rater sample sizes were: White ($n = 94$), Asian ($n = 90$), Latino/a ($n = 93$), and Black ($n = 107$). Each rater assessed 10 Black-identified CFD photos; these 10-photo-sets were randomly assigned from three tones of pinkness, yellowness, or neutral (30 Black-identified photos total). For PERLA, rater sample sizes were: White ($n = 56$), Asian ($n = 56$), Latino/a ($n = 54$), and Black ($n = 64$). For Massey-Martin, rater sample sizes were: White ($n = 59$), Asian ($n = 61$), Latino/a ($n = 59$), and Black ($n = 50$). PERLA and Massey-Martin raters assessed the same 10 stock photos reflecting Black-identified individuals.

Figure 1

The L'Oreal, PERLA, and Massey-Martin Skin Color Rating Scales

Sources.

L'Oreal: https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

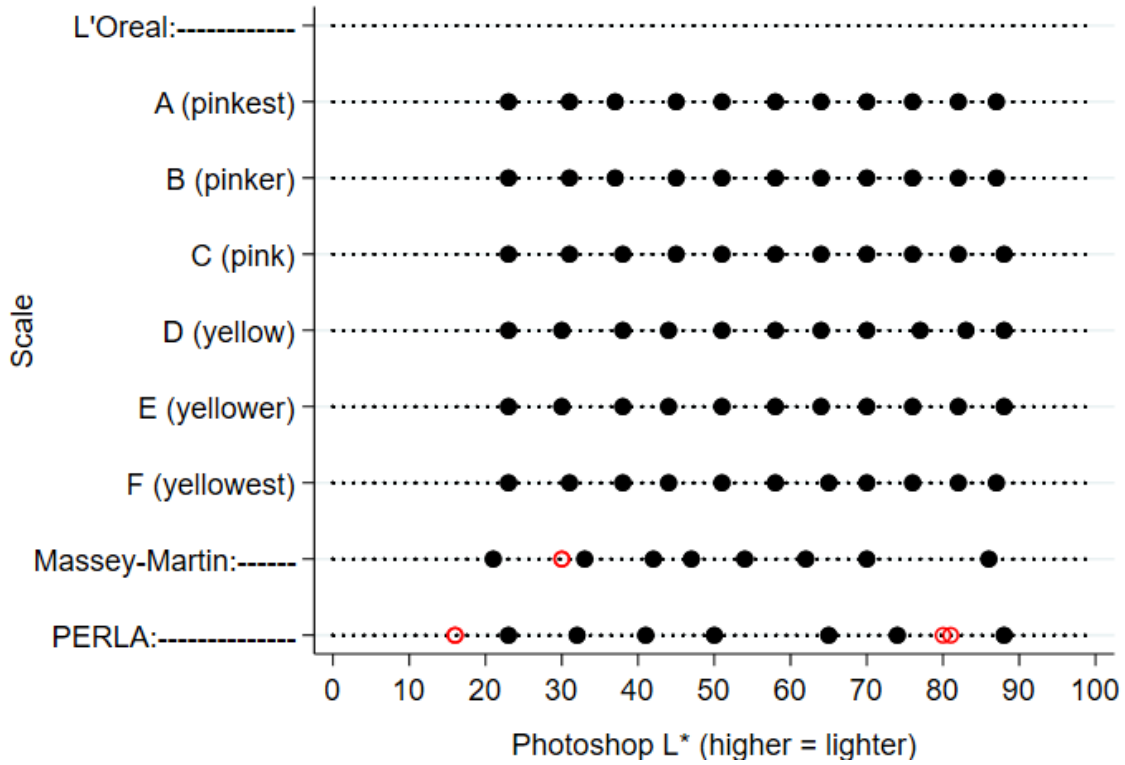
PERLA: https://perla.soc.ucsb.edu/data/color-palette

Massey-Martin: https://nis.princeton.edu/downloads/nis-skin-color-scale.pdf

Note. The L'Oreal scale is inverted to show lightness-to-darkness from top to bottom rather than left to right, to correspond with other scales. Like other recent uses of Massey-Martin, we removed the cuff-link and hand, focusing on the color.

Figure 2

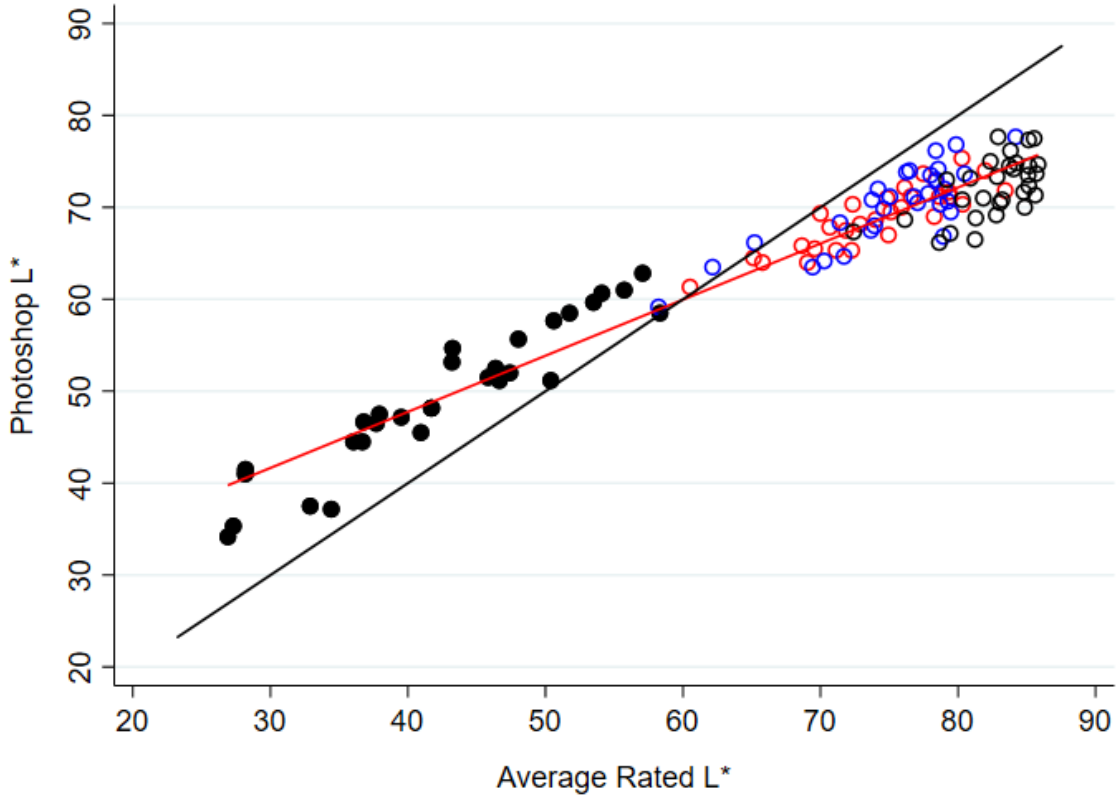
Dot Plots of Photoshop L Values for Swatches from Three Skin Color Rating Scales*



Note. Each marker represents a swatch. Hollow red markers reflect swatches that scale developers placed at higher (darker) categories but that have equal or higher (lighter) L* values.

Figure 3

Scatterplot of Photoshop L and Average Rater L* Values for L'Oreal Scale Among the 120 Chicago Face Database Photos*



Note. Higher L* values reflect lighter skin shades. Marker colors represent the photographed individuals' racial-ethnic identifications: Black-filled markers = Black-identified photos. Black-outlined markers = White-identified photos. Red-outlined markers = Latino/a-identified photos. Blue-outlined markers = Asian-identified photos. The red line reflects a best-fitting regression line ($b = 0.61$; $\beta = 0.97$, $R^2 = 0.94$). The black line is an identity line, reflecting the same values of L* on the Y and X axes.