SHORT ARTICLE

Measuring Perceived Skin Color: Spillover Effects and Likert-Type Scales

Marisa Abrajano, University of California, San Diego
Christopher S. Elmendorf, University of California, Davis
Kevin M. Quinn, University of Michigan

Discrimination based on skin color has been documented as a considerable problem in social science research. Most of this research relies on Likert-type ratings of skin color such as the Massey-Martin Scale (MMS). Scholars have raised questions about measurement error in such scales. We hypothesize that the coding of a person's skin color will vary depending on the race of persons previously coded. We find that the MMS is vulnerable to spillover effects: a person's skin is coded as darker, on average, if he is observed following a sequence of White persons than if he is observed following a sequence of Black persons. We also replicate previous work showing that Black and White coders use the scale differently. Finally, having coders cross-reference the palette at the time of coding, rather than recalling the palette from memory, fails to mitigate either race-of-coder or spillover effects.

A considerable body of evidence indicates that discrimination among Black Americans and Latinos on the basis of phenotype gradation, known as colorism, is a large and persistent social problem. Darker-skinned Black Americans, and darker Mexican and Cuban Americans, face greater discrimination in the labor market compared to their lighter-skinned counterparts (Espino and Franz 2002; Hersch 2008; Kreisman and Rangel 2015; Wade, Romano, and Blue 2004). Darker-skinned Black Americans are also worse off on a variety of socioeconomic and medical outcomes compared to their lighter-skinned counterparts (Hochschild and Weaver 2007; Klonoff and Landrine 2000). Additionally, several observational and experimental studies suggest that Black Americans with a more stereotypically Black appearance receive harsher criminal sentences (Blair, Judd, and Chapleau 2004; Burch 2015; Eberhardt et al. 2004, 2006). Even charitable giving and public support for natural-disaster relief seems contingent on the skin color of perceived beneficiaries (Iyengar and Hahn 2007; Jenq, Pan, and Theseira 2015). Studies conducted outside of the United States, particularly in Latin America, also identify colorism as a pressing social problem (Canache et al. 2014; Telles 2004; Telles, Flores, and Urrea-Giraldo 2015; Villareal 2010).

Over the past decade or so, political scientists have increasingly recognized colorism's influence on a vast array of political outcomes, ranging from assessments of candidates and political elites (Burge, Wamble, and Cuomo 2020; Hochschild and Weaver 2007; Weaver 2012) to feelings of political efficacy and desire to participate in politics (Canache et al. 2014; Garcia Bedolla 2005; Wilkinson and Earle 2012). Darker-skinned candidates face an electoral penalty when evaluated by White voters (Hunter 2005; Weaver 2012). Conversely, recent work by Burge et al. (2020) and Lerman, McCabe, and Sadin (2015) finds that Black voters prefer dark-skinned Black

Marisa Abrajano (mabrajano@ucsd.edu) is a professor of political science and provost of Earl Warren College at the University of California, San Diego, San Diego, CA 92093. Christopher S. Elmendorf (cselmendorf@ucdavis.edu) is the Martin Luther King Jr. Professor of Law at the University of California, Davis, Davis, CA 95616. Kevin M. Quinn (kmq@umich.edu) is a professor of political science at the University of Michigan, Ann Arbor, MI 48103.

We gratefully acknowledge support from the National Science Foundation through grant SES 16-59922. Replication files are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the *JOP* replication analyst. An appendix with supplementary material is available at https://doi.org/10.1086/720941.

Published online October 27, 2022.

candidates over lighter-skinned Black candidates. Black criminal defendants have been shown to face color-based discrimination in criminal sentencing (Burch 2015), which in turn affects their opportunities for political participation.

Given the growth in studies focusing on skin color and politics, which in large part is driven by the need to capture race as a social construction (Omi and Winant 1986), it is critical to assess the accuracy of existing metrics. To the extent that the measures are noisy, this is likely to bias toward zero the estimated effect of skin color on any outcome variable of interest; to the extent that the measures embody a more systematic source of error, this can bias results in other ways (Meijer, Oczkowski, and Wansbeek 2021).

The prevailing measure of skin color among survey researchers is a Likert-type scale. Often an interviewer or coder is asked to rate the subject's skin color against a palette (e.g., fig. 1). In some cases, survey respondents are asked to self-assess their color.[1] There is good reason to believe, however, that such Likert-type metrics suffer from considerable measurement error. Previous studies have found that one of the leading Likert-type skin color scales, the Massey-Martin Scale (MMS), is applied differently by White and Black coders, with White coders rating Black subjects darker in color than Black coders do (Hannon and DeFina 2014; Hill 2002).[2] In light of this and other issues, Hannon and DeFina (2016, 540) recommend "having interviewers (who are asked to code the respondent's skin color) reference a simplified color chart during the interview."

We also suspect that the skin-color rating assigned to a given person may vary depending on the race of persons previously coded. A coder who has just rated a sequence of Black persons may become more attentive to differences among Blacks. Conversely, after evaluating a sequence of Whites, a coder who encounters a Black person may see that

Figure 1. Massey-Martin Scale (palette)

person as "dark" by contrast and assign a darker rating than would otherwise be the case.

If such "spillover affects" occur in the coding of skin color, they could systematically bias Likert-type ratings obtained through surveys that rely on cluster sampling, such as the American National Election Study. Under cluster sampling, the geographic concentration of racial/ethnic groups is likely to result in Black Americans in the sample being more likely to be interviewed immediately after another Black respondent than immediately after a White respondent. If spillover effects cause Black Americans to be coded as darker when they are observed following Whites than when they are observed following other Black Americans, then spillover effects may systematically bias toward "dark" the coded skin color of Blacks who live in predominantly White neighborhoods and toward "light" the coded skin color of Black Americans who live in predominantly Black neighborhoods. This would confound any research design that seeks to understand how neighborhood conditions affect socioeconomic and political outcomes for persons with a given skin color.

In light of these considerations, we undertake to test three conjectures:

Spillover Effects. Application of Likert-type skin color scales is affected by the race of persons whom the coder previously coded, such that persons will be rated as lighter in skin color if they are observed following a sequence of Black Americans than if they are observed following a sequence of Whites.

Race-of-Coder Effects. Relative to Black coders, White coders rate Black Americans as darker skinned (as previously found by Hannon and DeFina [2014] and Hill [2002]).

Attenuation with Palette. Providing coders with a skin-color palette alongside the images to be coded (as recommended by Hannon and DeFina [2016]) attenuates spillover and race-of-coder effects, relative to the baseline condition in which the coder is asked to memorize

1. We conducted a literature review of political science studies focusing on skin color over the past two decades and found that 75% of these articles use a Likert-type scale to measure skin color. See the appendix for details.

2. Other problems include that the color palette does not capture much of the variation in skin tone among Whites (Branigan et al. 2013, 1659), and the numbered shades on the palette are not equidistant per objective measures of reflectivity (Hannon and DeFina 2016). Moreover, interviewer codings of skin tone using the standard protocol are not reliable. In the General Social Survey panel, the intraclass correlation for skin tone between 2012 and 2014 was 0.451 for Black Americans and 0.079 for Latinos (Hannon and DeFina 2016). Among respondents known to have had different interviewers in each year, the correlations were lower yet: 0.279 for Blacks and 0.003 for Latinos (Hannon and DeFina 2016). "Less than a quarter of the Black respondents listed in the top three categories of skin darkness in 2012 fell into these same top three categories in 2014" (535).

the scale before observing the sequence of persons to be coded.

We test these conjectures by analyzing the "consensus" measure of skin color used in contemporary survey research, the MMS (Hannon and DeFina 2016, 535).[3] Initially developed to characterize the color spectrum among college freshmen (Massey et al. 2003) and immigrants (Massey and Martin 2003), the MMS was quickly adopted in a variety of applications by sociologists, economists, political scientists, and law professors (Dávila, Mora, and Stockly 2011; Frank, Akresh, and Lu 2010; Hannon 2015; Herman 2011; Hersch 2008, 2009, 2011; Jenq et al. 2015; Kreisman and Rangel 2015; Ostfeld and Yadon 2020). Both the 2010–14 General Social Survey panel (Hannon and DeFina 2016) and the 2012 American National Election Study time-series panel include the MMS, as does the US Bureau of Labor Statistics' National Longitudinal Survey of Youth (Hannon and DeFina 2014). The Bureau of Labor Statistics has also gathered MMS measurements for a nationally representative panel survey of young workers (Kreisman and Rangel 2015). Given its widespread use, the MMS serves an ideal case study for testing our hypotheses.

We find that the MMS is vulnerable to spillover effects, and we also corroborate the race-of-coder moderator effect. Both effects are about half a point on the 10-point MMS. Contrary to Hannon and DeFina's (2016) conjecture, providing coders with the MMS palette alongside the images to be coded does not attenuate these effects.

STUDY DESIGN

The data for this study come from a single, institutional-review-board-approved survey experiment in which Amazon Mechanical Turk (MTurk) workers in the United States (443 Black Americans and 457 Whites) rated the skin color of a number of Black Americans and Whites depicted in head-shot photographs.[4] MTurk workers gave consent and were compensated $1.50 for participating in the study. Our design, hypotheses, and analysis were registered with Evidence in Governance and Politics before fielding the experiment (https://doi.org/10.17605/OSF.IO/S4HB8).

Coders and coding tasks

All respondents self-identify as Black or non-Hispanic White and were randomly assigned to one of two coding tasks. Respondents given task 1 rated a sequence of 24 images using the Massey-Martin palette (see fig. 1) as it is conventionally employed. The respondents studied the palette at the beginning of the survey and were asked to commit it to memory. Respondents assigned to coding task 2 rated a sequence of 24 images using the Massey-Martin palette, with the palette shown on the same screen as the photo. The latter task implements the recommendation of Hannon and DeFina (2016), who suggest that interviewers discreetly reference the palette when recording the subject's apparent skin tone.

Photographs

Our study uses 48 photographs in total, 24 images of Whites and 24 images of Black Americans. These are subsets of the 100 photographs of Black Americans and 100 photographs of non-Hispanic Whites that Eberhardt et al. (2004) collected and had college students rate for "stereotypicality" using a Likert scale. We ordered each set of Eberhardt photos by their mean stereotypicality ratings (per her students) and selected quantiles.[5]

Each respondent rated a sequence of 24 photos, with each photo displayed individually on the screen. A position in the sequence of 24 photos shown to the respondent is indexed by $s \in \mathcal{S} = \{1, 2, \ldots, 24\}$. Sequence $\mathcal{S}$ is partitioned between treatment and outcome positions, $\mathcal{S}^{\mathrm{T}}$ and $\mathcal{S}^{\mathrm{O}}$. We set the treatment and outcome positions in $\mathcal{S}$ by fixing the first five positions as treatments and then randomly assigning an "outcome" status to 10 of the remaining 20 positions, holding the resulting sequence constant for all respondents. This resulted in $\mathcal{S}^{\mathrm{O}} = \{6, 7, 10, 12, 15, 17, 19, 20, 22, 23\}$.

The sampling protocol for photos in the treatment condition, $\mathcal{S}^{\mathrm{T}}$, is determined by the treatment condition to which the respondent was assigned. In the Black-photos treatment condition, the photo in each treatment position is chosen by randomly sampling a photo from the set of 24 Black photos. In the White-photos treatment condition, the photo assigned to the treatment position was chosen by randomly sampling a photo from the set of 24 White photos. Outcome-position

3. The MMS is collected as follows. A trained interviewer receives a numbered palette showing human hands in 10 hues and is told to memorize each hue and its associated number (see fig. 1). During interviews, the interviewer recalls the palette from memory and matches the respondent's skin color to one of the colored hands on the palette. Each respondent is coded by a single interviewer.

4. To obtain approximately equal numbers of Black and White MTurk workers, we used a demographic prescreening survey.

5. The terms of use of the Eberhardt photo data prohibit distribution and use of photos. Researchers interested in replication may contact the Eberhardt lab directly to request access to the photos. It is also important to note that these images are the faces of people with no criminal history. Many were Stanford University students or employees at the time they were photographed.

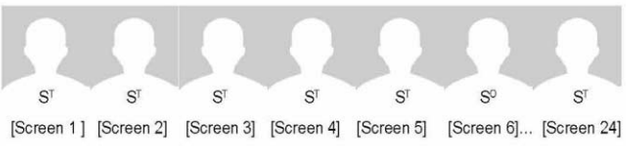

Figure 2. Representation of photo sequence, illustrating treatment and outcome positions.

photos are also randomly sampled.[6] Figure 2 provides a visual representation of the experimental design.

RESULTS

Photo-spillover effects

As stated above, we hypothesize that White treatment photos will cause photos in the outcome positions to be coded darker (a higher number on the MMS), relative to the codings that result when Black photos are in the treatment positions. More specifically we have the following hypothesis.

> **H1.** The average MMS rating among respondents assigned to the White-photos treatment condition will be greater (darker) than the average MMS rating among respondents assigned to the Black-photos treatment condition, where the averages are taken over all respondents assigned to task 1 or 2, all photos, and all outcome positions.[7] We call the average MMS rating among respondents assigned to the White-photos treatment condition minus the average MMS rating among respondents assigned to the Black-photos treatment condition the treatment effect over all outcome positions.

We also estimate photo spillover effects on the coding of the first photograph in an outcome position, which follows five treatment photos. Although we have less statistical power to detect this effect, the size of the treatment effect is likely to be largest for the photos in this position, which is immediately preceded by five consecutive treatment photos. By contrast, no other outcome position is immediately preceded by more than two consecutive treatment photos. Specifically, our hypothesis is as follows.

> **H2.** The average MMS rating of the sixth photo rated (the first outcome photo) among respondents assigned to the White-photos treatment condition will be greater (darker) than the average MMS rating of the sixth photo rated (the first outcome photo) among respondents assigned to the Black-photos treatment condition, where the averages are taken over all respondents assigned to task 1 or 2 and all photos. We call the average MMS rating among respondents assigned to the White-photos treatment condition minus the average MMS rating among respondents assigned to the Black-photos treatment condition the treatment effect for the first outcome position.

We further hypothesize that providing respondents with the MMS palette alongside the photos to be coded (coding task 2) will attenuate the treatment effect of the treatment photos relative to the conventional MMS task where the respondent must recall the palette from memory (coding task 1).[8] Let the treatment effect in task 1 be defined as the average MMS rating among respondents assigned to the White-photo treatment condition and task 1 minus the average MMS rating among respondents assigned to the Black-photo treatment condition and task 1. Similarly, let the treatment effect in task 2 be defined as the average MMS rating among respondents assigned to the White-photo treatment condition and task 2 minus the average MMS rating among respondents assigned to the Black-photo treatment condition and task 2. Here the averages are taken over all respondents assigned to task 1 or task 2 as appropriate, all photos, and all outcome positions.

With those definitions in place, the hypothesis can be stated as follows.

> **H3.** The treatment effect in task 2 (MMS coding with template) will be less than the treatment effect in task 1 (conventional MMS coding).

Photo-spillover results

Given the random assignment of manipulations to respondents, all of the estimands described above can be consistently estimated by sample averages with clustered standard errors to account for the fact that there are multiple observations from each respondent. Results are summarized in table 1.

Averaging across all 48 photos in our sample, the effect of the White treatment photos (relative to the Black treatments) increases the average MMS rating of a photo in the outcome positions by about 0.5 points on the 10-point MMS. The effect is somewhat larger on the first outcome position,

6. The appendix provides more details.

7. We also average over the treatment photo selection vectors, which means that, in expectation, each photo is equally likely to be coded by each respondent.

8. We did not include an attenuation-effect estimand for the first outcome position in our preanalysis plan, figuring that we might lack power to pin down that effect.

Table 1. Treatment-Photo Effects on MMS Coding

Hypothesis	Description	Estimate	*p*
1	Treatment effect, all outcome positions	.48	1.9e−07
		(.09)	
2	Treatment effect, first outcome position	.63	.006
		(.23)	
3	Treatment effect from task 1 minus treatment effect from task 2	−.14	.417
		(.17)	

Note. The *p*-values are calculated using a two-tailed test against the null hypothesis that the estimand equals zero, with clustered standard errors (in parentheses) to account for multiple observations from each respondent and of each photograph.

which follows five treatments, than on the average outcome position (0.63 vs. 0.48).

There is no evidence that providing respondents with the MMS palette alongside the photographs to be coded reduces the treatment-photo effect. Rather, the treatment effect is actually slightly larger when respondents see the palette displayed alongside the photograph to be coded, although the difference is not close to statistically significant.[9]

Race-of-coder moderator effects

Race-of-coder estimands and hypotheses. In light of previous work finding that White interviewers using the MMS rate Black subjects as darker on average than Black interviewers do (Hannon and DeFina 2014; Hill 2002), we hypothesize that White MTurk workers will similarly code photographs of Blacks as darker on average than Black MTurk workers do. This hypothesis can be stated as follows.

> **H4a.** The average MMS rating of Black photos among White respondents will be greater (darker) than the average MMS rating of Black photos among Black respondents, where the averages are taken over all respondents assigned to task 1 or 2, all Black photos, and all outcome positions. We call the average MMS rating among White respondents minus the average rating among Black respondents the race-of-coder effect for Black photos.

We also separately investigate the effect for all White photos.

> **H4b.** The average MMS rating of White photos among White respondents will be greater (darker) than the average MMS rating of White photos among Black respondents, where the averages are taken over all respondents assigned to task 1 or 2, all White photos, and all outcome positions. We call the average MMS rating among White respondents minus the average rating among Black respondents the race-of-coder effect for White photos.

We hypothesize that providing respondents with the MMS palette alongside the photographs to be coded will attenuate the race-of-coder effect, as follows.

> **H5a.** The race-of-coder effect for Black photos within task 1 (standard MMS coding) will be greater than the race-of-coder effect for Black photos within task 2 (MMS coding with template).

Again, we separately investigate the effect for the photos of White individuals.

> **H5b.** The race-of-coder effect for White photos within task 1 (standard MMS coding) will be greater than the race-of-coder effect for White photos within task 2 (MMS coding with template).

Race-of-coder results. We find a modest but highly statistically significant difference in the average MMS ratings of Black and White coders. The difference mainly registers in the coding of Black photographs. The average difference between Black and White coders in the rating of the Black photos in our sample is about 0.5 points on the 10-point MMS, with White coders perceiving Blacks to have darker skin than Black coders do (see table 2). This is essentially the same effect size that we found with treatment photos. Averaging

9. Displaying the template (task 2) does increase the average MMS rating of outcome-position photos by a little more than half a point. In the White-treatment condition, the average rating goes from 4 to 4.7, and in the Black-treatment condition, it goes from 3.6 to 4.2. This suggests that coders applying the scale from memory may tend to err in the "lighter" direction. But providing the template does not attenuate the treatment-photo effect.

Table 2. Race-of-Coder Effects on MMS Coding of Photos of Blacks

Hypothesis	Description	Estimate	p
4a	Race-of-coder effect (Black photos)	.52 (.08)	1.4e−10
5a	Race-of-coder effect from task 1 minus race-of-coder effect from task 2 (Black photos)	.03 (.14)	.809

Note. The p-values are calculated using a two-tailed test against the null hypothesis that the estimand equals zero, with clustered standard errors (in parentheses) to account for multiple observations from each respondent.

over White photos instead of Black photos, we see that White coders assign an MMS rating 0.2 points higher (darker) than the average rating of Black coders (see table 3).

The Black-White coder difference is virtually identical whether coders are asked to apply the MMS from memory or are presented with the MMS palette alongside the photos to be coded. Just as providing the palette failed to attenuate photo-spillover effects, so too does it fail to attenuate the race-of-coder moderator effect. This is true for both Black and White photos.

An alternative approach based on pairwise comparisons

Research in several fields indicates that measurement via pairwise comparisons is generally superior to Likert-type scales (Dittrich et al. 2007; Oishi et al. 2005; Phelps et al. 2015). As the MMS is a Likert-type scale, one might reasonably wonder whether a pairwise alternative would prove less vulnerable to spillover and race-of-coder effects.

A random subset of our survey respondents was assigned to view pairs of randomly selected photos and asked to indicate which person in the pair has the darker skin color. More specifically, we construct various data-collection scenario contrasts' and derive photo rankings from pairwise and MMS data corresponding to each scenario.[10] For example, to evaluate the robustness of each method to variation in the race of the coder, we construct one ranking of the photographs using data from Black coders applying the MMS, another ranking using data from Black coders making pairwise comparisons, a third ranking using data from White coders applying the MMS, and a fourth ranking using data from White coders making pairwise comparisons. The question of interest is whether the ranking inferred from White coders' observations is closer to the ranking inferred from Black coders' observations when the observations consist of pairwise comparisons than when the observations consist of MMS ratings.

We find that rankings inferred from pairwise data are generally noisier (higher variance) than rankings inferred from the same amount of MMS data, at least for smaller samples, but that if one standardizes the distance between rankings inferred from two data-collection scenarios, the pairwise methods are generally more robust to the data-collection contrast. However, in only one of our comparisons does the difference reach the conventional 5% threshold for statistical significance. These results suggest the potential usefulness of pairwise comparisons to mitigate some of the issues associated with using the MMS scale to measure skin color.[11]

DISCUSSION

Our results corroborate and extend the body of work on measurement problems with the MMS. Using a convenience sample of MTurk workers as coders, we provided the first test of spillover effects in application of the MMS, finding that treatment photographs of Whites cause "outcome" persons to be rated about 0.5 points darker (on a 10-point scale) than if the treatment photos are of Black Americans. We also replicated an earlier finding that White coders applying the MMS rate Black subjects as darker on average than Black coders do (Hannon and DeFina 2014; Hill 2002). The size of the Black-White coder difference in coding of Black Americans' skin color is about the same as the size of the spillover effect. When respondents were provided with the MMS palette, a modest shift in skin-color ratings toward the darker end of the scale occurred, but race-of-coder and photo-spillover effects were not attenuated. Finally, we proposed an alternative approach that relies on pairwise comparisons of photos to evaluate skin color; while

10. For the pairwise comparisons data, we fit a Bayesian version of Thurstone's (1927) model. For the MMS data, we fit a Bayesian ordinal probit model in which the only covariates are photo-specific dummy variables. We then convert the latent measures of skin color from each model into ordinal rankings of photographs by skin color and use the rankings to compare the two methods of encoding perceptual data.

11. A more detailed discussion is found in app. sec. 3.2. Results are presented in app. table 2.

Table 3. Race-of-Coder Effects on MMS Coding of Photos of Whites

Hypothesis	Description	Estimate	*p*
4b	Race-of-coder effect (White photos)	.23 (.04)	3.3e−08
5b	Race-of-coder effect from task 1 minus race-of-coder effect from task 2 (White photos)	−.10 (.08)	.195

Note. The *p*-values are calculated using a two-tailed test against the null hypothesis that the estimand equals zero, with clustered standard errors (in parentheses) to account for multiple observations from each respondent.

the results show promise, further work and data-collection efforts are needed to fully evaluate this approach.

One potential lesson from our study is that researchers using Likert-type measures of skin color should try to integrate or average the perceptions of multiple observers from different racial groups, at least if the goal it to see how a person is perceived "on average." Another lesson is to avoid coding or interviewing protocols in which some persons are more likely than others to be observed and coded in a racially homogeneous context, that is, at a moment when the coder has recently observed many other people of one race and few or no people of other races. Yet even if the subjects to be coded are presented in random order, with each subject equally likely to be appear in each position in the sequence, our results imply that the average skin-color ratings of Black and White persons are likely to depend on the relative frequency of Blacks and Whites in the pool of persons to be coded, as well as the relative frequency of Blacks and Whites in the pool of coders. This means that ratings on the same Likert-type scale may not be comparable across studies.

In combination with the burgeoning body of applied work on colorism—the vast majority of which relies on Likert-type scales[12]—we hope our results will motivate further work on the underlying measurement issues. One might be tempted to give up on attempts to measure subjective perceptions of skin color and instead focus on obtaining objective measures of skin color—such as those from spectrophotometer readings—as a way of minimizing measurement error. This is the approach taken by Branigan et al. (2013), who argue that perceived skin color is simply too difficult to measure reliably. However, this focus on objective measures is not without costs, as it effectively closes off questions about how individuals perceive skin color and how those perceptions effect political and social outcomes.

Social scientists now understand race as a social construction (Omi and Winant 1986), which means that measurement of race ought to be grounded in perceptions. In the political world, there is strong evidence that candidates' electability and viability can hinge on their skin color (Burge et al. 2020; Hochschild and Weaver 2007; Weaver 2012) and that citizens' experiences with racial discrimination affect their sense of political efficacy and desire to participate in the polity (Canache et al. 2014; Ostfeld and Yadon 2020; Wilkinson and Earle 2012). However, if the underlying measures of skin color or race on which such studies rely are contaminated by serious measurement error, the estimates may be biased. It is therefore incumbent on researchers to carefully consider how best to measure skin color and to understand the limitations of scales that are now in widespread use.

12. See the appendix, which documents this trend.

REFERENCES

Blair, Irene V., Charles M. Judd, and Kristine M. Chapleau. 2004. "The Influence of Afrocentric Facial Features in Criminal Sentencing." *Psychological Science* 15 (10): 674–79.

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.

Burch, Traci. 2015. "Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing." *Journal of Empirical Legal Studies* 12 (3): 395–420.

Burge, Camille D., Julian J. Wamble, and Rachel R. Cuomo. 2020. "A Certain Type of Descriptive Representative? Understanding How the Skin Tone and Gender of Candidates Influences Black Politics." *Journal of Politics* 82 (4): 1596–601.

Canache, Damarys, Matthew Hayes, Jeffery J. Mondak, and Mitchell A. Seligson. 2014. "Determinants of Perceived Skin-Color Discrimination in Latin America." *Journal of Politics* 76 (2): 506–20.

Dávila, Alberto, Marie T. Mora, and Sue K. Stockly. 2011. "Does Mestizaje Matter in the US? Economic Stratification of Mexican Immigrants." *American Economic Review* 101 (1): 593–97.

Dittrich, Regina, Brian Francis, Reinhold Hatzinger, and Walter Katzenbeisser. 2007. "A Paired Comparison Approach for the Analysis of Sets of Likert-Scale Responses." *Statistical Modeling* 7 (1): 3–28.

Eberhardt, Jennifer L., Paul G. Davies, Valerie J. Purdie-Vaughns, and Sheri Lynn Johnson. 2006. "Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes." *Psychological Science* 17 (5): 383–86.

Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies. 2004. "Seeing Black: Race, Crime, and Visual Processing." *Journal of Personality and Social Psychology* 87 (6): 876–93.

Espino, Rodolfo, and Michael Franz. 2002. "Latino Phenotypic Discrimination Revisited: The Impact of Skin Color on Occupational Status." *Social Science Quarterly* 83 (2): 612–23.

Frank, Reanne, Ilana Redstone Akresh, and Bo Lu. 2010. "Latino Immigrants and the US Racial Order: How and Where Do They Fit In?" *American Sociological Review* 75 (3): 378–401.

Garcia Bedolla, Lisa. 2005. *Fluid Borders: Latino Power, Identity, and Politics in Los Angeles*. Berkeley: University of California Press.

Hannon, Lance. 2015. "White Colorism." *Social Currents* 2 (1): 13–21.

Hannon, Lance, and Robert DeFina. 2014. "Just Skin Deep? The Impact of Interviewer Race on the Assessment of African American Respondent Skin Tone." *Race and Social Problems* 6 (4): 356–64.

Hannon, Lance, and Robert DeFina. 2016. "Reliability Concerns in Measuring Respondent Skin Tone by Interviewer Observation." *Public Opinion Quarterly* 80 (2): 534–41.

Herman, Melissa R. 2011. "Methodology and Measurement in the Study of Multiracial Americans: Identity, Classification, and Perceptions." *Sociology Compass* 5 (7): 607–17.

Hersch, Joni. 2008. "Profiling the New Immigrant Worker: The Effects of Skin Color and Height." *Journal of Labor Economics* 26 (2): 345–86.

Hersch, Joni. 2009. "Skin Color Discrimination and Immigrant Pay." *Emory Law Journal* 58 (2): 357–77.

Hersch, Joni. 2011. "The Persistence of Skin Color Discrimination for Immigrants." *Social Science Research* 40 (5): 1337–49.

Hill, Mark E. 2002. "Race of the Interviewer and Perception of Skin Color: Evidence from the Multi-City Study of Urban Inequality." *American Sociological Review* 67 (1): 99–108.

Hochschild, Jennifer L., and Vesla Weaver. 2007. "The Skin Color Paradox and the American Racial Order." *Social Forces* 86 (2): 643–70.

Hunter, Margaret. 2005. *Race, Gender, and the Politics of Skin Tone*. London: Routledge.

Iyengar, Shanto, and Kyu S. Hahn. 2007. "Natural Disasters in Black and White: How Racial Cues Influenced Public Response to Hurricane Katrina." Working paper.

Jenq, Christina, Jessica Pan, and Walter Theseira. 2015. "Beauty, Weight, and Skin Color in Charitable Giving." *Journal of Economic Behavior and Organization* 119:234–53.

Klonoff, Elizabeth A., and Hope Landrine. 2000. "Is Skin Color a Marker for Racial Discrimination? Explaining the Skin Color–Hypertension Relationship." *Journal of Behavioral Medicine* 23 (4): 329–38.

Kreisman, Daniel, and Marcos A. Rangel. 2015. "On the Blurring of the Color Line: Wages and Employment for Black Males of Different Skin Tones." *Review of Economics and Statistics* 97 (1): 1–13.

Lerman, Amy E., Katherine T. McCabe, and Meredith L. Sadin. 2015. "Political Ideology, Skin Tone, and the Psychology of Candidate Evaluations." *Public Opinion Quarterly* 79 (1): 53–90.

Massey, Douglas S., Camille Z. Charles, Garvey F. Lundy, and Mary J. Fischer. 2003. *The Source of the River: The Social Origins of Freshmen at America's Selective Colleges and Universities*. Princeton, NJ: Princeton University Press.

Massey, Douglas S., and Jennifer A. Martin. 2003. "The NIS Skin Color Scale." Office of Population Research, Princeton University.

Meijer, Erik, Edward Oczkowski, and Tom Wansbeek. 2021. "How Measurement Error Affects Inference in Linear Regression." *Empirical Economics* 60 (1): 131–55.

Oishi, Shigehiro, Jungwon Hahn, Ulrich Schimmack, Phanikiran Radhakrishan, Vivian Dzokoto, and Stephen Ahadi. 2005. "The Measurement of Values across Cultures: A Pairwise Comparison Approach." *Journal of Research in Personality* 39:299–305.

Omi, Michael, and Howard Winant. 1986. *Racial Formation in the United States: From the 1960s to the 1980s*. London: Routledge.

Ostfeld, Mara, and Nicole Yadon. 2020. "The Gravity of Color: Skin Color and Power in Contemporary American Politics." Unpublished manuscript.

Phelps, Andrew S., David M. Naeger, Jesse L. Courtier, Jack W. Lambert, Peter A. Marcovici, Javier E. Villanueva-Meyer, and John D. MacKenzie. 2015. "Pairwise Comparison versus Likert Scale for Biomedical Image Assessment." *American Journal of Roentgenology* 204 (1): 8–14.

Telles, Edward. 2004. *Race in Another America: The Significance of Skin Color in Brazil*. Princeton, NJ: Princeton University Press.

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.

Thurstone, L. L. 1927. "A Law of Comparative Judgment." *Psychological Review* 34:273–86.

Villareal, Andres. 2010. "Stratification by Skin Color in Contemporary Mexico." *American Sociological Review* 5 (75): 652–78.

Wade, T. Joel, Melanie Romano, and Leslie Blue. 2004. "The Effect of African American Skin Color on Hiring Preferences." *Journal of Applied Social Psychology* 34 (12): 2550–58.

Weaver, Vesla. 2012. "The Electoral Consequences of Skin Color: The 'Hidden' Side of Race in Politics." *Political Behavior* 34 (1): 159–92.

Wilkinson, Betina Cutaia, and Emily Earle. 2012. "Taking a New Perspective on Latino Racial Attitudes." *American Politics Research* 41 (5): 783–818.