RELIABILITY CONCERNS IN MEASURING RESPONDENT SKIN TONE BY INTERVIEWER OBSERVATION

LANCE HANNON

ROBERT DEFINA*

Abstract The current study assesses the intercoder reliability of one of the most important skin tone measurement instruments—the Massey–Martin scale. This scale is used in several high-profile social surveys, but has not yet been psychometrically evaluated. The current evaluation is only possible because, for the first time, the General Social Survey’s 2010–2014 panel used the instrument to guide interviewers’ skin tone observation of the same respondents in two different years (2012 and 2014). Despite the widespread use of the Massey–Martin scale to investigate potential effects of skin tone on social attitudes and outcomes, the data suggest that the measure has low intercoder reliability. Implications for researchers and survey practitioners are discussed.

A number of social surveys include measures of respondent skin tone obtained from interviewer observation. A key reason for including such measures is the recognition that broad racial categories such as “Black” inadequately capture socially meaningful differences in racial appearance that can affect important social outcomes. For example, some research has suggested that darker-skinned African Americans are more likely to suffer harsher penalties in the criminal justice system (Blair, Judd, and Chapleau 2004) and receive significantly lower wages than lighter-skinned African Americans (Goldsmith, Hamilton, and Darity 2006). Other research has indicated that skin tone affects attitudes among Latinos/as regarding commonality with Whites (Wilkinson and Earle 2012) as well as self-esteem and self-identity (Hunter 2005). Forecasted demographic shifts related to Latino/a immigration have also led to predictions that a new racial order will emerge in the United States, where

LANCE HANNON and ROBERT DEFINA are professors in the Sociology and Criminology Department at Villanova University, Villanova, PA, USA. *Address correspondence to Robert DeFina, Villanova University, Sociology and Criminology Department, 800 Lancaster Avenue, Villanova, PA 19085, USA; e-mail: robert.defina@villanova.edu.

“preference for people who are light-skinned will become a more important factor in all kinds of social transactions” (Bonilla-Silva 2003, 352; see also 2009).

Furthermore, numerous social commentators have argued that discrimination based on skin tone, or “colorism,” can persist despite progress in reducing racism (Harrison 2010; Hochschild 2012). This may occur, in part, because government agencies charged with monitoring and reducing discrimination typically do not record or have access to skin color data, while they usually do have access to information regarding traditional racial categories. For example, Title VII of the 1964 Civil Rights Act prohibits employment discrimination based on “race, *color*, religion, sex and national origin.” However, claims of colorism to the Equal Employment Opportunity Commission are complicated by the absence of any remotely standardized definition of what constitutes “light” or “dark” skin (Jones 2010).

While social surveys in the United States have also historically been without a standard method for measuring skin tone, in recent years a consensus appears to have developed around one skin tone survey instrument—the Massey–Martin scale (2003). This popular interviewer-assessed skin tone measure incorporates a professionally designed color chart to anchor interviewer judgments. The scale is used in the National Longitudinal Study of Freshmen, New Immigrant Survey, National Longitudinal Survey of Youth, American National Election Studies, and most recently, the General Social Survey. Until now, it has been impossible to evaluate the intercoder reliability of the Massey–Martin scale with publicly available data. Potentially reflecting the belief that the color palette associated with the scale would eliminate interviewer subjective bias, a single interviewer has traditionally assessed a respondent’s skin tone (in a single year). The recently released panel component of the 2012–2014 General Social Survey provides an opportunity to test cross-interviewer levels of agreement in skin tone measurement. The results of this test should inform how researchers interpret the magnitude and statistical significance of skin tone effects, and inform survey practice regarding the administration of the Massey–Martin scale in the future.

The Massey–Martin Scale

Introduced in 2012, skin tone measurement in the General Social Survey approximated a procedure initially developed by Douglas Massey and Jennifer Martin (2003). Ten shades of skin color were mapped to a pictorial guide of male hands corresponding to the points 1 to 10 (see figure 1). Importantly, the interviewer was told to record the respondent’s skin color after the interview was complete and was not allowed to directly compare the color palette to the respondent’s skin. More specifically, the interviewer was instructed as follows:

Scale of Skin Color Darkness

Figure 1. Massey–Martin Skin Color Guide for Interviewers.

Please complete this section as soon as possible after the interview. Never bring the card with you into the house. Familiarize yourself with the card before entering the location of the interview and remember that we are asking for the number of the color most closely matched to the respondent's facial features. It might be helpful for you to keep the card handy in your car so you can refer to it shortly after leaving the interview. Please record the color from the color card that most closely corresponds to the respondent's facial color.

In cases where the General Social Survey was conducted over the telephone, the interviewer was instructed to select "Don't know" in response to this skin color question.

Results

Table 1 displays average intraclass correlations for skin tone measurements in 2012 and 2014 for Black and Latino/a individuals in the most recent General Social Survey longitudinal subset.¹ Panel A includes all respondents with skin tone assessments for both time periods and valid data for self-reported race/ethnicity in 2012. As can be seen, the correlations are quite modest, given the expectation that skin tone would be relatively fixed for individuals. The

1. We focus on these two specific minority populations because they are the primary subjects in colorism research in the United States, and the General Social Survey provides reasonably large sample sizes in each case. Although not as relevant to most of the discourse on colorism, the overall intraclass correlation for non-Hispanic Whites ($n = 627$) was also quite low (.142). The instructions for the basic race question (variable name *Racecen1*) are What is your race? Indicate one or more races that you consider yourself to be. Available responses are White, Black or African American, American Indian or Alaska native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Guamanian or Chamorro, Samoan, Other Pacific Islander, Some other race, Don't know, no answer, and not applicable. In cases involving the selection of multiple races, we used the first selected race. The instructions for the Latino/a question (variable name *Hispanic*) are If R is female, read Latina; if R is male, read Latino. Are you Spanish, Hispanic, or Latino/Latina? (Yes/No).

Table 1. Intraclass Correlations Between the Respondent’s 2012 and 2014 Skin Tone Rating by Self-Reported Race/Ethnicity

Panel A: All Black and Latino/a respondents		
Self-reported race/ethnicity of respondent	<i>n</i>	<i>ICC</i>
Black	160	.451
Latino/a	88	.079
Panel B: Only respondents known to have a new interviewer in 2014		
Self-reported race/ethnicity of respondent	<i>n</i>	<i>ICC</i>
Black	115	.279
Latino/a	76	.003

NOTE.—The source is the 2012–2014 General Social Survey (2010 Panel). The equivalent Pearson correlation coefficients were all below .5 (with or without non-response panel sampling weights).

intraclass correlation for Latinas/os was only .079, and the correlation for African Americans, while noticeably higher at .451, was still well below acceptable levels for inter-rater reliability. In terms of coding agreement, the data ($n = 160$) indicate that less than a quarter of the African American respondents listed in the top three categories of skin darkness in 2012 fell into these same top three categories in 2014, and only slightly more than half of the respondents classified in the bottom three and middle four categories in 2012 were in those same groupings two years later.² For Latinos/as, skin tone variability was concentrated in a narrower range at the lower end of the skin darkness scale (mostly between 1 and 4), but less than 25 percent of respondents had the same rating in 2014 as in 2012.

However, even the modest levels of association suggested by the correlations in panel A likely overstate the reliability of the Massey–Martin scale, since some respondents probably had the exact same interviewer in both years. While the data do not contain an identifying code for interviewers that is consistent across the panel years, researchers can isolate those respondents who had a new interviewer by using available information concerning interviewer age, sex, race/ethnicity, and time employed with the National Opinion Research Center. Here, new interviewers are those with patterns in these variables that are clearly inconsistent with a respondent having the

2. In other words, even if a researcher decided to sacrifice the more fine-tuned variation and recode the Massey–Martin scale into a simpler light-medium-dark scheme, the level of reliability would still be unacceptably low in this case. About half of the Black respondents would have an inconsistency in skin tone measurement using this collapsed three-category measure.

same interviewer in both panels. So, for example, when a given respondent's interviewer is listed as being five years younger in 2014 compared to the one in 2012, or having 10 years' more experience in 2014 than 2012, that interviewer was considered definitively new. These comparisons reduce the non-Hispanic Black sample size from 160 to 115 and the Latino/a sample size from 88 to 76.

The results for the group of respondents with known different interviewers in 2012 and 2014 are displayed in panel B. As expected, the correlations drop to even lower levels when the samples are restricted to cases where assessments clearly involve different interviewers. Indeed, the intraclass correlation for Latinas/os is close to zero and the association for Black respondents is reduced by nearly 40 percent when this sample filter is applied.

Overall, our findings are logically consistent with other studies demonstrating that samples of Black respondents assessed by White interviewers have higher average skin darkness ratings than samples of Black respondents assessed by Black interviewers (Hill 2002; Hannon and DeFina 2014).³ These studies have argued that the overall intercoder reliability of skin tone measures will hinge on the degree to which respondents are matched to interviewers of the same race, particularly if one assumes that, for example, Black interviewers are better able to reliably judge variation in Black skin tone than White interviewers. Still, a close look at the data suggests that race-of-interviewer effects can only partially account for the low levels of intercoder agreement reported here.

Focusing on Black respondents, table 2 provides descriptive data on both the prevalence of skin tone rating disagreement and, in cases with disagreement, the average absolute difference between ratings by racial composition of interviewer pairs. While the data suggest that the prevalence of disagreement is slightly lessened when interviewers are of the same race, and this is especially true when interviewers are of the same race as the respondent, the prevalence of disagreement remains very high (about 80 percent) regardless of the racial composition of interviewer pairs. For the Massey–Martin instrument, the most meaningful benefit from matching Black respondents to Black interviewers appears to be a reduction in the absolute difference between inconsistent values (the mean discrepancy is reduced from about 2 units to 1 on the 10-point scale).⁴

3. This general finding can also be seen in the General Social Survey panel data used here. For example, in 2014 the mean value for the Massey–Martin scale for African American respondents with African American interviewers was 4.73 ($n = 41$), while the mean value for African American respondents matched to a White interviewer was 5.91 ($n = 109$).

4. It is important not to exaggerate the significance of this result, especially considering the relatively small sample size ($n = 18$) for Black respondents solely interviewed by Black interviewers. Moreover, the intraclass correlation for this group is still low (.453), and even if researchers were to use the collapsed three-category recoded scale mentioned earlier for this group (instead of the original 10-point scale), one-third of this sample would still have inconsistent values.

Table 2. Levels of Disagreement in the Assessment of Black Respondent Skin Tone by Racial Composition of Interviewer Pairs

Summary measure of disagreement	Black-White interviewer combo (n = 43)	White-White interviewer combo (n = 73)	Black-Black interviewer combo (n = 18)
Percent in disagreement	81.4	78.1	77.8
Average amount of disagreement	2.1	2.2	1.1

NOTE.—The source is the 2012–2014 General Social Survey (2010 Panel).

Conclusion

Almost all surveys that measure respondent skin tone with the Massey–Martin scale, regardless of cross-sectional or longitudinal design, do so only once. This may be both a matter of cost efficiency as well as faith in the ability of the color guide to help standardize interviewer observations. Recently released panel data from the General Social Survey provide a rare opportunity to examine the intercoder reliability of the Massey–Martin scale. Our analysis of those data indicates that the assumption that skin tone assessments will be consistent across raters is not tenable.

One of the primary motivations for including a measure of skin tone in social surveys is to enable researchers to uncover a form of racialized discrimination—colorism—that is often more nuanced than discrimination based on broad racial categories. However, the ability of researchers to discern significant skin tone effects within racial categories is contingent on the degree of reliability of the skin tone measure. Low levels of reliability will bias the coefficients for the skin tone variable toward zero and reduce the likelihood of finding statistical significance in research investigating the possibility of colorism.⁵ Additionally, research using the Massey–Martin scale as a control variable may misinterpret the coefficients of primary interest as being free of spurious association with the respondent’s skin color.

We recommend that, at a minimum, future surveys should implement a different procedure for administering the Massey–Martin scale. The current procedure requires interviewers to refrain from looking at the color chart when the respondent is in front of them, presumably to avoid awkwardness or direct questions about the interviewer’s observation. Instead, interviewers are asked

5. As noted earlier, levels of reliability may systematically vary by respondent and interviewer characteristics. In particular, respondents assessed by two interviewers of a different race will likely have lower levels of consistency in skin tone values than those assessed by two interviewers of the same race. Regardless of the source of the inconsistency, if the average skin tone assessment in a sample is of significantly low reliability, the average effect of skin tone in analyses based on that sample will be significantly biased toward zero.

to memorize highly nuanced distinctions in color and recall them later.⁶ In our opinion, this is not a reasonable request. In the face of such a challenging demand, many interviewers, but not necessarily all, will simply revert to an idiosyncratic definition of light or dark skin. We recommend having interviewers reference a simplified color chart during the interview to hopefully produce more consistent and generalizable observations.⁷

Alternatively, a case could be made that, while not particularly useful as a generalizable indicator of respondent skin tone beyond the survey interview setting, the values from the Massey–Martin scale as currently implemented can still tell us something about survey interviewer perceptions. So, for example, a researcher might ask if an interviewer’s perception of respondent cooperation is related to that interviewer’s judgment of respondent skin tone. However, this was clearly not the intended purpose of the Massey–Martin scale, and in our view, a reliable and more generalizable measure of respondent skin tone is still possible and desirable.

As Branigan et al. (2013, 1662) and others have argued, a “promising alternative to interviewer coding is the objective measurement of light reflectivity provided by a spectrophotometer.”⁸ Indeed, in recent years such devices have become much more accessible in terms of cost (less than \$2,000) and much more user-friendly (equivalent to a digital camera). Nevertheless, we suspect that few survey organizations will be swayed toward funding this type of measurement approach. A less costly strategy might involve taking a digital photograph of a respondent holding a color palette for calibration purposes (Marguier et al. 2007).

Reliably measuring variation in skin tone is clearly a difficult task that involves a variety of complex issues. However, the task is also clearly an important one, especially considering the evolving racial and ethnic diversity of the United States and increased attention to legal protection against colorism (Equal Employment Opportunity Commission 2007).

Appendix. Survey Description

The National Opinion Research Center administers the widely referenced General Social Survey (GSS). This nationally representative US-based survey is conducted every other year, mainly via face-to-face interviewing. It includes both English and Spanish speakers. Since 2006, the GSS has augmented its repeating cross-section with a panel study component to allow assessment of

6. Indeed, even when we used a spectrophotometer (Xrite RM200QC) to quantify the lightness and darkness of colors depicted on the Massey–Martin scale, the device barely distinguished the graphics for points 8 and 9 (while discerning a much larger separation between points 9 and 10).

7. With computer-assisted personal interviewing, it may be possible to do this discreetly.

8. Branigan et al. (2013) report significant positive relationships between skin lightness (assessed with a spectrophotometer) and educational attainment within racial groups.

change over time in the same individuals. The GSS utilizes a multistage, stratified probability sample of the non-institutionalized adult population (18 years old and older). The primary sampling units are Metropolitan Statistical Areas and non-metropolitan counties. Interviews were typically conducted between March and October, and the overall response rate (AAPOR RR5) for 2010 was 70.3 percent. For the GSS panel, among the 2,044 cases that were first interviewed in 2010, 1,551 were re-interviewed in 2012 and 1,304 were re-interviewed in 2014. The GSS is funded by the National Science Foundation.

References

- Blair, Irene, Charles Judd, and Kristine Chapleau. 2004. "The Influence of Afrocentric Facial Features in Criminal Sentencing." *Psychological Science* 15:674–79.
- Bonilla-Silva, Eduardo. 2003. "New Racism, Color-Blind Racism, and the Future of Whiteness in America." In *White Out: The Continuing Significance of Racism*, edited by Ashley Doane and Eduardo Bonilla-Silva, 345–60. New York: Routledge.
- . 2009. *Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States* (2nd ed.). Boulder, CO: Rowman and Littlefield.
- Branigan, Amelia J., Jeremy Freese, Assaf Patir, Thomas W. McDade, Kiang Liu, and Catarina Kiefe. 2013. "Skin Color, Sex, and Educational Attainment in the Post-Civil Rights Era." *Social Science Research* 42:1659–1674.
- Equal Employment Opportunity Commission (EEOC). 2007. *EEOC Takes New Approach to Fighting Racism and Colorism in the 21st-Century Workplace*. <http://www.eeoc.gov/eeoc/initiatives/e-race/index.cfm>.
- Goldsmith, Arthur, Darrick Hamilton, and Sandy Darity. 2006. "Shades of Discrimination: Skin Tone and Wages." *American Economic Review* 96:242–45.
- 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* 6:356–64.
- Harrison, Matthew S. 2010. "Colorism: The Often Undiscussed "ism" in America's Workforce." *The Jury Expert* 22:67–72.
- Hill, Mark E. 2002. "Race of Interviewer and Perception of Skin Color: Evidence from the Multi-City Study of Urban Inequality." *American Sociological Review* 67:99–108.
- Hochschild, Jennifer L. 2012. "Lumpers or Splitters: Analytic and Political Choices in Studying Colour Lines and Colour Scales." *Ethnic and Racial Studies* 35:1132–1136.
- Hunter, Margaret L. 2005. *Race, Gender, and the Politics of Skin Tone*. New York: Routledge.
- Jones, Trina. 2010. "Intra-Group Preferencing: Proving Skin Color and Identity Performance Discrimination." *NYU Review of Law and Social Change* 34:657–707.
- Marguer, Joanna, Nina Bhatti, Harlyn Baker, Michael Harville, and Sabine Susstrunk. 2007. "Assessing Human Skin Color from Uncalibrated Images." *International Journal of Imaging, Systems, and Technology* 17:143–51.
- Massey, Douglas S., and Jennifer A. Martin. 2003. *The NIS Skin Color Scale*. Princeton, NJ: Princeton University Press.
- Wilkinson, Betina Cutaia, and Emily Earle. 2012. "Taking a New Perspective to Latino Racial Attitudes: Examining the Impact of Skin Tone on Latino Perceptions of Commonality with Whites and Blacks." *American Politics Research* 41:783–818.