# Measuring Skin Tone: A Test of Common Approaches[1]

Mary E. Campbell
*Texas A&M University*

Verna M. Keith
*University of Alabama at Birmingham*

[1] Draft; please do not quote or cite without permission. Please address correspondence to the first author at m-campbell@tamu.edu. The authors would like to thank Lance Hannon for the spectrometer loan, Adrienne Carter-Sowell and Phia Salter for their help with the project development, and Katie Constantin, Vanessa Gonlin, Emily Knox, Gabe Miller, David Orta, and Jesus Smith for their help with data collection.

**Abstract**

In this experiment, we compare three different ways of asking raters to evaluate skin tone, testing whether methods created to reduce interrater variation (such as using a color palette to create a "norm" for responses) are effective. We compare two popular scales: a text-based 5-point skin color scale (which asks raters to classify pictures on a scale from very light to very dark) and a 10-point palette-based skin color scale (which asks raters to choose a number from 1 to 10 with pictures associated with each number). We also ask raters to use a more complex two-axis color chart to rate pictures, in order to test whether addressing common criticisms of the palette-based scales improves ratings. White and Latinx experiment participants complete a demographic questionnaire and rate a randomly selected set of 16 pictures. We find that characteristics of the raters such as gender, race, the amount of contact with diverse racial groups, and immigration status affect skin tone ratings that observers assign, no matter what type of measure is used. We discuss the implications of the differences between the measures for designing studies of inequality and demography.

**The Need for a Systematic Study of Skin Tone Measures**

Racialized inequality in the United States is not only shaped by self-identified racial categories (e.g. White, Black, etc.), but also by other experiences of race such as how one is perceived by others (Campbell, Bratter, and Roth 2016; Saperstein 2012; Vargas 2015). An important aspect of this is *colorism*, the privileging of light-skinned minorities over dark-skinned minorities. Colorism is especially important to our understanding of racialized social processes and experiences for African Americans and Latinxs (Hunter 2005). Darker complexion among African Americans and Latinxs has been linked to disadvantages in domains such as status attainment (Allen, Telles, and Hunter 2000; Bailey, Saperstein, and Penner 2014; Drake and Cayton 1962; Frank, Akresh, and Lu 2010; Keith and Herring 1991; Monk 2014), educational attainment (Hersch 2006; Monk 2014), perceptions in politics (Caruso, Mead, and Balcetis 2009; Weaver 2012), mate selection (Bond and Cash 1992; Hunter 2002; Ross 1997), body image dissatisfaction (Bond and Cash 1992), self-worth (Keith and Thompson 2003), discriminatory experiences (Klonoff and Landrine 2000), and health (Dressler 1993, 1991).

Unfortunately, the *measures* of skin tone that are used in this literature are largely untested. Different measures are used in different studies, with little evidence to suggest which method is the most valid or reliable. Thus it is possible that the importance of colorism is incorrectly estimated, weakening research that attempts to measure or modify discrimination based on color. Here, we present a comparison of different methodologies for measuring skin color, testing whether measures designed to increase reliability are accomplishing that goal. We ask participants to rate the same photographs using different measures, and then compare the properties of these different measures in order to assess the impact of measurement changes.

## Common Skin Tone Measures in Social Science

### *Text-based measures*

Early measures of skin tone in modern survey-based social science generally asked interviewers to rate respondents' skin tones using a Likert-scale style measure. The most frequently used strategy to measure skin tone in social science studies for many years was asking interviewers to rate respondents on 3, 5 or 7-point continua (e.g., categorizing respondents as very dark, dark, medium, light, or very light). The text-based measurement methods are often critiqued for lacking a way to benchmark the responses of various interviewers (i.e. to make sure everyone is thinking of the same shade when they select "very dark"). Nonetheless, measures like these have been used across a huge range of studies. Some, like the National Longitudinal Study of Adolescent Health, have the interviewer rate the respondent's skin tone. Others, like the Detroit Area Study and the National Survey of American Life, have both the interviewer and the respondent rate their skin tone. Interestingly, discrepancies between interviewer ratings and self-ratings are less common among the more educated (Uzogara et al. 2014).

### *Palette-based measures*

Many skin tone measures today have introduced palettes, or graphics that show a range of skin tones and assign them to values. These graphics are used to train interviewers, with the hope that all of the interviewers using the same graphic will create a norm for each assigned value, increasing reliability and reducing the variation in results across different types of observers, compared to measures that simply use text like "very dark" to "norm" the observers' responses. Graphic-based skin color scales (using a palette that shows a range of colors) are growing in popularity today, such as the Massey and Martin (2003) scale showing images of 10 hands ranging from light to dark (see Appendix A).

The various graphic-based methods have also been critiqued, however, for only including certain types of skin tone (for example, emphasizing a White-to-Black continuum that does not adequately represent Asian American skin tones, or providing a wide range of light-skinned examples and few dark-skinned examples). There also is concern that these measures might not be reliable and consistent across observers. For example, Hannon and DeFina (2014) found that interviewers from different racial backgrounds categorize respondents differently even when interviewers are provided a color guide to standardize their responses. There is some evidence from other countries to suggest that despite the subjective nature of these interviewer assessments, interviewers assess the same respondent similarly over time (Villarreal 2010), but this reliability has not been tested in the U.S. in the same way.

*Light reflectance measures*

Another popular measure, especially in public health, is a reflectance spectrophotometry measure of skin pigmentation. In order to reduce variation in skin tones that results from the season (e.g., more sun exposure in summer than winter), these measurements are often taken on underarm skin. Major studies like the CARDIA data have used these measures (e.g. Borrell et al. 2013; Krieger, Sidney, and Coakley 1998). Although there are advantages of this measurement style for creating year-round consistency in the data, there is considerable debate about whether or not using such a measure captures the ways in which skin color is socially meaningful, since others *are* seeing your face, so sun exposure might have real effects on those perceptions that are not captured in underarm measures. One study in Puerto Rico found, for example, that social classification into color categories was related to blood pressure, but skin pigmentation measures were not (Gravlee, Dressler, and Bernard 2005).

## Data Collection and Method

In this experimental study, we test the strengths and weaknesses of the most popular methods of measuring skin color. In particular, we test: 1) how reliable the ratings of a single person's face are across different observers using each measure; and 2) how similarly Whites and Latinxs categorize faces with a broad range of skin tones using each measure.

### *Photographs*

To collect photographs of models with a wide range of skin tones, at the beginning of the project we took pictures of Texas A&M graduate students between the ages of 22 and 30.[2] We contacted graduate student organizations and members of the Africana Studies certificate program. All pictures were taken in one location with the same background, consistent lighting and similar attire. Models were paid $20 for their time. We hired a professional artist to alter the resulting set of sixteen pictures to have a range of skin tones; we used this manipulation to ensure we covered a full range of skin tones for both male and female models. Each photo had up to five versions total: the original photograph, plus up to four with altered skin tone. We chose only photos that a panel of four viewers agreed looked realistic, discarding any that were seen as obviously altered. Each participant (rater) viewed one randomly selected photo of each model, so each rater viewed a series of 16 pictures, each of a different person. See Appendix B for an example photograph.

### *Skin Tone Measures*

See Appendix A for all three skin tone measures tested in this study. We tested two popular scales used extensively in large surveys today (the 5-point text scale and the 10-point

[2] We did not recruit any graduate students who had taught classes, in order to ensure that undergraduate participants would not recognize the graduate students from class. We also asked after the experiment was completed if the rater recognized anyone from the pictures. No participants reported having recognized anyone from the photographs used.

graphical scale). We also created a new skin tone measure to address common critiques of current graphic-based scales. The skin tone measure we designed is based on the makeup gradient designed by L'Oréal, supplemented with Fashion Fair Fast Finish® Stick Foundation in order to add greater range to the darker skin tones. The L'Oréal gradient, which includes 66 skin shades (and to which we added 3), was created based on their research measuring the skin tone of women around the world and designed to capture a range of color undertones as well as a broader spectrum of color.

In addition to these measures, we also collected reflectance spectrophotometry data for the sixteen individuals, but we do not discuss those results here, because this study focuses on both altered and unaltered photographs, making the light reflectance results less relevant for this particular analysis.

*Raters*

We recruited Latinx and White undergraduate raters from Texas A&M University's campus to come to the Hysom Social Psychology Lab at Texas A&M University and view photographs on a set of iPads with standardized display settings. Recruiters visited large classes and student organizations to recruit a subject pool for a range of studies being conducted at the time, all of which offered compensation. We selected Whites and Latinxs for this experiment because their populations on campus are large, they represent the two largest racial and ethnic groups in the United States, and past research has hypothesized that members of ethnic minority groups will perceive more variation in skin tone than Whites will. Data collection is ongoing this semester, in order to increase the size of the Latinx sample.

In our study, we first asked raters to fill out a self-identification questionnaire. We asked raters their racial/ethnic self-identification; skin color self-identification; the amount of contact they

have had with members of different racial and ethnic groups in different domains of interaction (e.g. school, work, home, etc.); and demographic characteristics including educational status, family income, gender, age, and place of birth for themselves and their parents.

Each rater then classified their 16 photographs (randomly selected from the available photographs, so that they saw one of each model) using all three skin tone measurement scales: a popular 5-item text-based skin tone scale, a widely-used 10-point graphically-based scale (Massey and Martin 2003), and the 69-option grid-based scale we developed for this study. See Appendix A for all three skin tone measures. Pictures were displayed to raters in a randomized order. Raters were paid $15 for their time.

## Results

### *Descriptives*

As Table 1 shows, non-Latinx Whites are the largest response group, and the sample is dominated by women. Most of the sample were born in the United States and have two parents born in the United States, but as Table 2 shows, there are difference in the ratings that immigrants assign and the ratings that the native-born assign. More than 10 percent of the sample identify as lesbian, gay, bisexual or trans*, and almost all of our respondents are in their second year of college or later. The students come from a wide range of majors, with 31 percent in Sociology or Psychology and the other 69 percent distributed across a wide range of majors. About one-third of the sample are first-generation college students, another third come from a family where at least one parent has a bachelor's degree, and another third come from a family where at least one parent completed a postgraduate degree.

Table 1. Characteristics of the sample, N=107

| | Proportions |
|---|---|
| Female | 0.79 |
| U.S. born | 0.81 |
| Parents both U.S. born | 0.75 |
| LGBT | 0.13 |
| Race | |
| White | 0.63 |
| Latinx | 0.18 |
| White and Latinx | 0.07 |
| Multiracial and Other | 0.12 |
| Year in school | |
| First | 0.01 |
| Sophomore | 0.33 |
| Junior | 0.43 |
| Senior | 0.23 |
| Major | |
| Psychology | 0.16 |
| Sociology | 0.15 |
| Allied Health | 0.09 |
| All others | 0.60 |
| Highest level of parental education | |
| Less than a high school degree | 0.07 |
| HS diploma or some college | 0.25 |
| Bachelor's degree | 0.37 |
| Graduate degree | 0.31 |

We hypothesized that having more contact with other racial groups should help individuals perceive more variation within racial groups, rather than seeing them as a homogenous outgroup. We tested their level of contact with other groups before college by asking respondents to describe the racial makeup of a typical class for them at age 15. Table 2 shows significant differences in this by race. The average White rater said that they attended classes that were roughly 70 percent White, while the average Latinx rater estimated that their classes were 35 percent White. The average White rater estimated their classes were 20 percent Latinx, while the average Latinx rater estimated that more than half their peers were Latinx. These numbers suggest a surprising amount of diversity in high school classes, but note that 87

percent of the respondents reported that they lived in Texas at age 15. Texas has roughly equal size White and Latinx populations (about 40 percent for each),[3] making the school contexts more diverse than those found in many other parts of the country. Of course, it is also likely that respondents' memories are less accurate for this type of question, so we also asked respondents to report the race, gender, and age of their four closest friends. Almost ¾ of Whites named a White friend as their closest friend, while about half of Latinx respondents named a Latinx closest friend.

Table 2. Raters' Teenage Social Context, by Race

| | White raters | Latinx raters | All other raters |
|---|---|---|---|
| At age 15, my typical class was… | | | |
| …percent White | 69 | 35 | 33 |
| …percent Latinx | 20 | 51 | 17 |
| …percent Black | 11 | 8 | 12 |
| …percent other race | 7 | 5 | 30 |

Each photograph had a minimum of 17 raters and a maximum of 53.

*Analytic results*

Fixed-effects models for the ratings from the two popular scales (the 10-point graphic scale and the 5-point text scale) and the column number from the grid we created (columns indicating how dark the person is, and rows indicating the undertone of the skin) show that the the relationships between the social backgrounds of the observer and their ratings of the pictures are similar for all three scales. (All scales are coded so that higher numbers refer to darker skin.) Table 3 shows, for example, that women rate pictures as darker with all scales (although the finding is less robust with the grid measure, where the relationship is only significant in some model specifications). Immigrants rate the pictures as darker as well, across all three scales.

[3] https://www.census.gov/quickfacts/fact/table/tx/PST045217

Latinx raters, White/Latinx raters, and our heterogeneous other race category classify pictures as lighter than Whites do. Those who attended a school at age 15 with a greater percentage of Whites also rate pictures as significantly lighter. Interestingly, how long the students have been in college (and therefore how much of the college environment they have been exposed to) does not relate to their ratings on any scale, but having educated parents relates to darker classifications on the grid scale. Those who see themselves as having darker skin relative to their own group also rate others as darker when they are using the 10-point palette scale, but not when they use the other two scales.

Table 3. Fixed effects models of rater characteristics

| | 10-point graphic scale | 5-point text scale | Grid column |
|---|---|---|---|
| Female | 0.328*** | 0.061* | 0.118 |
| | (0.056) | (0.032) | (0.073) |
| U.S. born | -0.413*** | -0.106*** | -0.348*** |
| | (0.060) | (0.034) | (0.077) |
| White (ref.) | | | |
| Latinx | -0.160** | -0.071* | -0.398*** |
| | (0.069) | (0.039) | (0.089) |
| White/Latinx | -0.139* | 0.006 | -0.310*** |
| | (0.084) | (0.048) | (0.109) |
| Other race | -0.739*** | -0.259*** | -0.856*** |
| | (0.072) | (0.041) | (0.093) |
| Junior/senior | 0.004 | 0.018 | 0.046 |
| | (0.048) | (0.028) | (0.062) |
| Parents BA+ | 0.028 | -0.005 | 0.260*** |
| | (0.051) | (0.029) | (0.066) |
| Percent school White | -0.003*** | -0.001** | -0.004*** |
| | (0.001) | (0.000) | (0.001) |
| Skin tone (5=Very Dark) | 0.050* | 0.007 | 0.003 |
| | (0.030) | (0.017) | (0.039) |
| Constant | 4.711*** | 2.789*** | 6.458*** |
| | (0.127) | (0.073) | (0.166) |
| Observations | 1,660 | 1,673 | 1,674 |
| R-squared | 0.087 | 0.031 | 0.072 |
| Number of Photos | | 60 | |

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

## Discussion and Conclusions

Examining *rho* for these fixed effects models shows that the *fraction* of the variance due to differences *within* the ratings of each separate picture are greater for the 10-point scale and the grid column rating than the 5-point text scale, so results do not support the idea that there is reduced rater variability when graphics are used to create a "norm" across raters. The next step will be to calculate inter-rater reliability to confirm this result with a measure that has better

properties for comparing scales with different ranges, once the new data collected in March 2019 is added to the existing data. We will calculate Krippendorff's alpha (Hayes and Krippendorff 2007) for each measure. Krippendorff's alpha has significant advantages for this comparison (for example, it generalizes across different levels of measurement and it discounts the amount of agreement across raters that occurs simply by chance).

**References**

Allen, Walter, Edward Telles, and Margaret Hunter. 2000. “Skin Color, Income and Education: A Comparison of African Americans and Mexican Americans.” *National Journal of Sociology* 12(1):129–80.

Bailey, Stanley, Aliya Saperstein, and Andrew Penner. 2014. “Race, Color, and Income Inequality across the Americas.” *Demographic Research* 31:735–56.

Bond, Selena and Thomas F. Cash. 1992. “Black Beauty: Skin Color and Body Images among African-American College Women.” *Journal of Applied Social Psychology* 22(11):874–88.

Borrell, Luisa N., C. I. Kiefe, A. V Diez-Roux, D. R. Williams, and P. Gordon-Larsen. 2013. “Racial Discrimination, Racial/Ethic Segregation, and Health Behaviors in the CARDIA Study.” *Ethnicity & Health* 18(3):227–43.

Campbell, Mary E., Jenifer L. Bratter, and Wendy D. Roth. 2016. “Measuring the Diverging Components of Race: An Introduction.” *American Behavioral Scientist* 60(4):381–89.

Caruso, E. M., N. L. Mead, and E. Balcetis. 2009. “Political Partisanship Influences Perception of Biracial Candidates’ Skin Tone.” *Proceedings of the National Academy of Sciences* 106(48):20168–73.

Drake, St Clair and Horace R. Cayton. 1962. *Black Metropolis: A Study of Negro Life in a Northern City*. 2nd ed.

Dressler, William W. 1993. “Health in the African American Community: Accounting for Health Inequalities.” *Medical Anthropology Quarterly* 7(4):325–45.

Dressler, William W. 1991. “Social Class, Skin Color, and Arterial Blood Pressure in Two Societies.” *Ethnicity & Disease* 1(1):60–77.

Frank, Reanne, Ilana Redstone Akresh, and Bo Lu. 2010. “Latino Immigrants and the U.S.

Racial Order: How and Where Do They Fit In?" *American Sociological Review* 75(3):378–401.

Gravlee, Clarence C., William W. Dressler, and H. Russell Bernard. 2005. "Skin Color, Social Classification, and Blood Pressure in Southeastern Puerto Rico." *American Journal of Public Health* 95(12):2191–97.

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.

Hayes, Andrew F. and Klaus Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1(1):77–89.

Hersch, Juni. 2006. "Skin-Tone Effects among African Americans: Perceptions and Reality." *American Economic Review* 96(2):251–55.

Hunter, Margaret L. 2002. "'If You're Light You're Alright' Light Skin Color as Social Capital for Women of Color." *Gender & Society* 16(2):175–93.

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

Keith, Verna M. and Cedric Herring. 1991. "Skin Tone and Stratification in the Black Community." *American Journal of Sociology* 97(3):760.

Keith, Verna M. and Maxine S. Thompson. 2003. "Color Matters: The Importance of Skin Tone for African American Women's Self-Concept in Black and White America." Pp. 116–35 in *In and Out of Our Right Minds: The Mental Health of African American Women*, edited by D. R. Brown and V. M. Keith.

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.

Krieger, N., S. Sidney, and E. Coakley. 1998. “Racial Discrimination and Skin Color in the CARDIA Study: Implications for Public Health Research. Coronary Artery Risk Development in Young Adults.” *American Journal of Public Health* 88(9):1308–13. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/9736868).

Massey, Douglas S. and Jennifer A. Martin. 2003. “The NIS Skin Color Scale.” Retrieved August 3, 2015 (http://nis.princeton.edu/downloads/NIS-Skin-Color-Scale.pdf).

Monk, Ellis P. 2014. “Skin Tone Stratification among Black Americans, 2001–2003.” *Social Forces* 92(4):1313–37.

Ross, Louie E. 1997. “Mate Selection Preferences Among African American College Students.” *Journal of Black Studies* 27(4):554–69.

Saperstein, Aliya. 2012. “Capturing Complexity in the United States: Which Aspects of Race Matter and When?” *Ethnic and Racial Studies* 35(8):1484–1502.

Uzogara, Ekeoma E., Hedwig Lee, Cleopatra M. Abdou, and James S. Jackson. 2014. “A Comparison of Skin Tone Discrimination among African American Men: 1995 and 2003.” *Psychology of Men & Masculinity* 15(2):201–12.

Vargas, Nicholas. 2015. “Latina/o Whitening? Which Latina/Os Self-Classify as White and Report Being Perceived as White by Other Americans?” *Du Bois Review: Social Science Research on Race* 12(01):119–136.

Villarreal, Andrés. 2010. “Stratification by Skin Color in Contemporary Mexico.” *American Sociological Review* 75(5):652–78.

Weaver, Vesla M. 2012. “The Electoral Consequences of Skin Color: The ‘Hidden’ Side of Race in Politics.” *Political Behavior* 34:159–92.

## Appendix A. Skin tone measures

### *Text-based 5-point scale*

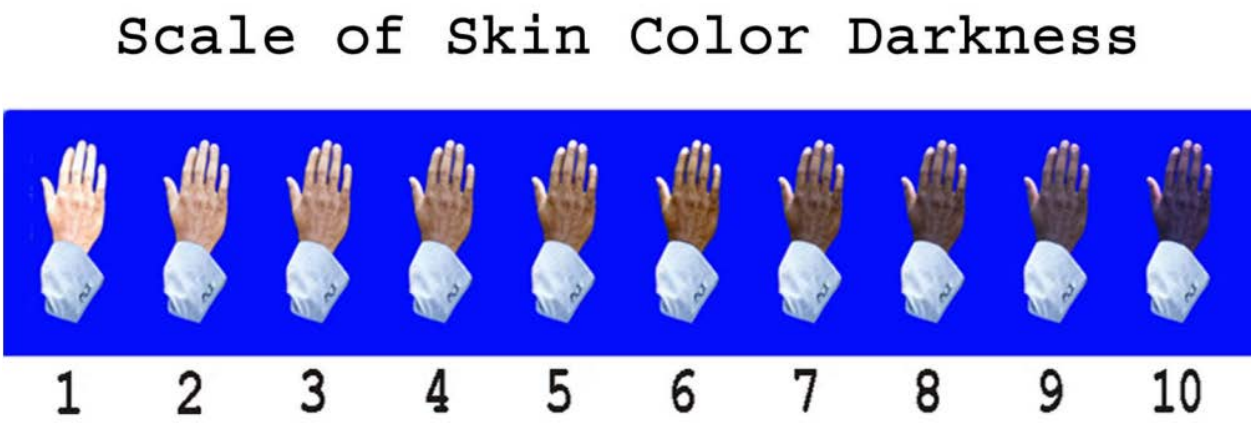
Probably the most commonly used scale in the social sciences is the text-based scale, which asks interviewers to rate the respondent's skin color on a scale from very light to very dark. We asked participants to rate the photographs on this scale:

The subject's skin color is:

1. Very Dark
2. Dark
3. Medium
4. Light
5. Very Light

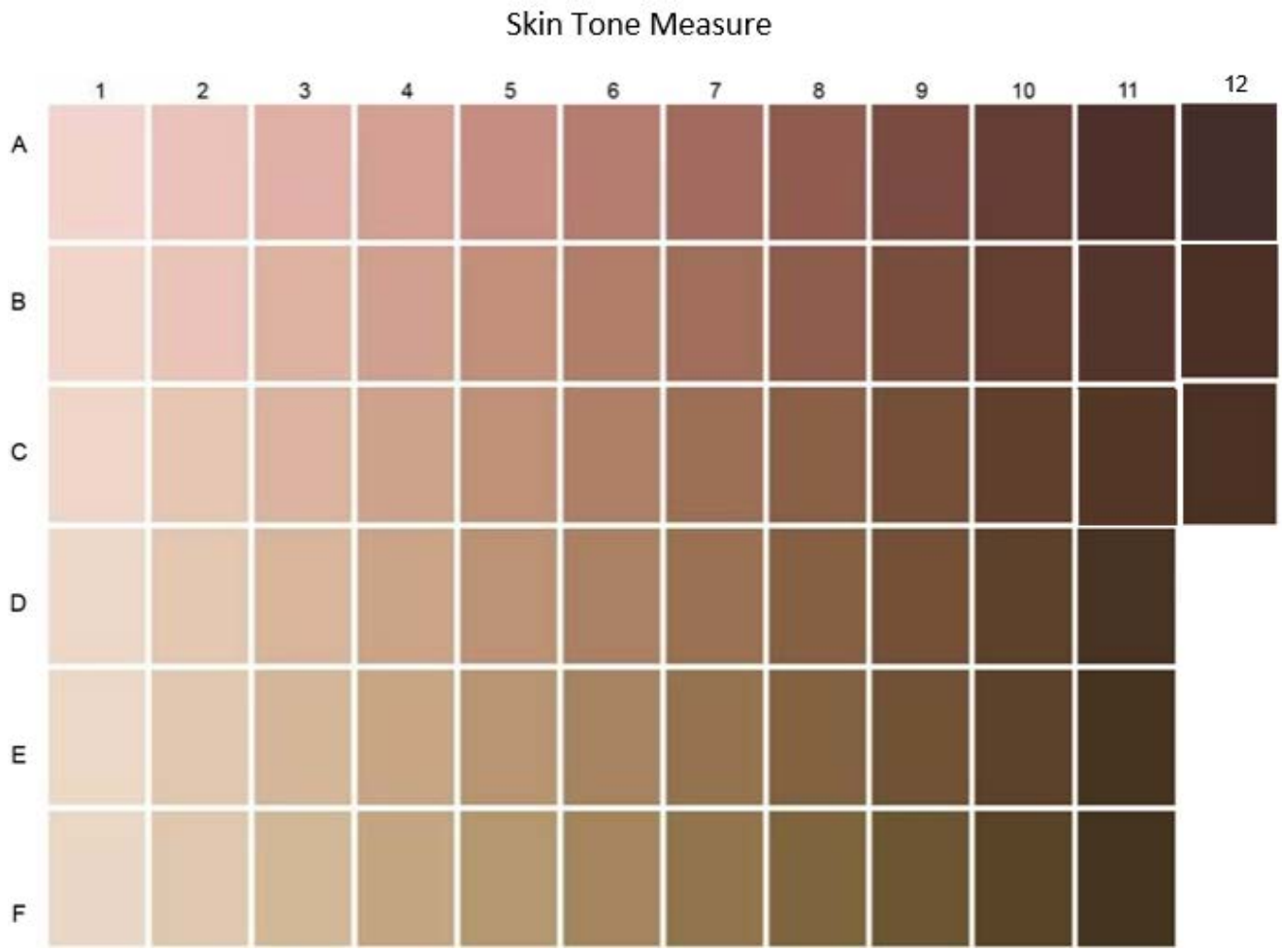
### *Graphic-based 10-point scale*

Another widely used scale is the 10-point Massey and Martin (2003) scale, which asks interviewers to classify the skin tone of every respondent using this graphic (which was to be memorized and never to be shown to the people they were classifying) as a guide:




### *Graphic-based 69-point scale*

This scale also asked the respondents to classify each person based on a graphic of skin tones. This measure was adapted from L'Oréal's grid of skin tones based on their collection of measurements from around the world (http://www.loreal.com/research-and-innovation/when-the-diversity-of-types-of-beauty-inspires-science/expert-in-skin-and-hair-types-around-the-world), supplemented with shades from http://shop.fashionfair.com/ProductDetails.asp?ProductCode=FAST+FINISH+FOUNDATIONS in order to broaden the range of darker skin tones available. The lettered rows indicate the individual's undertone, which varies from red (A) to yellow (F). The numbered columns vary in skin tone darkness, with the lightest colors on the left (1) and the darkest (12) on the right. Respondents were instructed to choose the cell they believe best matches the skin color of the individuals photographed (e.g. "E5").


Skin Tone Measure
1
2
3
4
5
6
7
8
9
10
11
12
A
B
C
D
E
F

**Appendix B: Sample photograph**