Citation and Link to Published Version (Open Source): Gordon, Rachel A., Amelia R. Branigan, Mariya Adnan Khan, and Johanna G. Nunez. 2024. "Best Practices for Measuring Skin Color in Surveys." Survey Practice 17 (May). https://doi.org/10.29115/SP-2024-0005

Title Best Practices for Measuring Skin Color in Surveys

Authors:

Rachel A. Gordon, College of Health and Human Sciences, Northern Illinois University
Amelia R. Branigan, Department of Sociology, University of Maryland, College Park
Mariya Adnan Khan, Department of Sociology, University of Illinois at Chicago
Johanna G. Nunez, Department of Sociology, University of Wisconsin, Madison

Introduction

Most U.S. based surveys assess racial-ethnic identities and are increasingly asked to better capture skin color as an aspect of racialized appearance (Telles 2018). Such survey data can importantly inform how skin color relates to social and health outcomes (Adams et al. 2016; Dixon and Telles 2017). Doing so depends on reliable measurement of skin color, however. The typical approach to skin color measurement in survey data has been interviewer or respondent ratings using categorical skin color scales (e.g., Campbell et al. 2020). The potential for using mechanical instruments to assess skin color has grown as handheld devices have become increasingly affordable and user-friendly (e.g. Gordon et al. 2022). We compared these two strategies for skin color measurement— a) ***handheld devices*** and b) ***rating scales***— offering empirical findings and practical guidance for future survey efforts to collect skin color data.

Handheld devices, including colorimeters and spectrophotometers, measure color via light reflectance. Historically, such instruments were used primarily by bench scientists in biology and chemistry fields because they were too expensive and too large and delicate for easy transportation outside of laboratory settings. These instruments are now small and inexpensive enough to be feasible for a wide range of in-person survey contexts. Handheld devices measure consistently across varying lighting conditions, but technical settings, such as the size of the opening (*aperture*) through which light passes, can affect readings. Survey methodologists need evidence regarding: a) the reliability of new low-cost devices relative to well-established yet larger and more expensive devices, b) where on the body and how to take color readings using these devices, and c) whether and how field staff can be effectively trained to use the devices.

The current study builds on prior research by comparing three devices, examining how consistent their readings are across *repeated measures* at *four locations* (forehead, cheek, inner arm, outer arm) and varying *technical settings* (size of aperture for light transmission; simulated lighting conditions). In prior work (Gordon et al. 2022) we compared two devices at a single location with a single device setting. The new results offer comparison with a more sophisticated and expensive instrument certified to perform at industry standards for reliability and validity (Konica Minolta 2007). The new results also inform survey methodologists about where on the body to take readings with what device settings. Results are also translated into practical guidance, including lessons learned for creating measurement protocols and training staff.

Skin color rating scales build on a long tradition asking people to select from images (e.g., colored porcelain tiles) or words (e.g., lightest, lighter, darker, darkest). Scales developed based on color science emerged only recently, however. The widely-used *Massey-Martin scale* (Massey and Martin 2003) was developed in the early 2000s for interviewers to rate participants' skin shade (lightness-darkness) and has been used in many large surveys (see Figure 1). The *L'Oreal scale* was developed for the cosmetics industry using color science (de Rigal et al. 2007) and has since been used in surveys (e.g., Campbell et al. 2020; Garcia and Abascal 2016; Khan et al. 2023).

The current study builds on prior research by offering evidence regarding how reliably interviewers and participants can choose from the more numerous L'Oreal versus fewer Massey-Martin options. In prior work (Gordon et al. 2022) we compared the Massey-Martin with another scale, the PERLA. The new results are important because the Massey-Martin and the PERLA offer ten or eleven choices arrayed along a single dimension primarily reflecting lightness-to-darkness. In contrast, the L'Oreal offers sixty-six choices arrayed along eleven levels of lightness-to-darkness each within six levels of redness-to-yellowness. Consideration of the L'Oreal choices for skin undertone (redness, yellowness) is important given skin undertone has been less studied than skin darkness. Adding to a recent study examining undertone in photographs (Branigan et al. 2023)[1], we compared in-person human ratings of undertone to handheld device readings of redness and yellowness. The current study also extends prior results by using a specialized room with equalized conditions such as lighting.

Method

Sample. Undergraduate students (n = 102) were recruited through flyers, emails, and class visits by pairs of undergraduate research assistants. Consistent with the university's designation as an Asian American Native American Pacific Islander and Hispanic Serving Institution, the majority of study participants identified as Asian (54%) and about one-fifth each identified as Latinx (23%) and White (18%); 3% each identified as Black and Other race-ethnicities (see Table 1). Over two-thirds of participants identified as Cisgender Woman (71%), over one-quarter as Cisgender Man (27%); two participants identified as Transgender Man and Nonbinary. Most participants were ages 18 to 21 (12% to 30% each single age). Study team members also represented multiple genders and race-ethnicities, including Black, White, Latino, and Asian.

Procedures. A dedicated room equalized background and lighting conditions for each participant's ratings. In consultation with a color measurement expert, we selected an interior room (to reduce temperature fluctuation), determined the appropriate number (four) and placement of luminaire lighting fixtures for the room size and shape, and selected grey paint

[1] Branigan and colleagues (2023) drew upon prior theory and research to conceptualize the importance of undertone for colorism research. Skin redness and yellowness, for instance, can be perceived as signals of attractiveness and health, although these colors' momentary fluctuations due to emotions, diet, and sleep may mean that their social signaling is less stable than is skin's darkness. Color science assessments of skin color commonly use dimensions of darkness-lightness, greenness-redness, and blueness-yellowness. Measured values for skin color fall within the red and yellow ranges of the latter dimensions.

color, furniture, and covering for participants' clothing. The luminaires simulated outdoors mid-day light during data collection. Participants sat in a chair at a desk across from two interviewers.

Handheld devices are based on color science which aims to understand and replicate how humans see color. One widely used color space separates three dimensions of darkness-lightness (L*), greenness-redness, (a*), and blueness-yellowness (b*). The devices operate by emitting light out of a small opening, placed flush against the area of measurement, and recording the light reflected by the object. *Spectrophotometers* capture the full light spectrum whereas *colorimeters* focus on certain wavelengths. The instruments are used in a range of applications from house painting to constructing craniofacial prosthetics.

An advantage of handheld devices for survey methodologists is consistency of measurement across many conditions—e.g., a single reading can simulate various *illuminants* (lighting conditions) from *outdoors mid-day* (known as *D65*), to *outdoors sunrise/sunset* (known as *D50*), to *indoors incandescent* (known as *A*). [2] Formulas translate to various illuminants using recorded values that reference industry standard "black" (no light) and "white" (all wavelengths visible to humans). Technical features can affect readings, however, and each device uses a somewhat different design, often proprietary. Some devices allow for changing technical settings, such as the size of the aperture letting light pass through. Our protocol used two devices' options to compare aperture size and lighting conditions (see Appendix A).

The first of three devices we used is a spectrophotometer from the commercial company *Konica Minolta*. The Konica Minolta *CM-700d* has been widely used for a range of applications yet is expensive and cumbersome to maneuver due to size. At the time of our study, the device cost about $14,000, was just over 8 inches tall, and weighed about 1 pound. We compared two aperture options which could be readily toggled (a larger aperture, labelled *MAV*; a smaller aperture, labelled *SAV*). For survey purposes, the device is sturdy with a built-in screen, easy-to-use calibration checks, and computer connected software to take and export multiple readings at once. Yet, at the time of our study, the device required wired connection to a computer and required wall plugin when batteries ran low.

We also considered two less expensive and smaller devices. *Nix*, like Konica Minolta, is a commercial company. Nix has specialized in small colorimeters intended for everyday use in painting and design (a spectrophotometer is now also available). The Nix device had no built-in screen, but was sturdily encased to resist regular fall damage and worked wirelessly with a user-friendly smart phone app. The device was inexpensive, small, and light. We used a $349 Nix *Pro 2* approximately 2 x 2 inches in size and weighing 1.5 ounces. The device arrived pre-calibrated and reliability-tested but possessed no built-in features for users to run calibration checks.

The *Labby* spectrophotometer was developed for low-resource contexts with open-source specifications and readily purchased components, including assembly in a 3D printed case. The

[2] Color science aims to understand and reproduce the ways humans see color (Logvinenko and Levin, 2023). One important construct is illumination, the relative intensity of light across the spectrum of wavelengths. How people perceive an object's color depends on its illumination. Various illuminants have been defined to represent different scenarios, such as those listed in the narrative (i.e., outdoors mid-day, known as D65; outdoors sunrise/sunset, D50; and, indoors incandescent, A).

company built the device used in our study, costing about $1,200. The version of Labby we used lacked a built-in screen, had a single aperture, and had limited pre-programmed readings for a single illuminant. The open-source nature of the device made fully transparent the hardware and calculations used to obtain final outputs but presented a steeper learning curve, greater potential for human error, and less protection from accidental damage.

We took readings in the *L*a*b** color space from each device. L* readings can range from 0 to 100, with higher scores indicating lighter skin. For human skin tone, a* and b* values are positive with higher scores indicating darker shades of redness (a*) or yellowness (b*). In our sample, L* (lightness) values ranged from about 25 to 80, averaging around 60 with a standard deviation of about 6 (see Appendix B). The b* (yellowness) values ranged from about 5 to 25, averaging about 16 with a standard deviation of about 3. The a* (redness) values ranged from about 2 to 20 with an average of around 10 and standard deviation of about 2. We had 10 fewer readings from Labby than the other devices due to missing data when we waited for a replacement device. One participant also refused use of the Konica Minolta when informed of the brief flash it would emit during readings. We also recoded to missing one set of outlying L*a*b* readings for three participants for Labby and for one participant for Nix.

The original *Massey-Martin* (2003) *rating scale* included 11 images of lighter to darker colored hands, each with visible cuffs. We followed recent studies using a circular portion extracted from 10 images (see again Figure 1). The 66-color L'Oreal palette was created using color science readings taken from the faces of over 1,000 women worldwide (France, United States, Mexico, Brazil, Japan, Korea, China, Thailand, Africa; de Rigal et al. 2007). L'Oréal scientists selected the colors using color science's definition of the minimum difference that the human eye can detect. The resulting palette includes 11 levels of lightness-darkness and 6 levels of redness-yellowness. Respondents used 8 of the 10 Massey-Martin color swatches, all but the top 2 values (see Appendix B). Respondents used nearly all of the 66 L'Oreal color swatches, with values covering the full range of 1 to 11 for lightness-darkness and all but 1 (reddest) of the 6 levels of redness-yellowness.

Analyses. We considered *absolute agreement* of individual scores (i.e., were two readings or two ratings identical in value?) using the intraclass correlation (*ICC*; Koo and Li 2016). For survey methodologists, absolute agreement is important for studies considering mean differences in skin color. We also presented Pearson correlations (*r*; i.e., were scores higher on one reading/rating when higher on another?), which are important for studies considering correlations of other variables with skin color. We considered ICCs above .60 as *good* and above .75 as *excellent* agreement (Cicchetti 1994; Lance, Butts, and Michels 2006). We used similar guidance for Pearson correlations, which will be equal to or larger than ICCs and also have a shared variation interpretation ($r = .75$ reflects 56% shared variation).

Findings and Implications for Best Practices

We organized key findings around focal questions of interest to survey methodologists.

How many handheld device readings are needed? Additional readings take time, yet that time may be warranted if test-retest reliability is low and extra readings could thus considerably reduce measurement error.

Our results showed that test-retest reliability was excellent (see Table 2). Konica Minolta edged out the other two devices, with its repeated readings nearly identical, especially for lightness (L*) and yellowness (b*). Nix and Labby showed slightly more variation between their repeated readings, and each had some outlying values.

For practice, our results indicated that one reading would generally be sufficient. Given a second reading took little time, however, two readings could protect against the few instances of outlying readings.

To achieve these results in practice, however, training is recommended. Our staff training supported consistent device use, such as about how much pressure to apply and how to avoid skin features such as veins and freckles (notes available from authors). Some of the difference between repeated readings seen for Nix and Labby may also reflect their technical construction. Using a more recently developed Nix attachment may reduce sensitivity of readings to varying pressure applied by field staff during readings.

How do technical settings affect readings? Survey methodologists are faced with many choices for device technical settings, yet, the impact of such choices for measuring skin color has not been well documented to date.

Our findings, shown in Table 3, indicate that these choices matter the least for assessments of lightness (L*). Yet, their impact is somewhat greater for assessing undertones of yellowness (b*) and particularly important for redness (a*). Consistently higher redness values of readings taken with a larger than smaller aperture are illustrated in Figure 2.

In practice, when undertone is focal to a study's research questions, taking readings with multiple technical settings may be advised. Ensuring that surveys' documentation and publications clearly report what settings they used would also facilitate comparisons across studies. Encouraging device manufacturers to be transparent about relevant technical details for scientific communities might also counterbalance their proprietary interests for commercial applications.

How much does body location matter? Medical and anthropological uses of spectrophotometry have long recognized the importance of body location, such as sun-exposed (facultative) and sun-protected (constitutive) skin (Neville, Palmieri and Young 2021). Participants in large scale field surveys may also decline measurements in private body locations.

We documented the importance of body location for skin undertone, in addition to its recognized importance for skin shade. Within face and arm, readings were highly correlated, but differed somewhat in absolute levels, being somewhat lighter (higher L*) on the cheek than forehead and on the inner versus outer arm (Table 4). Readings were also redder and yellower (higher a* and b*) on the outer than inner arm, but more consistent between cheek and forehead. Comparing

cheek and inner arm, although lightness (L*) and yellowness (b*) were fairly consistent, redness (a*) was considerably higher on the cheek. Figure 3 illustrates the consistently redder readings on the outer than inner arm.

In practice, survey methodologists would want to carefully consider the substantive goals of a project when choosing body locations. For example, for questions about implicit bias due to colorism, the forehead location might be chosen as the front of the face is generally visible across day-to-day interactions. For a different question, such as an individual's biochemical vulnerability to seasonal affective disorder, sun-protected skin, such as the inner arm location, might be chosen (e.g., Stewart et al. 2014).

How well do recent handheld devices work relative to well-established yet larger and more expensive devices? Smaller size and lower cost facilitate taking devices into the field when budgets are limited.

Our results showed that although values were highly correlated between readings taken by different devices at the same body location, absolute levels differed (Table 5, Figure 4). Labby tended to produce lighter (L*) and yellower (b*) readings than Konica Minolta. The reverse was true for Nix. Aperture size seemed important, including for the longer wavelengths of redness, as illustrated in Figure 4. The Nix aperture size was closest to Konica Minolta's larger aperture, where consistency was highest (bottom right, Figure 4).

In practice, substantive goals should inform survey methodologists' choices. Studies focused on questions correlating skin color with other variables would expect similar results regardless of device choice. Here, smaller and less expensive devices may be sufficient. Results for absolute levels of skin color would be more sensitive to device choice, including aperture size, warranting more research into when and where these differences matter most.

How do rating scale scores relate to handheld device readings? Collecting both rating scale and device readings increases respondent burden and survey cost, making important evidence about the relative similarity and difference of their scores.

Our findings showed that correlations were considerably higher between device readings and human ratings of skin shade (lightness-darkness) than skin undertone (redness, yellowness; Table 6). Single ratings correlated with darkness nearly as highly as three-rating averages. However, these correlations were somewhat lower for participants than interviewers.

For practice, if skin darkness is the focus, our findings suggest that correlational results would be similar if either a handheld device or a rating scale were used. At the same time, for studies aiming to distinguish how humans assess skin color from its color science calculated value, both human ratings and device readings would be needed. These studies could further examine self and other perceptions by having multiple ratings (including from photographs of participants; Khan et al. 2023). Cognitive interviews might also inform why humans are better at choosing swatches that align with color science calculated skin darkness than its redness or yellowness.

Conclusion

Modern technology offers survey methodologists new options for responding to calls to better capture skin color in surveys (Telles 2018). Our findings document the advantages of using handheld devices to reliably assess skin color, supporting substantive questions about how skin shade and skin undertone affect social inequalities in human health and well-being. We offer guidance to survey methodologists for such uses. In some cases, recommendations are clear—e.g., just one or two device readings at any given location; any device can similarly capture the relative darkness of skin. In other cases, recommendations are less certain—e.g., skin undertones of redness and yellowness being sensitive to device choices and body locations. We encourage future studies that pursue why such variability exists and for which substantive questions it matters most.

Acknowledgments: 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 Dahlya El-Adawe and Hai Nguyen.

References

Adams, Elizabeth A., Beth E. Kurtz-Costes, and Adam J. Hoffman. 2016. "Skin Tone Bias among African Americans: Antecedents and Consequences across the Life Span." *Developmental Review* 40 (June): 93–116. https://doi.org/10.1016/j.dr.2016.03.002.

Branigan, Amelia R., Johanna G. Nunez, Mariya Adnan Khan, and Rachel A. Gordon. 2023. "Variation in Skin Red and Yellow Undertone: Reliability of Ratings and Predicted Relevance for Social Experiences." *Social Psychology Quarterly*. OnlineFirst. https://doi.org/10.1177/01902725231196851

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. https://doi.org/10.1007/s12552-020-09294-0.

Cicchetti, Domenic V. 1994. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment* 6 (4): 284–90. https://doi.org/10.1037/1040-3590.6.4.284.

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. https://doi.org/10.1111/j.1600-0846.2007.00223.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. https://doi.org/10.1146/annurev-soc-060116-053315.

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

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* 10 (2): 337–64. https://doi.org/10.1093/jssam/smab046.

Khan, Mariya Adnan, Hai Nguyen, Amelia R. Branigan, and Rachel A. Gordon. 2023. "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." *Research in Human Development* [Online first]: 1–17. https://doi.org/10.1080/15427609.2023.2224318.

Konica Minolta. 2007. *Spectrophotometer CM-700d/600d: Instruction Manual*. Available at https://sensing.konicaminolta.us/wp-content/uploads/cm-700d_instruction_eng-54pn5p743t.pdf

Koo, Terry K., and Mae Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

Lance, Charles E., Marcus M. Butts, and Lawrence C. Michels. 2006. "The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?" *Organizational Research Methods* 9 (2): 202–20. https://doi.org/10.1177/1094428105284919.

Logvinenko, Alexander D., and Vladimir L. Levin. 2023. *Foundations of Colour Science: From Colorimetry to Perception.* Chichester, UK: Wiley.

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

Neville, Jonathan J., Tommaso Palmieri, and Antony R. Young. 2021. "Physical Determinants of Vitamin D Photosynthesis: A Review." *JBMR Plus* 5 (1): e10460-n/a. https://doi.org/10.1002/jbm4.10460.

Stewart, Alan E., Kathryn A. Roecklein, Susan Tanner, and Michael G. Kimlin. 2014. "Possible Contributions of Skin Pigmentation and Vitamin D in a Polyfactorial Model of Seasonal Affective Disorder." *Medical Hypotheses* 83 (5): 517–25. https://doi.org/10.1016/j.mehy.2014.09.010.

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. https://doi.org/10.1177/0002716218766463.

Table 1. Characteristics of study participants (n = 102)

		N	%
Race-Ethnicities	White	18	18
	Asian	55	54
	Latinx	23	23
	Black	3	3
	Other	3	3
Sex-Genders	Cisgender Man	28	27
	Cisgender Woman	72	71
	Transgender Man	1	1
	Nonbinary	1	1
Age in Years	18	29	28
	19	31	30
	20	13	13
	21	12	12
	22+	16	16
	Missing	1	1
Undergraduate Major	Life Sciences and Health	31	30
	Social Sciences	22	22
	Business, Engineering, and Computer Science	29	28
	Humanities	1	1
	Undeclared	19	19
Total		102	100

Table 2. Intraclass (ICC) and Pearson (r) correlations for repeated device readings of the cheek

Device	Lightness (L*)	Yellowness (b*)	Redness (a*)
Konica Minolta	1.00 (.99, 1.00)	1.00 (.99, 1.00)	.98 (.97, .99)
	[r = 1.00]	[r = 1.00]	[r = .98]
Nix	.98 (.96, .98)	.98 (.97, .99)	.94 (.91, .96)
	[r = .98]	[r = .98]	[r = .94]
Labby	.95 (.93, .97)	.97 (.95, .98)	.92 (.88, .95)
	[r = .95]	[r = .97]	[r = .92]

ICC confidence intervals in parentheses.

Values above .60 considered *good;* above .75 *excellent*

Table 3. Intraclass (ICC) and Pearson (r) correlations for aperture and illuminant settings in readings of the cheek

Device		Lightness (L*)	Yellowness (b*)	Redness (a*)
Konica Minolta	(Larger vs smaller aperture)	.96 (.85, .99)	.90 (.42, .96)	.65 (-.05, .87)
		[r = .98]	[r = .95]	[r = .84]
Nix	(Noon daylight vs incandescent)	.93 (.90, .95)	.96 (.94, .97)	.87 (.82, .91)
		[r = .93]	[r = .96]	[r = .88]

ICC confidence intervals in parentheses.

Values above .60 considered *good;* above .75 *excellent*

Table 4. Intraclass (ICC) and Pearson (r) correlations for body locations using Konica Minolta larger aperture readings

		Lightness (L*)	Yellowness (b*)	Redness (a*)
Within Face	Cheek vs forehead	.79 (.07, .93)	.74 (.54, .84)	.56 (.40, .68)
		[r = .90]	[r = .78]	[r = .60]
Within Arm	Inner vs outer arm	.79 (.00, .93)	.61 (.07, .82)	.52 (.01, .76)
		[r = .92]	[r = .76]	[r = .68]
Face vs Arm	Cheek vs inner arm	.79 (.54, .89)	.67 (.55, .84)	.06 (-.05, .20)
		[r = .85]	[r = .68]	[r = .22]

ICC confidence intervals in parentheses.

Values above .60 considered *good;* above .75 *excellent*

Table 5. Intraclass (ICC) and Pearson (r) correlations for Konica Minolta (larger aperture) vs Nix (Noon daylight) and Labby readings of the cheek

Konica Minolta	Comparison	Lightness (L*)	Yellowness (b*)	Redness (a*)
Smaller aperture	Nix	.84 (.09, .95)	.79 (.00, .93)	.67 (.15, .85)
		[r = .93]	[r = .92]	[r = .80]
Smaller aperture	Labby	.82 (.66, .90)	.72 (-.06, .92)	.19 (-.05, .52)
		[r = .89]	[r = .94]	[r = .72]
Larger aperture	Nix	.76 (-.05, .92)	.64 (-.07, .89)	.76 (.67, .83)
		[r = .92]	[r = .92]	[r = .77]
Larger aperture	Labby	.85 (.77, .90)	.84 (.12, .95)	.34 (-.09, .68)
		[r = .89]	[r = .94]	[r = .73]

ICC confidence intervals in parentheses.

Values above .60 considered *good;* above .75 *excellent*

Table 6. Pearson (r) correlations of average ratings and Konica Minolta Larger Aperture device reading of the cheek

Dimension	Rating Scale	Average	1st Interviewer	2nd Interviewer	Participant
			Lightness (L*)		
Lightness-to-Darkness	Massey-Martin	[r = -.89]	[r = -.84]	[r = -.87]	[r = -.76]
Lightness-to-Darkness	L'Oreal	[r = -.85]	[r = -.81]	[r = -.80]	[r = -.76]
			Yellowness (b*)		
Redness-to-yellowness	L'Oreal	[r = .25]	[r = .20]	[r = .19]	[r = .10]
			Redness (a*)		
Redness-to-yellowness	L'Oreal	[r = -.30]	[r = -.24]	[r = -.08]	[r = -.24]

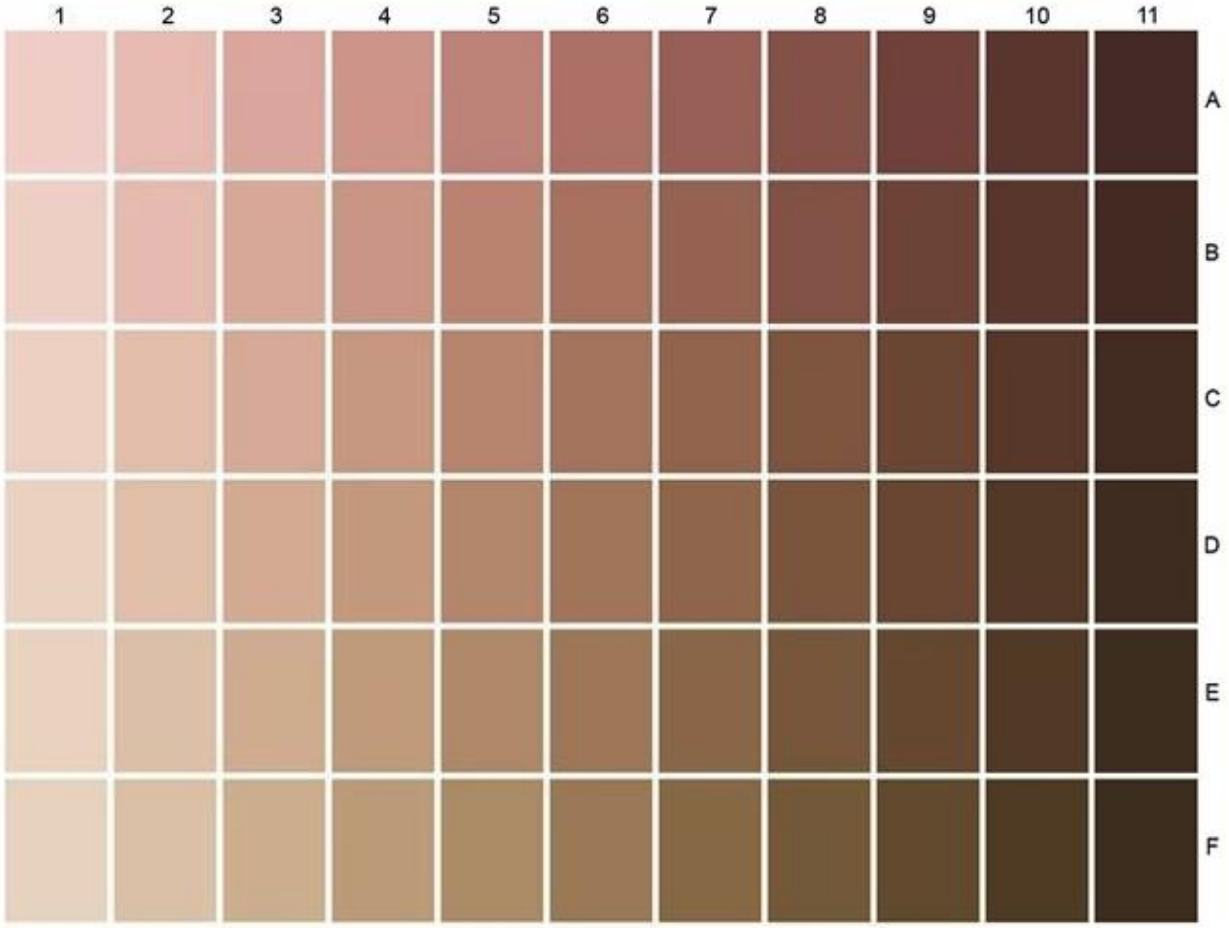
Average = average of 1st interviewer, 2nd interviewer, and participant ratings.

Values above .60 considered *good;* above .75 *excellent*

Figure 1. Rating Scales

a) Massey Martin

b) L'Oreal



Source: Massey & Martin (2003). De Rigal et al. (2007). Since its development 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). We programmed the 66 L'Oreal shades using the Qualtrics hot spot question type to consider its potential for similar use in large scale surveys. See also: https://nis.princeton.edu/downloads/nis-skin-color-scale.pdf; https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

Figure 2. Illustration of redder readings with larger vs smaller aperture

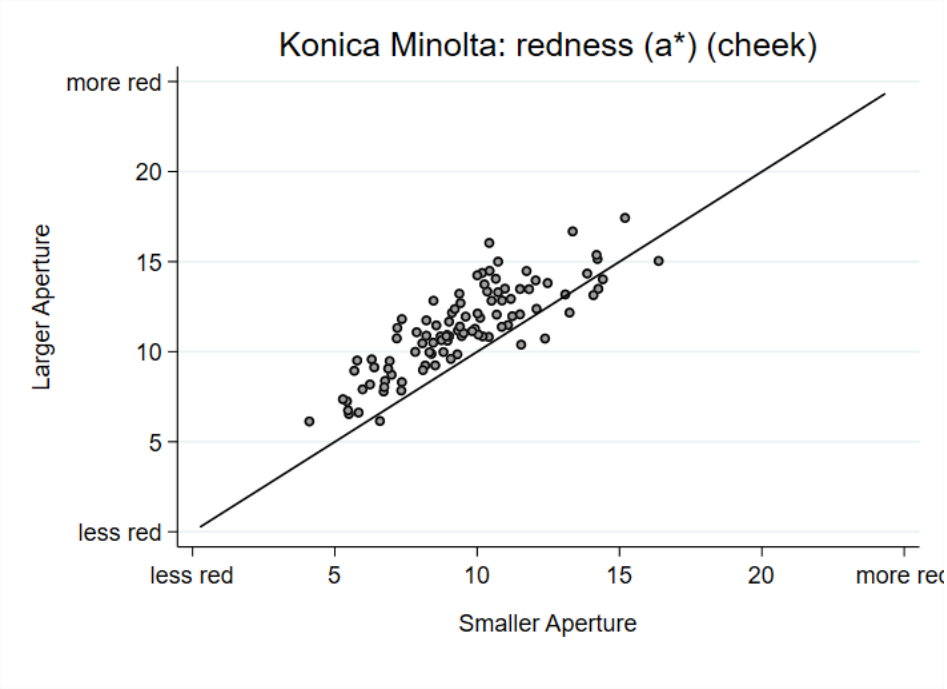


Figure 3. Illustration of redder readings on outer than inner arm

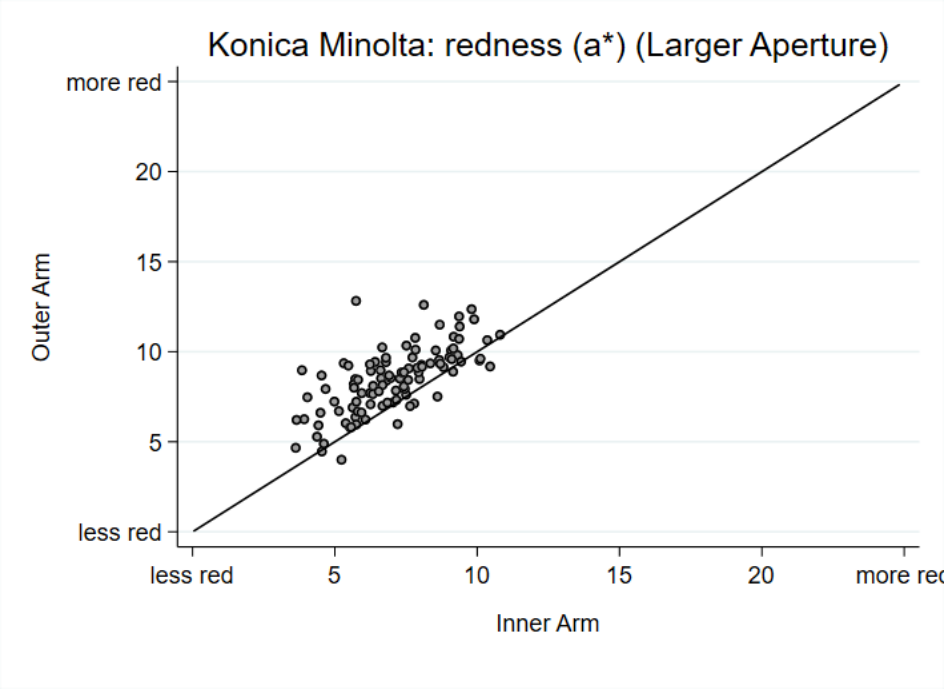
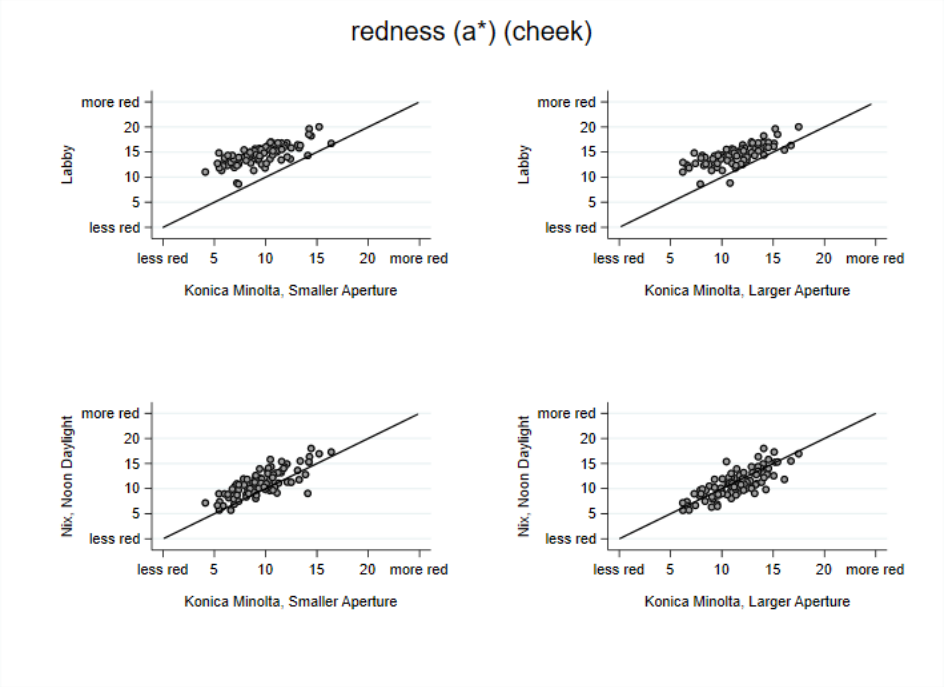


Figure 4. Illustration of redder readings by device and aperture



Links to Supplementary Materials

Appendix A:

https://www.surveypractice.org/article/94628-best-practices-for-measuring-skin-color-in-surveys/attachment/220460.xlsx

Appendix B:

https://www.surveypractice.org/article/94628-best-practices-for-measuring-skin-color-in-surveys/attachment/220459.xlsx