Assessing Intercoder Reliability of Skin Tones in Medical Textbook Imagery

OAKLAND UNIVERSITY WILLIAM BEAUMONT

Kaitlin Pataroque[1], Ryan Wu[1], Malli Barremkala, M.D.[2]

[1]Oakland University William Beaumont School of Medicine, Rochester, Michigan
[2]Department of Foundational Medical Studies, Oakland University William Beaumont School of Medicine, Rochester, Michigan

Introduction

As an independent risk factor for health outcomes and as a marker of disease presentation, an important consideration for physicians is skin tone representation in medical textbooks. However, substantiating colorism in these resources is contingent on the degree of reliability of the skin tone measure. Our goal is to assess intercoder reliability of one of the most widely used skin tone measures in social surveys – the Massey-Martin (MM) scale, in context of rating skin tones seen in medical textbook imagery.

Aims and Objectives

(I): Assess intercoder reliability among of the MM scale in rating skin tones seen in medical textbook imagery – using a survey that asks medical students to code images seen in popular preclinical anatomy textbooks ie. *Atlas, Bates', Clinically, and Gray's* using the scale.

(II): If present, examine and discuss potential sources of intercoder disparities in ratings such as participant race and self-identified skin tone.

Methods

M1 and M2 students attending OUWB were invited to complete an electronic survey. 78 responded, filling out a self-identification questionnaire detailing racial/ethnic group and self-identified skin tone. They were then asked to code 20 images selected from the most recent editions of popular preclinical anatomy textbooks, including *Atlas, Bates', Clinically, and Gray's* using the MM scale. We assessed intercoder reliability including measures such as average pairwise percent agreement and Krippendorff's alpha – stratified across respondent race and self-identified skin tone.

Fig. 1: Sample survey question with image from *Moore's Clinical Anatomy*.

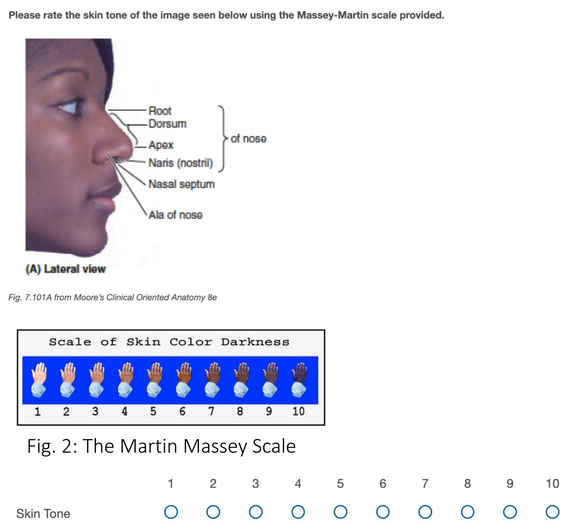


Fig. 2: The Martin Massey Scale

Results

91 responses to the survey were received. Of those responses, 3 did not fit the inclusion criteria because participants were color-blind. 10 responses were excluded due to incompletion

Table 1. Reliability of ratings by skin tone measure and (a) rater race, or (b) rater self-identified skin tone (20 photos)

Panel A

Reliability Measure	All raters (N=78)	White raters (N=54)	Asian raters (N=21)	Black raters (N=2)
% Agreement	36.80%	34.35%	31.68%	-
Krippendorff's alpha	0.261	0.246	0.225	-

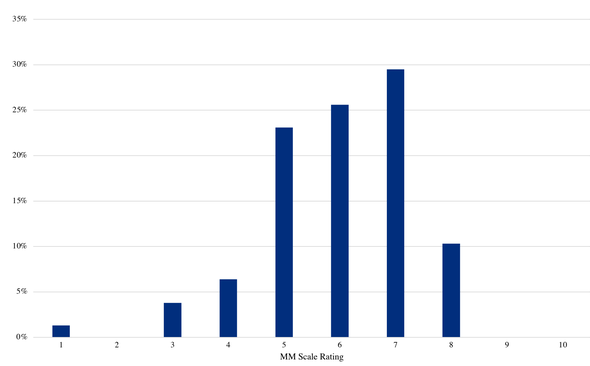
Panel B

Reliability Measure	Tone 1 (N=10)	Tone 2 (N=47)	Tone 3 (N=12)	Tone 4 (N=9)
% Agreement	41.78%	39.30%	35.15%	30.00%
Krippendorff's alpha	0.291	0.279	0.247	0.118

* N was insufficient to calculate reliability measures for Black (2) and Latinx (1) raters
* Based on 91 total survey responses. 3 did not fit inclusion criteria due to color blindness and 10 were excluded due to being incomplete

The use of the MM scale in rating skin tones seen in medical textbook imagery fails to pass acceptability standards for intercoder reliability. The aggregate average pairwise percent agreement was 36.8%. No group of coders, regardless of race or self-identified skin tones had an agreement above 45%. Furthermore, none all the associated Krippendorff's alphas passed the threshold of 0.8, falling <0.3 for each group.

Fig. 2: Sample distribution of responses for Fig. 7.118B from *Gray's*



Conclusion

While sample sizes were small for these analyses, especially respondents who identified as Black or of dark skin tone, the study provides evidence for low reliability of the MM scale in rating skin tones in medical textbook imagery – regardless of the coder's race or self-identified skin tone.

Acknowledgements

We would like to thank the anonymous medical students at OUWB who participated in this survey, as well as Michelle Jankowski for her assistance with data analysis.

References

1.) Campbell, M. E., Keith, V. M., Gonlin, V., Carter-Sowell, A. R. (2020). Is a picture worth a thousand words? An experiment comparing observer-based skin tone measures. Race and Social Problems, 12, 266–278. https://doi.org/10.1007/s12552-020-09294-0
2.) Feagin, Joe, Bennefield, Zinobia, 2014. Systemic racism and U.S. Health care. Soc. Sci. Med. 103, 7–14. http://www.sciencedirect.com/science/article/pii/S0277953613005121.
3.) 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
4.) Hannon, Lance, and Robert DeFina. 2016. "Reliability Concerns in Measuring Respondent Skin Tone by Interviewer Observation." *Public Opinion Quarterly* 80(2): 534–41.
———. 2020. "The Reliability of Same-Race and Cross-Race Skin Tone Judgments." *Race and Social Problems*. http://link.springer.com/10.1007/s12552-020-09282-4 (February 18, 2020).
5.) Hawley, Sarah Tropman, Anne Earp, Jo, O'malley, Michael, Ricketts, Thomas C., 2000. The role of physician recommendation in women's mammography use: is it a 2-stage process? Med. Care 38 (4), 392–403.
6.) 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.
7.) Louie P, Wilkes R. Representations of race and skin tone in medical textbook imagery. Soc Sci Med. 2018;202:38–42
8.) Mangione-Smith, Rita, et al., 2004. Racial/ethnic variation in parent expectations for antibiotics: implications for public health campaigns. Pediatrics 113 (5), e385–e394.
9.) Martin, Glenna, Julianna, Kirgis, Eric, Sid, Sabin, Janice, 2016. Equitable imagery in the pre-clinical medical school curriculum. Acad. Med. 91 (7), 1002–1006.
10.) Massey, Douglas S., Martin, Jennifer A., 2003. The NIS Skin Color Scale. Office of Population Research, Princeton University.