Guidelines for race and ethnicity data in linguistic studies

Robert Squizzero
Martin Horst
Alicia Beckford Wassink

UW Linguistics Colloquium Series
April 21, 2023

Presentation Roadmap

1. Motivations for the work
2. Summary Recommendations
3. Problems with using predefined labels
4. “Defining” race and ethnicity
5. Conceptualizations in linguistic subfields
6. Risks to study participants, researchers, and linguistics
7. Wrap up

Where we are headed

Submission to *Language*: *Commentary* “Collecting and using race and ethnicity data in linguistic studies” (Squizzero, Horst, Wassink, Panicacci, Jensen, Moroz, Conrod, and Bender, in prep.)

Motivations for the study

Colleagues' requests for recommendations and templates

Release of the US 2020 Census, with revised demographic categories

Two-quarter long Sociolinguistics Brown Bag series on Race and Ethnicity:

Racializing practices in Linguistics (Charity-Hudley, 2017)
Linguistic Society of America Statement on Race (2019)
Methodological techniques (Wassink, UW; Maya Smith, UW; Alex Panicacci, Queen Mary University)
Best practices from sister fields

January 20, 2020
Series: "Collecting race and ethnicity information in Linguistics"

May 25, 2020
Murder of George Floyd

Survey of Research Design Advice

62 textbooks found and reviewed

Publication years: 1951 and 2020

All subfields queried (including core subfields, applied linguistics, corpus linguistics, language documentation anthropological linguistics)

New researchers are most likely to find guidance about **conceptualizing (speech/language) community** in critical sociolinguistics, applied sociolinguistics, language documentation, language variation and change.

2 ACTUALLY "WENT THERE"

New researchers are most likely to find general guidance about **designing demographic prompts** in methods texts about studying language variation and change.

Summary Recommendations

1. **Thoroughly evaluate** the **complexity of race and ethnicity** as it relates to your study
    a. Should you include demographic data? Yes.
    b. Embed considerations into the study design itself
    c. Choose the appropriate research question

2. Recognize that **self-identification** is **fluid**, **context-specific, and SELF-identified**
    a. Use community-based labels
    b. DO include complex identities in analysis rather than striking/changing them

3. Commit to **linguistic social justice** at all stages of your study
    a. Foreground the linguistic community of study
    b. Bolster the kinds of information that may reduce certain negative practices.
    c. “All linguistic research has the potential to reproduce or challenge racial notions” (LSA Statement on Race)

Why?

1. Linguistics has been criticized for undertheorized application of the notions of race and ethnicity (from outside and inside)
    - 2019 LSA Statement on Race

Charity Hudley & Mallinson (2011)
García Sanchez (2014)
Cheshire (2016, pc)
Charity-Hudley (2017)
Lanehart (2022)

2. Better alignment with our sister fields (anthropology, archaeology, psychology, sociology)

López et al. (2017)
Fuentes et al. (2019)
García (2020)
Charity Hudley et al. (2020)

3. Improved research ethics & respect for communities we represent

Cameron et al. (1992)
Rice (2010, 2012)
Eckert (2013)

How the 2020 census asked about Hispanic origin and race

➜ **NOTE: Please answer BOTH Question 6 about Hispanic origin and Question 7 about race. For this census, Hispanic origins are not races.**

6. Are you of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

⚠ CAUTION

7. What is your race?

*Mark [X] one or more boxes **AND** print origins.*

- ☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*
- ☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*
- ☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*

- ☐ Chinese
- ☐ Filipino
- ☐ Asian Indian
- ☐ Vietnamese
- ☐ Korean
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Samoan
- ☐ Chamorro
- ☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*
- ☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*

- ☐ Some other race – *Print race or origin.*

Race & ethnicity defined

- Race and ethnicity are viewed as overlapping and used interchangeably

- In linguistic research, it is important to distinguish between:

1) Approaches that are static and essentializing (usually race-based)

2) Approaches that are practice-based (usually ethnicity-based)

Race

- Race refers to people sharing physical features, especially skin color, facial features, eye shape, and hair texture (Bobo 2001; Spears 2020)
- Race essentialism is the tendency to view race as biologically based, immutable, and informative (Haslam, Rothschild & Ernst 2000; Prentice & Miller 2007)
  - Race essentialism has been linked to racial stereotyping and prejudice (Levy & Dweck 2003; Williams & Eberhardt 2008)
- Essentialist racial classification schemes based in biology or genetics are:
  - unreliable (Garcia 2020; Relethford 2009)
  - severely flawed (Keita et al. 2004)
  - completely arbitrary (Omi & Winant 2014)

Race and Racism

- Defining race is pointless outside of an acknowledgement of racism; race implies and requires **hierarchy and the hegemonic positioning** of the categories it creates (Lanehart, 2023)
- "**Racial thinking, racial law, racial formation, and racialized behaviors** and phenomena in medieval Europe [emerged] **before** the emergence of a recognizable vocabulary of race…
- …a political hermeneutics of religion – so much in play again today – enabled the positing of fundamental human differences in biopolitical and culturalist ways to create **strategic essentialisms demarcating human kinds and populations**." (Heng, 2011)

Ethnicity

- Ethnicity refers to a group identified based on:
  - Shared signs (in the semiotic sense),
  - Shared aspects of a common culture, or
  - Shared practice

- Material manifestations of shared aspects of culture may include:
  - Following patterns of dress
  - Adhering to diets or eating particular foods
  - Observing holidays
  - Practicing religions, and crucially
  - **Speaking languages and language varieties**

(Garcia 2020, writing on behalf of the American Anthropological Association and the Society for Anthropology in Community Colleges)

So should I ask about race or ethnicity?

- *Probably ethnicity*. When describing language, use practice-based approaches
- Using community-based practices, linguists are less likely to essentialize speech communities
- Practice-based approaches are less exclusionary
    - Many individuals do not possess the phenotypic characteristics stereotypically associated with members of the community to which they belong
    - Practice-based approaches describe what people **do**, not what people look like

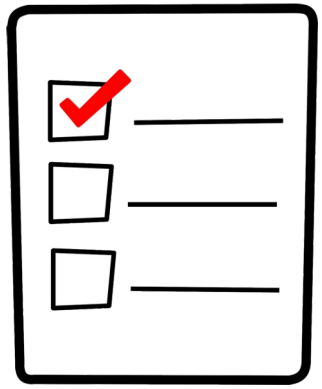
So should I ask about race or ethnicity?

- Main exception: if you are investigating racism, you might ask about race
- *Linguists should never talk about or imply biological predispositions towards language use as pertains to race and ethnicity*

No, no, I meant, should I ask about either one at all?

- Yes. (Usually)
- Authors writing on or characterizing a language or language variety may assume or imply that the language structure under discussion is invariant.
    - (1) I ain't not telling you nothing, no more. (Double negation)
    - Example (1) would likely appear preceded by an asterisk in an article discussing English syntax.
        - *Assumption*: "English" refers to mainstream, White US English
    - Would an author get away with presenting (1) as grammatical without any comment on social identity whatsoever?
    - This asymmetry indicates an entrenched assumption of Whiteness as default that is endemic in formal fields; bringing attention to it in our writing is a basic and necessary first step to fixing that

Back to categories – which categories should I use?

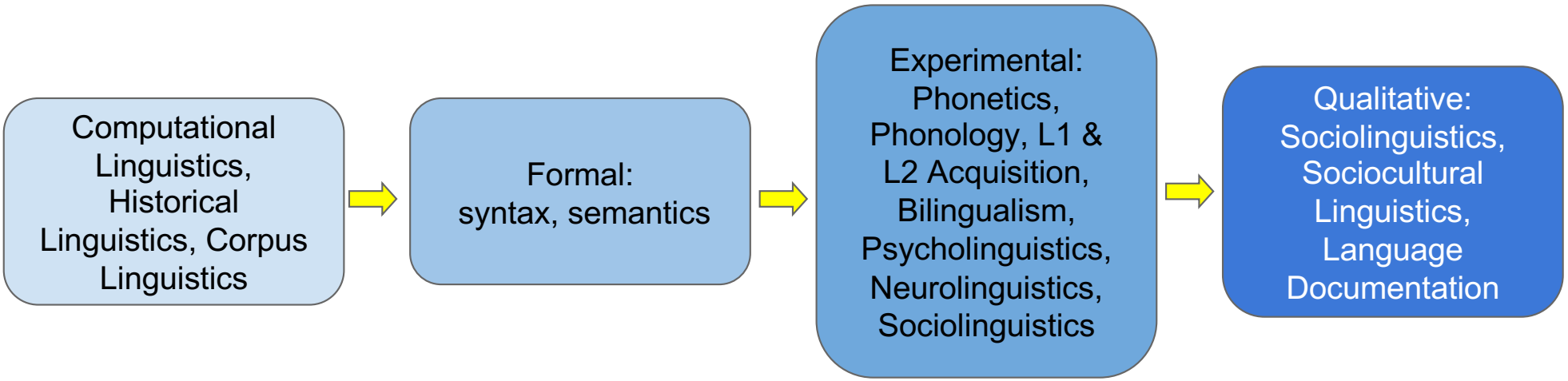


- Ask yourself: Do I need categories at all?
- Categories and labels help with:
  - Large-scale analysis
  - Providing a ‘big picture’ of social phenomena
  - Revealing inequities across social groups
- Categories and labels may fail to grasp:
  - Individuals’ authentic self-identifications & inter/intragroup variation
  - Outsiders’ categorization of the self
  - Fluidity of ethnic and racial identities
  - Multiethnic, multiracial, and ambiguous ethnic and racial identities
  - Intergroup perceptions and dynamics
- Example: If you aim to elicit grammaticality judgments from a handful of individuals, should you use categories at all?

Which categories should I use?

- <u>If you do need category labels, our recommendations are:</u>
  - Let the community studied guide the ethnic or racial categorization that you use.
  - Ask about participants' self-identifications when piloting the study to identify flaws in the categorization system you adopt.
  - Ideally, supplement category data with ethnographic or qualitative data.
  - When piloting, specifically seek feedback from multiracial or multiethnic participants.
    - 8.8% of American adults self-identified as multiracial on the 2020 Census.

How much demographic information might I need?



Computational linguistics

What differentiates CL from other experimental disciplines?

- CL data are immediate, digitized, and vast
- Largely anonymous, often internet-sourced language databases
  - Bender and Friedman (2018)

- The need for race and ethnicity data arises during specific tasks which implicate social identities
  - E.G., Hate speech detection, speech to text
  - Salience during studies of bias derived from CL tasks (Manzini et al 2019)
  - "Debiasing" models is possible, but may be ineffective/harmful (Gonen and Goldberg 2019)
    → Erasure of certain identities or people

Computational linguistics (cont.)

- Post-hoc attributing labels to individuals / language artifacts
  - Produce a small subset of data annotated with known identities
  - Predict remaining group's identifications using this classifier
    → Racializing participants without their consent or knowledge

Recommendations

1. Transparency if self-identity data are not known
2. Ethical data collection with as much information as possible
3. Avoiding essentializing linguistic features as THE markers of racialized language varieties (Charity Hudley 2017)
4. Beware of linguistic appropriation in datasets (Abreu 2015)

Special thanks to Dr. Emily Bender for a majority of the content for this slide :)

Grammaticality judgments

- Question from earlier: If you aim to elicit grammaticality judgments from a handful of individuals, should you use ethnic categories at all?

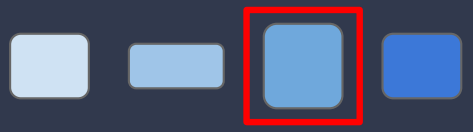
- Answer: Predefined categories are not as useful as a brief conversation with your sources.

Our Recommendations:

1. Identify your sources
2. Ask your sources what relevant social categories would apply if they were to write a brief positionality statement or bio for themselves. Include this information in a footnote or an appendix
3. If your source is yourself, write a brief positionality statement

Special thanks to Dr. Kirby Conrod for a majority of the content for this slide

Perception studies



Our recommendations apply to studies of both production and perception.

1. Collect demographic information from your stimulus speakers and especially from your study participants

- Demographic information allows the **generalizability** of results to be qualified.
  - Minority groups are frequently undersampled
- In designing a perceptual experiment, researchers typically devote a great deal of time and energy to avoiding potential **confounds**, such as the age and gender of participants
- Race and ethnicity can be confounding factors in that:
  - The stated or implied race or ethnicity of a stimulus speaker, signer, or writer can affect results (Squizzero 2020; 2022)
  - Participants may be users of language varieties associated with their race or ethnicity

Perception studies

Take special care to balance risks and benefits when:

- Investigating effects of a speaker's race or ethnicity on perception
- Investigating the perceived race or ethnicity of a speaker based on the language they produce
    - Example: in a within-subjects design asking listeners to rate the accentedness of speakers based on visual cues to the race of the speakers, participants may incorrectly infer that race is a valid and reliable predictor of accentedness

Our recommendations for mitigating or offsetting potential harm:

1. Debrief your participants
2. Follow up with an interventional study (see Kang, Rubin & Lindemann 2015)
3. Work in linguistic contexts where racial linguistic stereotyping is undocumented or under-documented (see Hanulíkova 2018; Squizzero 2022)

Quantitative/Experimental:

Typical types and amounts of data:

- Large judgement or random sample (primary data)
- Inferential, time series and descriptive analysis
- Coding for age, gender, language exposure, interlocutor type, treatment, group, etc.

Typical style of research question is **descriptive** or **correlational**:

- Sociolinguistics: Distribution of some linguistic feature, e.g., “Is use of Avertive *liketa* disfavored in constructions displaying multiple-negation?”
- Acquisition: “Does X (FL learning) occur differently in group Y (intervention children) than in group Z (CPC)?” (Ferjan-Ramirez & Kuhl, 2020)

Experimental:

Appropriate question type: multiple choice with well-justified categories, free-response or interview-style

Pros: Ease of summarization

Cons: Subjective or complex self-identifications, intersectionality need to be treated with care

Recommendations:

1. understand which demographic labels might be relevant in the community of interest.
2. allow participants/caregivers to not answer questions.
3. allow participants/caregivers to choose multiple options.
4. include an option where participants/caregivers can name one or more labels not already included which are relevant to them.
5. Report analyst's positionality.

Example

7d. What words would you use to describe your ethnic or cultural background? _____

7e. Please consider the words you just used to describe your ethnic self-identification and respond to the following. Which of the following describe "ethnic" practices that might place you in this group: [Fieldworker instruction: check as many as apply]

__ I use particular dialect or language forms.
__ I cook/eat particular foods.
__ I attend worship services/ceremonies.
__ Other: _____
__ I participate in particular events.*
__ I listen to particular music.
__ I travel to particular places.

7f. Considering your own cultural or family background, do you consider yourself to be a member of any group(s) outside your ethnicity or local community? What would you call that group/those groups?

8. The following are the US Census Bureau's 2020 categories for racial self-identification. For the purposes of this study, how would you prefer to be categorized?

__ American Indian or Alaska Native
__ Asian
__ Black or African American
__ Hispanic or Latino
__ Native Hawaiian or Other Pacific Islander
__ White

Language Documentation

Typical types and amounts of data:

- Qualitative - Ethnographic observation of “lived routines of daily living”
- Period of observation spans years
- Large amounts of data recorded by the analyst

Research Questions vary. Characterization of some linguistic phenomenon within an individual language.

- Community participation
- Lapierre (p.c.): Age, gender, clan, familial and social roles (marriage), village of origin
- Analyst records key facts

Key: *Ethnographic approach*

`Ethnos > Gk. “belonging”`

Recommendations: See previous slide; include practices

Risks of non-critical treatment of race and ethnicity data

1. Social harm to study participants
    a. Restriction of human rights based on race, ethnicity, or nationality (Smith 2019)
2. Misleading study participants into believing that race and ethnicity are biologically founded
    a. Thus exacerbating inaccurate/harmful models of social identity in lay communities
3. Negative impact on data accuracy
    a. Unreplicatable, opaque, or porous conclusions
4. Potential harm to linguistic researcher
    a. Bad-actors who use your research may bend your statements to their own ends
5. Exacerbating structural inequity in linguistic scholarship
    a. Furthering hegemonic systems of “ivory tower”ness (LSA Statement on Race)

What you gain by critical treatment of race & ethnicity data

- ➔ Better representing speakers in NLP modeling
  - ◆ Transparency!
- ➔ Avoiding social harm to participants
- ➔ Appropriate level of generalization for theory building
- ➔ Better descriptive representation of undersampled groups
- ➔ Deeper understanding of how native language can impact perception and production
- ➔ … and more! (see Squizzero et al. 2021)

Concluding remarks

- We argue that linguistics has the opportunity to **more accurately and more responsibly model race and ethnicity** across all subfields
  - During dataset creation, study design, and after-study analysis
- We aim to **empower researchers** to apply these techniques
  - Complexity does not equal impossibility
  - Approaches will differ, but thoughtfulness should be normative
- We reiterate to reinforce the notion that **language structure and speakers/signers of that language** are **inextricably linked**
  - Centering community of study by design
  - Research on, for, and with the community (Cameron et al. 1992)

Acknowledgements

- Our co-authors who are not presenting with us today.
  - Alex Panicacci, Monica Jensen, Anna Kristina Moroz, Kirby Conrod, Emily M. Bender
- Editors who have read the work and offered critiques, praise, and constructive criticism
  - Anne Charity Hudley, Arthur Spears, Sonja Lanehart, Laada Bilaniuk, Maya Angela Smith, Sara B. Ng
- Research assistants who have helped format the work into a publishable piece
  - Laura Munger
- You! For hearing us and helping further our joint effort to better the fields of linguistics

References

Abreu, A. M. (2015). Online Imagined Black English. *Arachne*. Available at: https://arachne.cc/issues/01/online-imagined_manuel-arturo-abreu.html

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Bobo, Lawrence. 2001. Racial attitudes and relations at the close of the twentieth century. In *America becoming: Racial Trends and Their Consequences,* vol. 1, ed. Neil J. Smelser, William Julius Wilson, and Faith Mitchell, 264-301. Washington, DC: National Academy Press.

Bucholtz, M. 2020. Race, Research, and Linguistic Activism. *The Routledge Companion to the Work of John R. Rickford*.

Cameron, Deborah, Elizabeth Fraser, Penelope Harvey, M. B. H. Rampton, and Kay Richardson. 1992. *Researching language: issues of power and method*. London, UK/New York, NY: Routledge.

Charity Hudley, A. H. 2017. Language and racialization. *The Oxford Handbook of Language and Society.* Oxford, UK: Oxford Handbooks.

Charity Hudley, A. H., Mallinson, C., & Bucholtz, M. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language* 96(4). 200–235.

Comaroff, J., & Comaroff, J. (2009). *Ethnicity, Inc.* Chicago: The University of Chicago.

Eckert, P. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology,* 41, 87-100.

Fasold, R. 2019. Comment on: LSA Statement on Race

Fuentes, Agustín, Rebecca Rogers Ackermann, Sheela Athreya, Deborah Bolnick, Tina Lasisi, Sang-Hee Lee, Shay-Akil McLean, Robin Nelson (2019) AAPA Statement on Race and Racism, *American Journal of Biological Anthropology* 169*(3),* https://doi.org/10.1002/ajpa.23882

García, J. D. 2020. Race and Ethnicity. In N. Brown, T. Mcilwraith, & L. Tubelle De González (Eds.), P*erspectives: An Open Invitation to Cultural Anthropology*, 2nd edn. 444–455. American Anthropological Association. http://perspectives.americananthro.org/

Hanulíková, Adriana. 2018. The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguistics Vanguard* 4(1). 1–9. https://doi.org/10.1515/lingvan-2017-0029

Haslam, N., Rothschild, L., & Ernst, D. 2000. Essentialist beliefs about social categories. *British Journal of Social Psychology*, 39(1), 113–127.

Heng, Geraldine. (2011). The Invention of Race in the European Middle Ages I: Race Studies, Modernity, and the Middle Ages. *Literature Compass*, 8(5), 315-331. https://doi.org/10.1111/j.1741-4113.2011.00790.x,

Kang, Okim, Donald L. Rubin & Stephanie Lindemann. 2015. Mitigating U.S. Undergraduates' Attitudes Toward International Teaching Assistants. *TESOL Quarterly* 49(4). 681–706. https://doi.org/10.1002/tesq.192

References

Keita, S. O. Y., Kittles, R. A., Royal, C. D., Bonney, G. E., Furbert-Harris, P., Dunston, G. M., & Rotimi, C. N. (2004). Conceptualizing human variation. *Nature genetics*, *36*(Suppl 11), S17-S20.

Lanehart, S. 2023. *Language in African American Communities.* (Routledge guides to Linguistics). London: Routledge.

Levy, S. R., & Dweck, C. S. (1999). The impact of children's static versus dynamic conceptions of people on stereotype formation. *Child Development*, 70(5), 1163-1180.

Linguistic Society of America. 2019. Statement on Race. https://www.linguisticsociety.org/content/lsa-statement-rac

López, N., Vargas, D. E., Juarez, M. Cacari-Stone, L., & Bettez, S. 2017. What's Your "Street Race"? Leveraging Multidimensional Measures of Race and Intersectionality for Examining Physical and Mental Health Status among Latinxs, *Sociology of Race and Ethnicity*. 4(1):49-66.

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Omi, Michael & Howard Winant. 2014. *Racial Formation in the United States* 3rd ed. Routledge.

Pauker, K., Apfelbaum, E. P., & Spitzer, B. 2015. When societal norms and social identity collide: The race talk dilemma for racial minority children. *Social psychological and personality science*, 6(8), 887-895.

Pauker K., Meyers C. K., Sanchez D. T., Gaither SE, Young DM. 2018. A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and personality psychology compass.*

Prentice, D. A., & Miller, D. T. 2007. Psychological essentialism of human categories. *Current Directions in Psychological Science*, 16(4), 202-206.

Relethford, J. H. (2009). Race and global patterns of phenotypic variation. *American journal of physical anthropology*, *139*(1), 16-22.

Smith, Maya A. 2019. *Senegal Abroad: linguistic borders, racial formations, and diasporic imaginaries.* Madison: University of Wisconsin Press.

Spears, A. 2020. Racism, Colorism, and Language within their macro contexts. In *The Oxford Handbook of Language and Race* (H. S. Alim, A. Reyes, & P. Kroskrity, Eds.). New York: Oxford.

Squizzero, Robert. 2020. Attitudes toward L2 Mandarin Speakers of Chinese and non-Chinese Ethnicity. In Kaidi Chen (ed.), *Proceedings of the 32nd Meeting of the North American Conference on Chinese Linguistics*, 521–538. Storrs, CT.

Squizzero, Robert. 2022. *Sociolinguistic and Phonetic Perception of Second Language Mandarin Chinese*. University of Washington dissertation.

Squizzero et al. (2021) "Collecting and using race and ethnicity information in linguistic studies" *U Washington Working Papers in Linguistics.*

Williams, B. F. 1989. A Class act: anthropology and the race to nation across ethnic terrain. *Annual Review of Anthropology* 18, 401-444.

Williams, M.J., & Eberhardt J.L. 2008. Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology,* 94, 1033–1047. doi: 10.1037/0022-3514.94.6.1033