Implicit Attitudes, NLP, and the "Real World"

Rob Voigt

Stanford Linguistics

robvoigt@stanford.edu

Other Work

- Quantifying “modernity” in Chinese poetry

  NAACL Comp Ling for Literature 2013

- Discourse-level effects on reference

  who is “you” in reviews?

  A[illegible] 201[illegible]

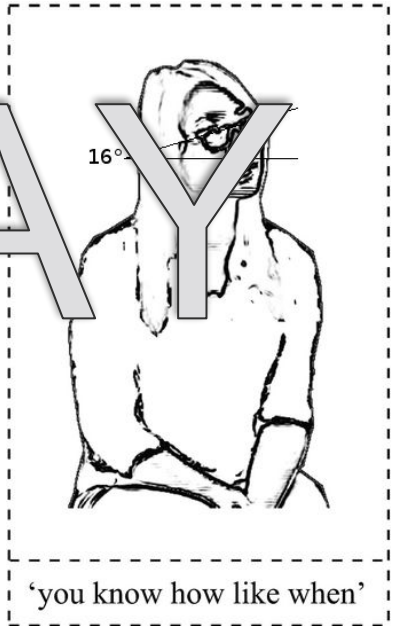
- Gender and pitch in bilinguals

  INTERSPEECH 2016

- Sociophonetic embodiment:

  Body movement and head positioning

  Journal of Sociolinguistics 2016



‘you know how like when’

Today - dissertation time!

Implicit Attitudes

"introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects"

Greenwald and Banaji 1995

NLP allows us to analyze linguistic "actions" at a large scale

...while accounting for crucial aspects of the "real world" social context

Project 1

Racial Disparities in Police Officer Respect

with Nick Camp, Camilla Griffiths, Will Hamilton, David Jurgens, Vinod Prabhakaran, Rebecca Hetey, Dan Jurafsky, and Jennifer Eberhardt

Our Question

Do officers treat White community members with a greater degree of respect than they afford to Blacks?

Police-Community Interaction

- Media focus on explosive incidents
- Research focus on outcomes

but:

- one quarter of adults have contact with the police during the course of a year
  - majority occurring in traffic stops

Respect is Important

- A person who is treated with respect
  - … has more trust in the individual officer's fairness
    Tyler and Ho 2001
  - … and the procedural fairness of the institution
    Tyler and Sunshine 2003
  - … and is more willing to support or cooperate with the police
    Tyler 1990, Mazerolle et al 2013

Previous work on procedural fairness

- Relies on:
  - citizens' recollection of past interactions

    Epp et al 2014
  - researcher observation of officer behavior

    Mastrofski et al 2009, Dai et al 2011, Jonathan-Zamir et al 2015

- These are invaluable but indirect
  - … and presence of researcher may influence police behavior

    Mastrofski and Parks 1990

Police body camera footage

- Oakland PD has been wearing body cameras since 2010
- Usually used only as evidence
- … but, a window into everyday behavior!

Our proposal: Footage as Data

- 981 stops by 245 officers in April 2014
  - Drivers: 682 black, 299 white
  - 183 hours of footage
- Professionally transcribed and diarized
- Resulting data set:
  - 36,738 officer utterances, 350k+ words

Sample transcription

0:00:00 0:00:09 OFFICER [to dispatch]: Unknown occupant and it's going to be for registration. It should be code four.
0:00:20 0:00:20 OFFICER: Hi.
0:00:20 0:00:20 FEMALE: Hi.
0:00:21 0:00:23 OFFICER: I pulled you over because your registration is expired by almost a year.
0:00:25 0:00:28 FEMALE: Okay, I have the paperwork for it, a moving permit?
0:00:28 0:00:28 OFFICER: I'm sorry?
0:00:29 0:00:30 FEMALE: I have the paperwork for it.
0:00:30 0:00:31 OFFICER: Okay.

Project 1

Study 1

Perceptions of Officer Treatment from Language

Study 1: Goals

- Can human raters judge respect from officers' language?
- Are there differences in officer respect towards Black versus White community members?

“Thin Slice” Utterance Rating Task

- Participants (N=70) blind to race labeled 414 officer utterances
  - 10 coders per utterance
  - 4-point Likert scales

    Respectful, Polite, Friendly, Formal, Impartial

    (high rater agreement αs=.73-.91)

Utterance Rating Task

Read the following interaction with a police officer:

The citizen just said:

It's in my glove compartment.

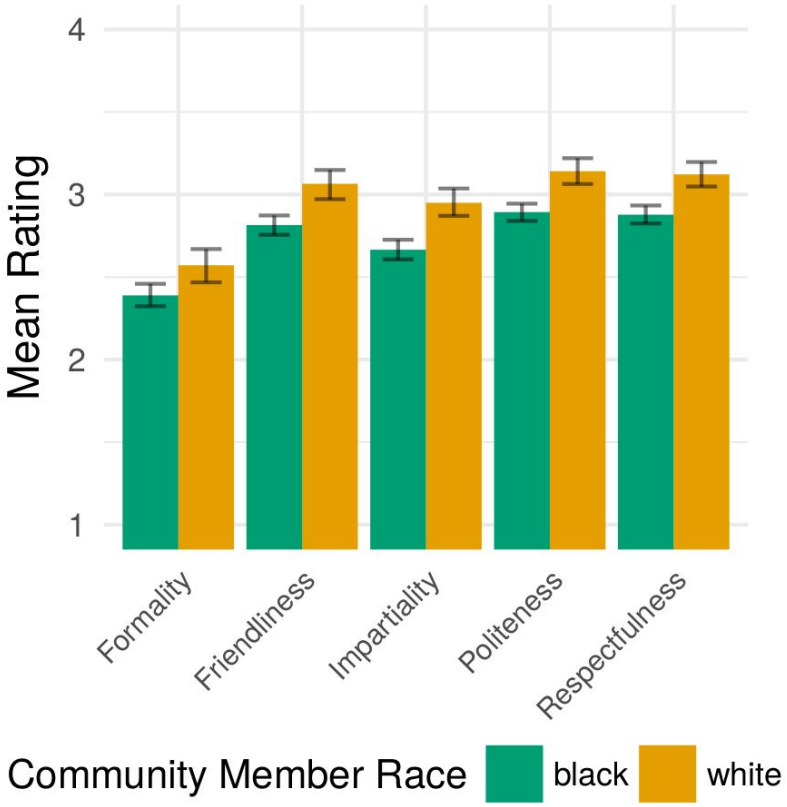
And then the officer says:

Let me take a look at it. How about insurance?

How ***impolite*** or ***polite*** was the officer?

Very Impolite	Somewhat Impolite	Somewhat Polite	Very Polite
○	○	○	○

Utterance Rating Task



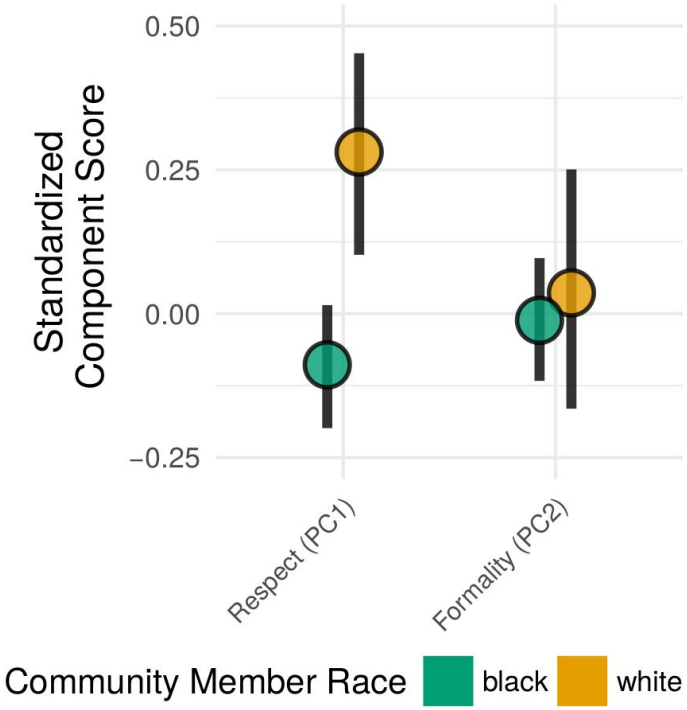
The Latent Space of Respect

Two PCs explain 93% of the variance:

	Respect	Formality
variance explained:	71%	22%
Formal	0.27	0.91
Friendly	0.47	-0.39
Polite	0.49	-0.04
Respectful	0.47	0.03
Impartial	0.50	-0.11

The Latent Space of Respect

- Race on these dimensions:



Project 1

Study 2

Modeling Respect with Computational Linguistics

Study 2: Goals

- Develop a computational linguistic model capable of estimating Respect
- Use the human labeled data as supervised training data to learn weights on interpretable features

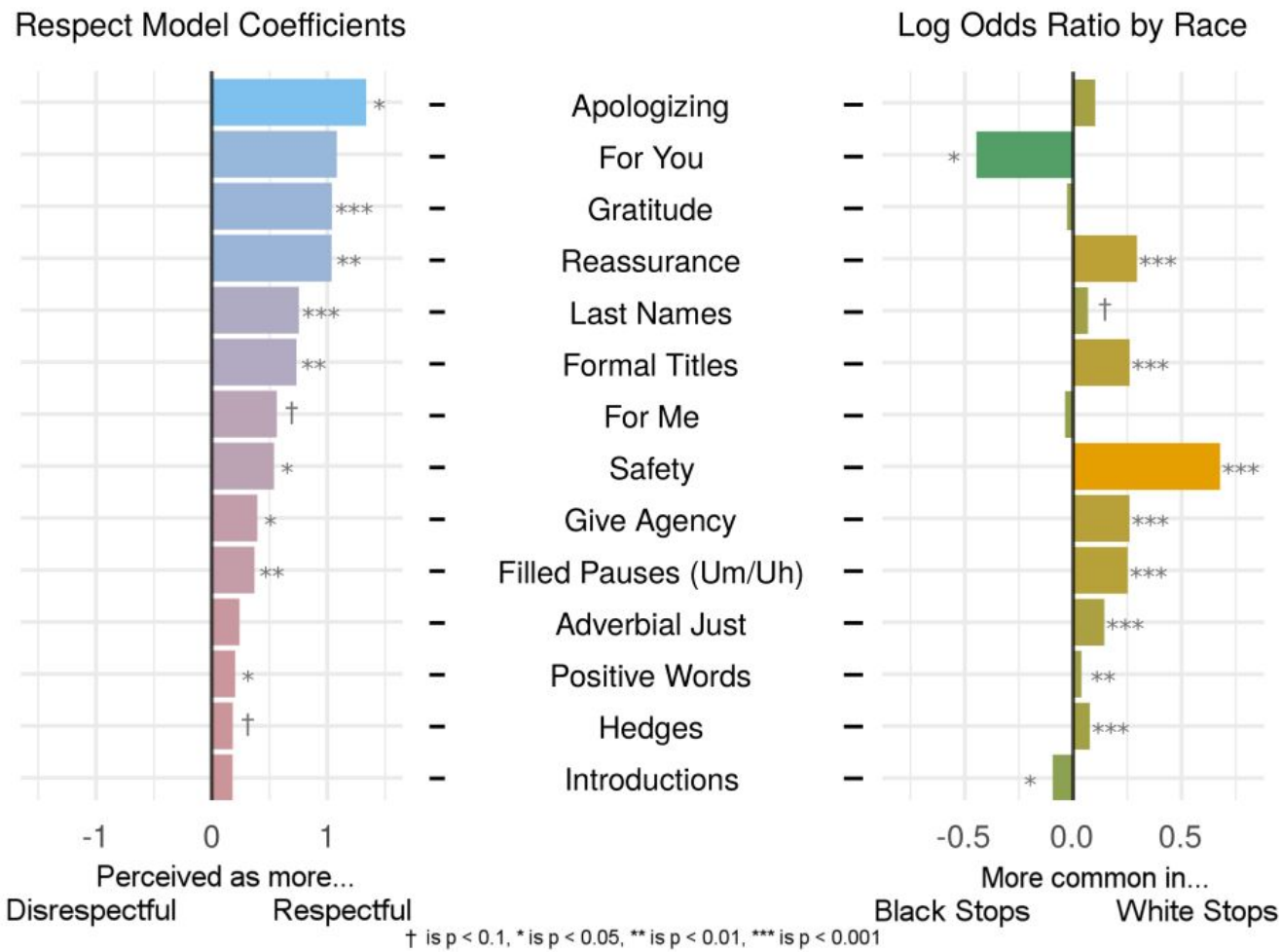
Methodology

- Hand-engineered features
  - Lexicons, gazetteers, regexes, dependencies, joint pattern matching ("bald commands")
  - Drawn primarily from linguistic and computational work on politeness

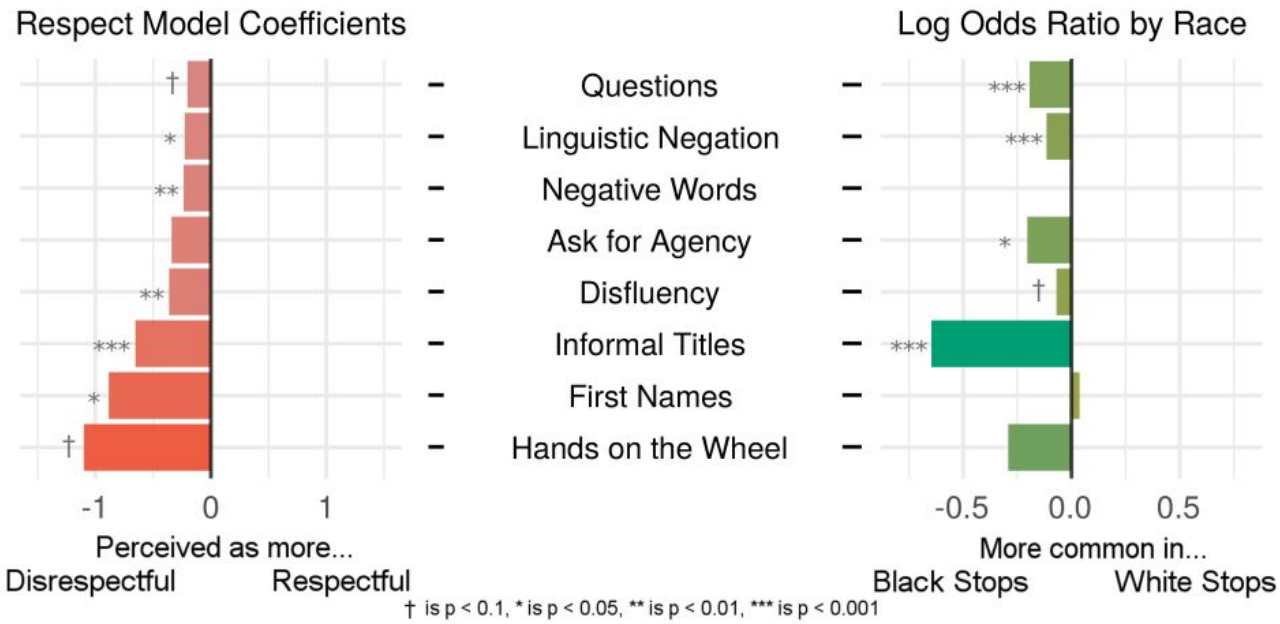
Goffman 1967, Lakoff 1973, Culpepper 1976, Brown and Levinson 1978
Prabhakaran et al 2012, Danescu-Niculescu-Mizil 2013, Krishnan and Eisenstein 2014

- Statistical Model: simple linear regression
  - log-tranformed counts of features per utterance

Feature Weights



Feature Weights



Example	Respect Score
First Name: [name] Ask For Agency: can I Questions: see that driver's license again? / It- it's showing suspended. Is that- that's you? Disfluency: It- Negative Word: suspended Disfluency: that-	-1.07
Informal Title: man Ask For Agency: Do me a favor. Adverbial "Just": Just All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick. "Hands On The Wheel": steering wheel	-0.51

Example	Respect Score
Apology (Sorry) to stop you. Introduction (My name's) Officer [name] Last Name (with the Police Department).	0.84
There you go, Formal Title (ma'am). Drive Safety (safe), Please (please).	1.21
It Adverbial "Just" (just) says that, Filled Pause (uh), you've fixed it. Reassurance (No problem). Thank you Gratitude (very much), Formal Title (sir).	2.07

Results

- *Respect* model is able to perform roughly like an average annotator

Model Adjusted R^2	0.258
Model RMSE	0.840
Average annotator RMSE	0.842 (range from 0.497 - 1.677)

- *Formality* model is worse but still reasonable

Model Adjusted R^2	0.190
Model RMSE	0.882
Average annotator RMSE	0.764 (range from .517 - 1.703)

Project 1

Study 3

Racial Disparity
Across the Entire Dataset

Study 3: Goals

- Do the results from Study 1 hold across an entire month of traffic stops?
- … even controlling for contextual factors?

Study 3: Results

	Respect			*Formality*		
	β	CI	p	β	CI	p
Arrest Occurred	-0.00	-0.03 – 0.03	.933	0.01	-0.02 – 0.04	.528
Citation Issued	0.04	0.02 – 0.06	**<.001**	0.01	-0.01 – 0.03	.209
Search Conducted	-0.08	-0.11 – -0.05	**<.001**	-0.00	-0.03 – 0.02	.848
Age	0.07	0.05 – 0.09	**<.001**	0.05	0.03 – 0.07	**<.001**
Gender (F)	0.02	-0.00 – 0.04	.062	0.02	0.00 – 0.04	**.025**
Race (W)	0.05	0.03 – 0.08	**<.001**	-0.01	-0.04 – 0.01	.236
Officer Race (B)	0.00	-0.03 – 0.04	.884	0.00	-0.03 – 0.03	.987
Officer Race (O)	-0.00	-0.04 – 0.03	.809	-0.00	-0.03 – 0.02	.783
Officer Race (B) : Race (W)	-0.01	-0.03 – 0.02	.583	0.01	-0.01 – 0.03	.188
Officer Race (O) : Race (W)	-0.01	-0.03 – 0.02	.486	-0.00	-0.02 – 0.02	.928

Interpretation

White community members are
57% more likely to hear an officer say one of the
top 10% most respectful utterances in our dataset

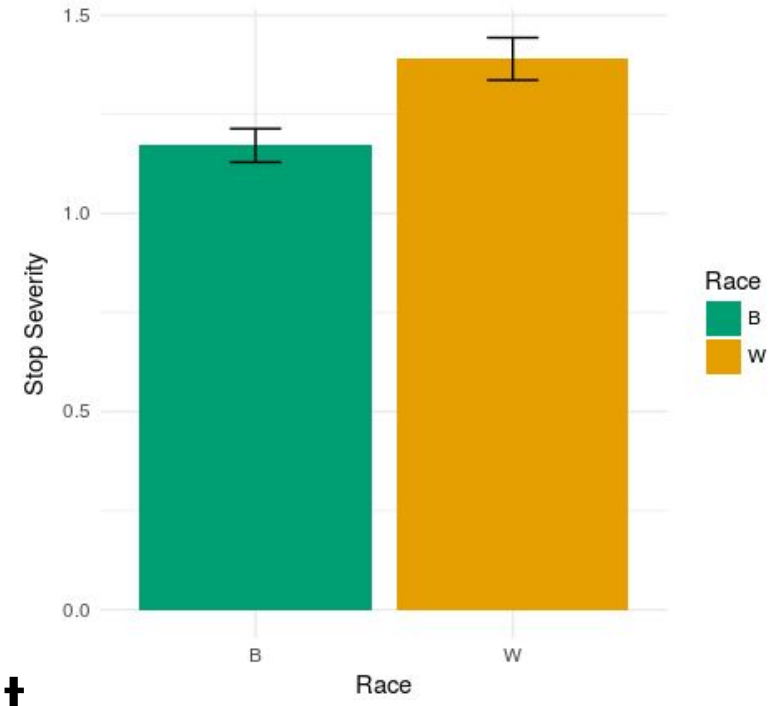
Black community members are
61% more likely to hear an officer say one of the
top 10% least respectful utterances in our dataset

Controls

- Holds even considering:
  - Only “everyday” interactions (no arrest, no search)
  - Crime rate in the area
  - Density of businesses in the area
  - Whether driver race was known before the stop
  - Officer years of experience

Controls - Severity

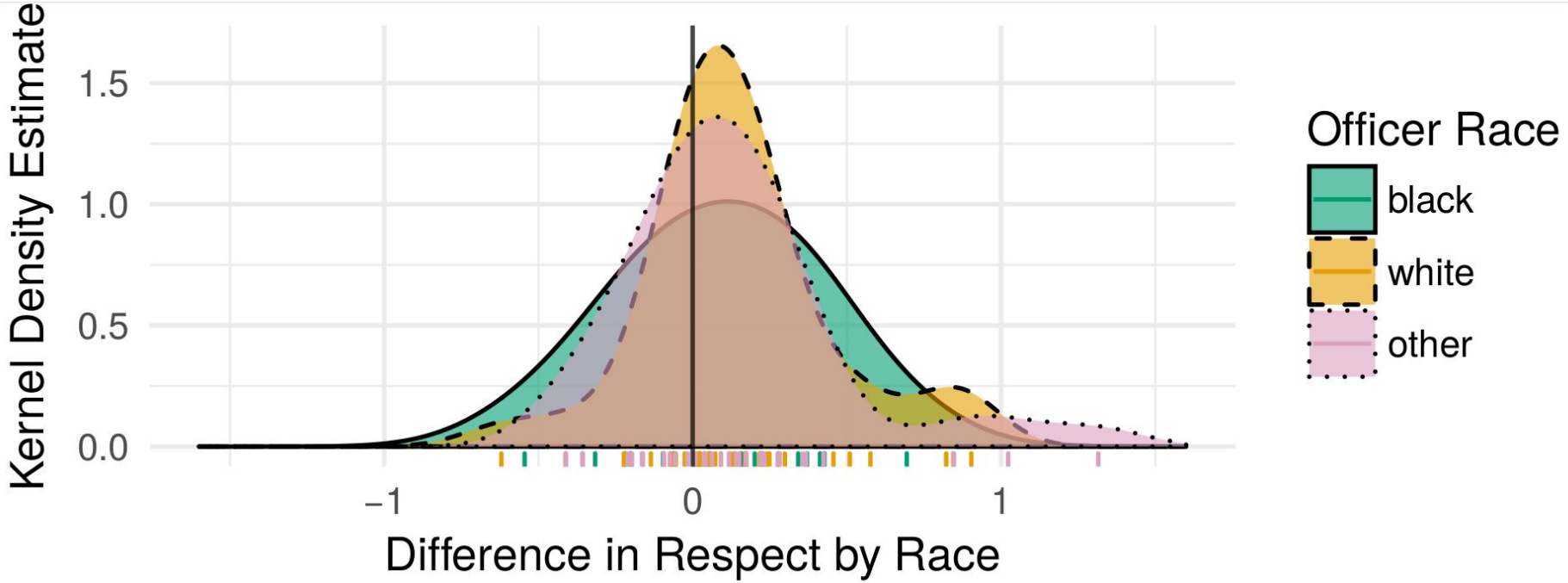
- We asked OPD officers to rate the stops for severity
  - 1 - very minor (expired registration)
  - 4 - very severe (speeding)



- Black drivers are stopped for less severe offenses
- … but no impact on respect

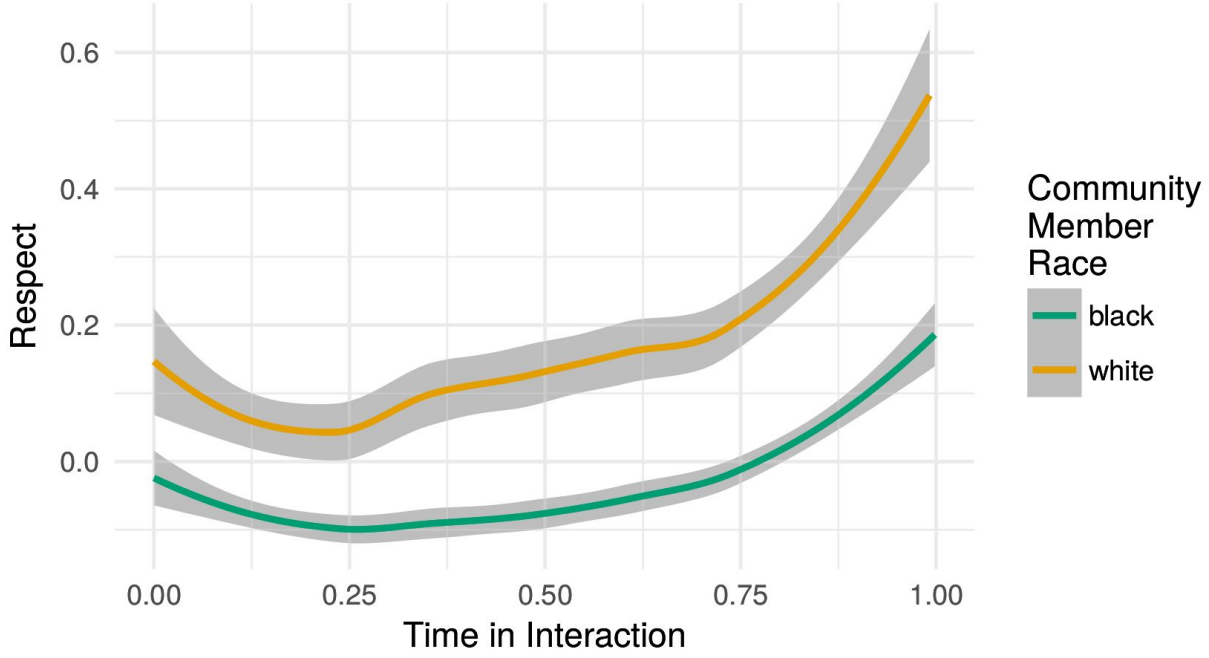
Controls - Officer Race

- Surprisingly, not a factor!



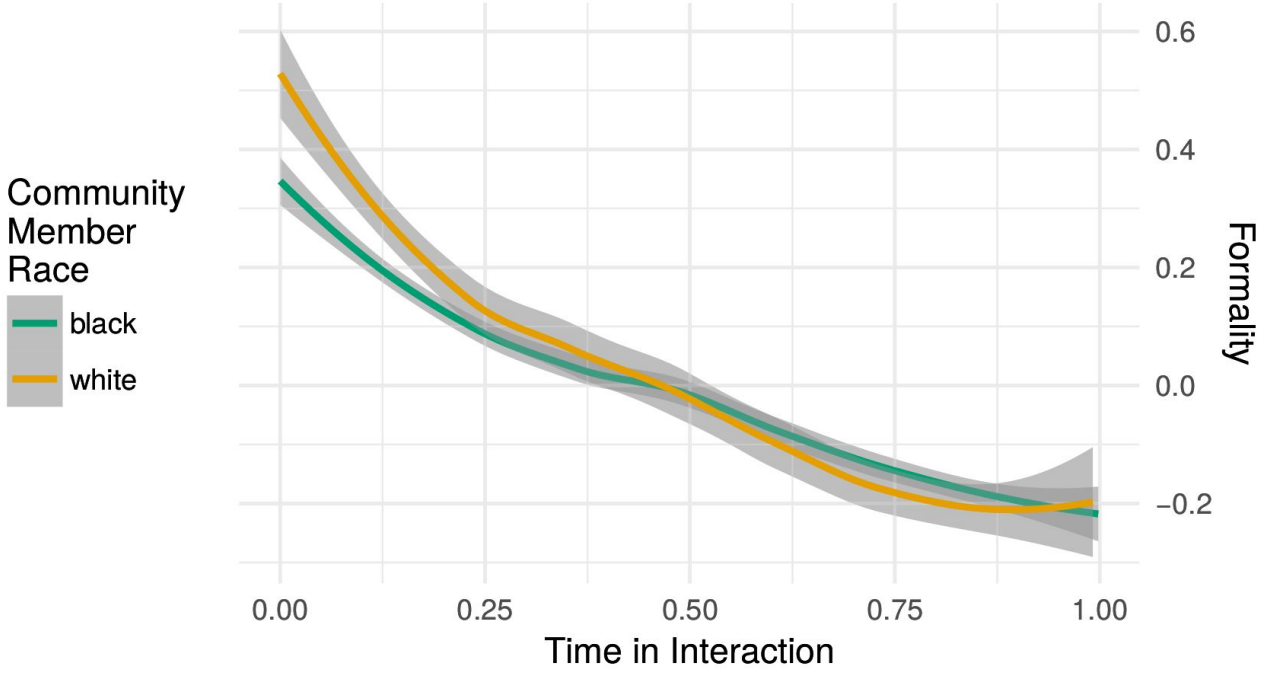
Across the Interaction

- Respect rises throughout the interaction
- … but rises faster for whites



Across the Interaction

- No race effect for Formality
- Officers less formal over the interaction



Conclusions from the first paper

- Confirms community reports: interactions with black community members are more fraught
- Provides concrete strategies for officers
- Cooperation with Oakland to integrate results into procedural justice training
  - … and we can measure impact

Moving Forward

- Tone of Voice:
  - Preliminary results suggest a similar trend
- Community member language:
  - Escalation
  - Compliance, politeness
- Other Departments