Language from police body camera footage shows racial disparities in officer respect

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt

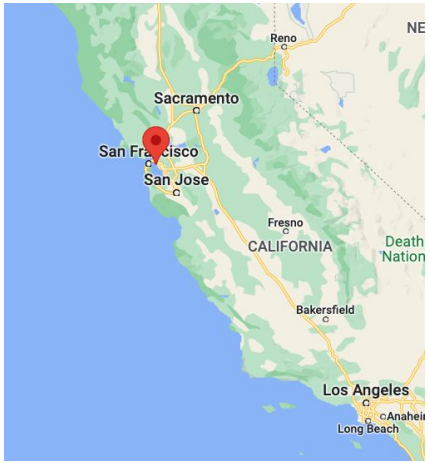
Presenter: Jiachen

Motivation

- Question:
  - Do officers treat white community members with a greater degree of respect than they afford to blacks?
- Significance:
  - Routine traffic stops are not only common, they are consequential, each an opportunity to build or erode public trust in the police. Being treated with respect builds trust in the fairness of an officer's behavior, whereas rude or disrespectful treatment can erode trust
  - Represent an important area of research of policing
  - Examine and improve police-community interaction
- Goal
  - Computational, large-scale analyses of the respectfulness of police officer language toward white and black community members during routine traffic stops

Data

- Source: transcribed body camera footage from vehicle stops
  - Oakland Police Department
  - during the month of April 2014
- Stats
  - **981 stops** (68.1% of the period) of black (N = 682) and white (N = 299) drivers
  - conducted by 245 different officers
  - 183h footage & **36,738** usable officer utterances



*Data processing

- Privacy preservation
  - data were kept on a central server, and transcribers (as well as all researchers) underwent background checks with the Oakland Police Department
- Transcription
  - video transcribed into text by professional transcribers
  - ‘diarized’: include labeling of who was speaking at each time point
  - kept only the speech from the officer directed toward the community member
- Cleaned up & processing (CoreNLP)
  - sentence & word segmentation
  - part-of-speech tags
  - dependency parses

Methods

- Study 1: Perceptions of Officer Treatment from Language.
  - Utterances were rated by human participants
- Study 2: Linguistic Correlates of Respect.
  - Statistical linguistic models are developed based on study 1
- Study 3: Racial Disparities in Respect.
  - Computational models applied to dataset

Study 1: Perceptions of Officer Treatment from Language.

- Objective:
  - whether human raters can reliably judge respect from officers' language
  - whether these judgments reveal differences in officer respect toward black versus white community members
- Data:
  - randomly sampled **414** unique officer utterances (1.1% of all usable utterances in the dataset)
    - black (N=312) or white (N=102)
  - all proper names and places were anonymized, and participants were not told the race or gender of the driver

transcribed language provides a sufficient and consensual signal of officer communication

- Method:
  - participants showed consistency in their perceptions of officer language, with reliability for each item ranging from moderate (**Cronbach's alpha** = 0.73) to high (**alpha** = 0.91) agreement

Batch	Formal	Friendly	Impartial	Polite	Respectful
1	0.82	0.86	0.84	0.86	0.83
2	0.88	0.89	0.86	0.86	0.87
3	0.80	0.87	0.73	0.84	0.78
4	0.85	0.91	0.79	0.88	0.87
5	0.77	0.89	0.81	0.87	0.87
6	0.91	0.82	0.81	0.87	0.86
7	0.85	0.86	0.84	0.84	0.84

Table 4: Annotator consistency (Cronbach's α) across batches and dimension for the utterance-level thin-slice judgments in Study 1.

whether participant ratings uncovered racial group differences

- Method:
  - On each trial, participants viewed the text of an officer utterance, along with the driver's utterance that immediately preceded it.
  - Participants indicated on four-point Likert scales how **respectful, polite, friendly, formal, and impartial** the officer was in each exchange.
  - Each utterance was rated by at least 10 participants.
- linear mixed-effects regression model
  - averaged scores across raters to calculate a single rating on each dimension for each utterance
  - estimate the fixed-effect of community member race across interactions

officer utterances directed toward black drivers were perceived as less respectful, polite, friendly, formal and impartial

Table 5 shows the results of linear mixed-effects models predicting score on each dimension as a function of the driver's race, sex, and age (standardized), with random intercepts for each stop.

	Respectful			*Polite*			*Impartial*			*Friendly*			*Formal*		
	b	CI	*p*	*b*	CI	*p*	*b*	CI	*p*	*b*	CI	*p*	*b*	CI	*p*
Fixed Parts															
Intercept	2.94	2.83 – 3.04	**<.001**	2.95	2.85 – 3.06	**<.001**	2.69	2.57 – 2.80	**<.001**	2.85	2.74 – 2.96	**<.001**	2.49	2.37 – 2.61	**<.001**
Driver Age	0.03	-0.02 – 0.08	.22	0.01	-0.04 – 0.07	.59	0.01	-0.05 – 0.07	.75	0.00	-0.05 – 0.05	1.00	0.08	0.02 – 0.14	**.01**
Driver Gender (F)	0.04	-0.07 – 0.16	.42	0.05	-0.07 – 0.16	.42	-0.0!	-0.13 – 0.12	.92	0.02	-0.10 – 0.14	.72	0.09	-0.04 – 0.22	.18
Driver Race (B)	-0.22	-0.33 – 0.10	**<.001**	-0.22	-0.34 – -0.11	**<.001**	-0.26	-0.39 – -0.13	**<.001**	-0.23	-0.36 – -0.11	**<.001**	-0.14	-0.28 – 0.01	**.04**
Random Parts															
σ^2		0.17			0.19			0.21			0.22			0.25	
$\tau_{00,\text{Stop}}$		0.05			0.04			0.07			0.05			0.06	
N_{Stop}		251			251			251			251			251	
ICC_{Stop}		0.22			0.19			0.24			0.17			0.18	
Observations		414			414			414			414			414	
R^2 / Ω_0^2		.52 / .39			.48 / .35			.56 / .42			.47 / .33			.47 / .34	

Table 5: Linear mixed-effects models results for judgements in Study 1.

officer utterances directed toward black drivers were perceived as less respectful, polite, friendly, formal and impartial

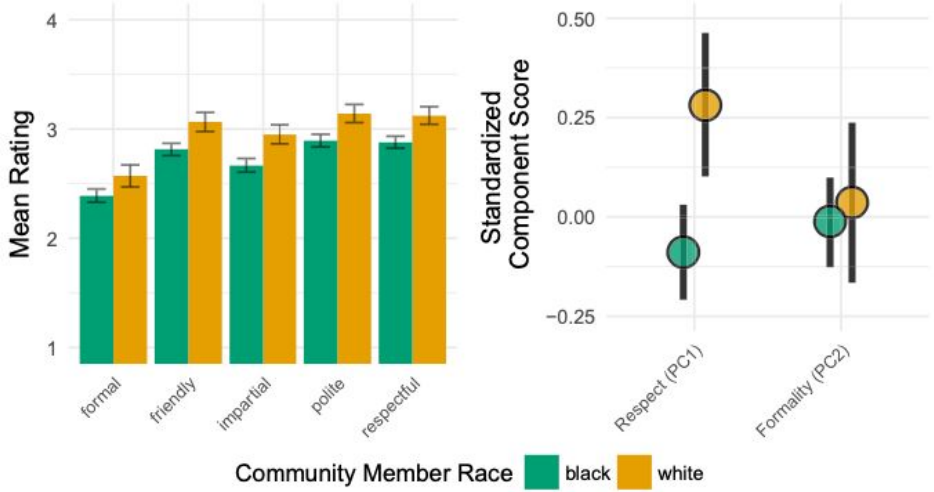
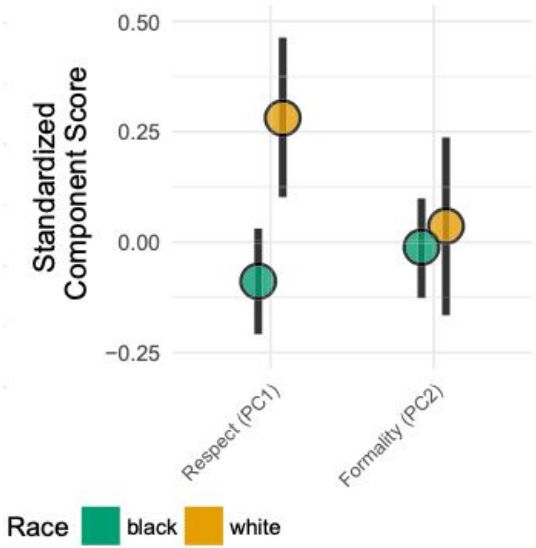


Fig. 1. (*Left*) Differences in raw participant ratings between interactions with black and white community members. (*Right*) When collapsed to two uncorrelated components, Respect and Formality, we find a significant difference for Respect but none for Formality. Error bars represent 95% confidence intervals. PC, principal component.

PCA to decompose ratings into underlying components

- Explaining 93.2% of the variance
- officers were equal in Formality with white and black drivers [beta = −0.01 (−0.19, 0.16)], but higher in Respect with white drivers [beta = 0.17 (0.00, 0.33)]

	PC1: Respect	PC2: Formality
Formal	0.272	0.913
Friendly	0.464	−0.388
Impartial	0.502	−0.113
Polite	0.487	−0.047
Respectful	0.471	0.026
% of Variance Explained	71.3%	21.9%



Study 1: Perceptions of Officer Treatment from Language.

- Conclusion:
  - key features of police treatment can be reliably gleaned from officer speech
  - Participant ratings from thin slices of police–community interactions reveal racial disparities in how respectful, impartial, polite, friendly, and formal officers' language to community members was perceived
  - Such differences were driven by differences in the Respect officers communicated toward drivers rather than the Formality with which officers addressed them
- Cons:
  - Scale (~26m vehicle stops made each year)
  - Sample too small

Study 2: Linguistic Correlates of Respect.

- Objective:
  - develop computational linguistic models of respect and formality and tune them on the 414 individual utterances
    - then apply these models to the full dataset of 36,738 utterances
- Method:
  - based on linguistic theories of respect that model how speakers use respectful language
  - extract features of the language of each officer utterance
  - log-transformed counts of these features are then used as independent variables in two linear regression models predicting the perceptual ratings of Respect and Formality from study 1

Linguistic Feature Engineering

Feature Name	Implementation	Source
Adverbial "Just"	"Just" occurs in a dependency arc as the head of an `advmod` relation	
Apologizing	Lexicon: `"sorry"`, `"oops"`, `"woops"`, `"excuse me"`, `"forgive me"`, `"apologies"`, `"apologize"`, `"my bad"`, `"my fault"`	[4]
Ask for Agency	Lexicon: `"do me a favor"`, `"let me"`, `"allow me"`, `"can i"`, `"should i"`, `"may i"`, `"might i"`, `"could i"`	[4]
Bald Command	The first word in a sentence is a bare verb with part-of-speech tag VB (`"look"`, `"give"`, `"wait"` etc.) but is not one of `"be"`, `"do"`, `"have"`, `"thank"`, `"please"`, `"hang"`.	
Colloquialism	Regular expression capturing `"y'all"`, `"ain't"` and words ending in `"in'"` such as `"walkin'"`, `"talkin'"`, etc., as marked by transcribers	
Conditional	Lexicon: `"if"`	
Disfluency	Word fragment ("Well I thi-") as indicated by transcribers	[5, 6]
Filled Pauses	Lexicon: `"um"`, `"uh"`	[7, 8]
First Names	Top 1000 most common first names from the 1990 US Census, where first letter is capitalized in transcript	[9, 10][1]
Formal Titles	Lexicon: `"sir"`, `"ma'am"`, `"maam"`, `"mister"`, `"mr*"`, `"ms*"`, `"madam"`, `"miss"`, `"gentleman"`, `"lady"`	[9, 10]
For Me	Lexicon: `"for me"`	
For You	Lexicon: `"for you"`	

Linguistic Feature Engineering

- linguistic features that received significant weights in our model of Respect
- The bars on the right show the log-odds of the relative proportion of interactions taken up by each feature

	Respect			*Formality*		
	β	CI	p	β	CI	p
Fixed Parts						
(Intercept)	-0.18	-0.36 – 0.00	.052	0.26	0.07 – 0.45	**.008**
Adverbial "Just"	0.24	-0.07 – 0.53	.118			
Apologizing	1.34	0.15 – 2.52	**.027**	-1.56	-2.80 – -0.32	**.014**
Ask for Agency	-0.34	-0.90 – 0.22	.230	0.37	-0.23 – 0.96	.225
Bald Commands				-0.25	-0.68 – 0.18	.255
Colloquialism				-1.10	-1.97 – -0.23	**.013**
Conditional				-0.27	-0.74 – 0.21	.271

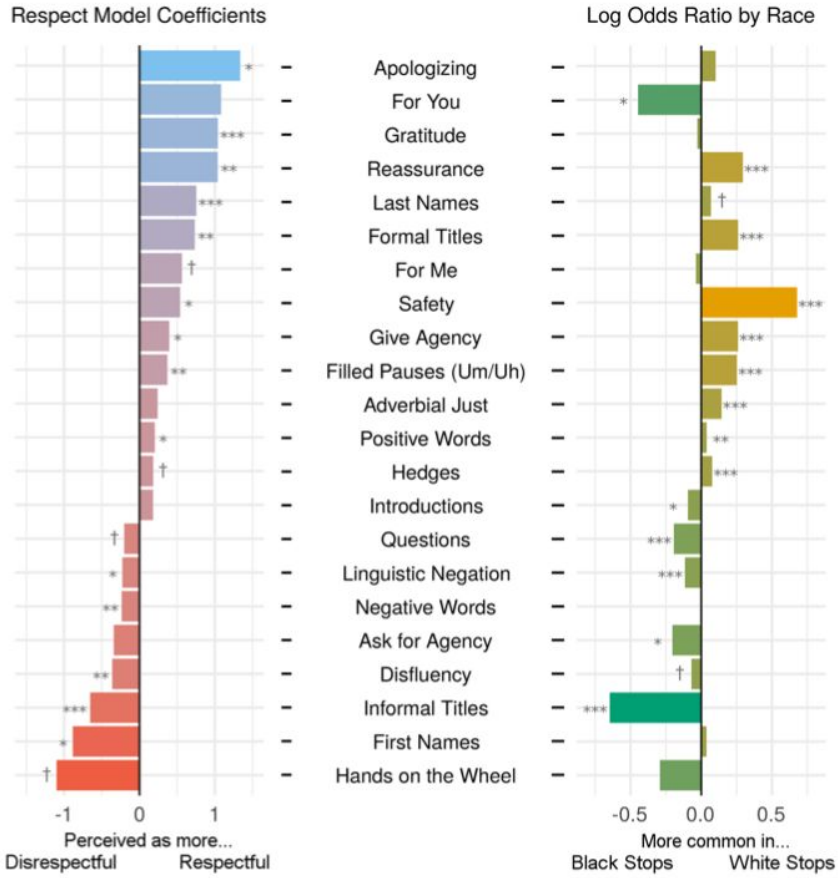
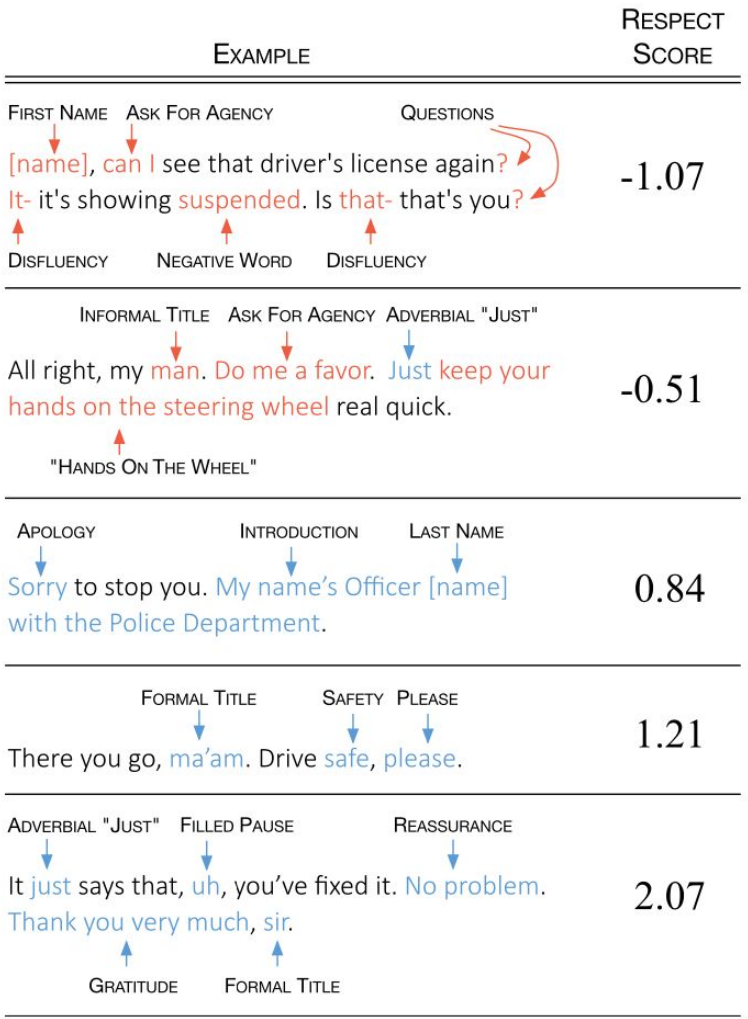


Fig. 2. (*Left*) Respect weights assigned by final model to linguistic features and (*Right*) the corresponding log-odds of those features occurring in officer speech directed toward black versus white community members, calculated using Fisher's exact test. $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Respect LR Model

- Sample sentences with automatically generated Respect scores.
- Features in blue have positive coefficients in the model and connote respect



Study 2: Linguistic Correlates of Respect.

- Results:
  - Our model for Respect obtains an adjusted R2 of 0.258 on the perceptual ratings obtained in study 1, and a root-mean-square error (RMSE) of 0.840, compared with an RMSE of 0.842 for the average rater relative to other raters.
  - Our model for Formality obtains an adjusted R2 of 0.190, and an RMSE of 0.882 com- pared with 0.764 for the average rater
- Conclusion:
  - Our model-assigned ratings agree with the average human from study 1 about as well as humans agree with each other
  - a constrained set of objectively measurable linguistic features can explain a meaningful portion of the variance in these ratings.

Study 3: Racial Disparities in Respect.

- Objective:
  - **Controlling for contextual factors** of the interaction, is officers' language more respectful when speaking to white as opposed to black community members?
- Method:
  - apply the LR models from study 2 to the entire corpus of the 36,738 utterances
  - build linear mixed-effects models
  - include covariates: community member race, age, and gender; officer race; whether a search was conducted; and the result of the stop (warning, citation, or arrest)
  - include random intercepts for interactions nested within officers

Study 3: Racial Disparities in Respect.

- Results:
  - Respect
    - utterances spoken by officers to **white** community members score higher in Respect [beta = 0.05 (0.03, 0.08)]
    - **older** [beta = 0.07 (0.05, 0.09)]
    - when a **citation** was issued [beta = 0.04 (0.02, 0.06)]
    - lower in stops where a **search** was conducted [beta = −0.08 (−0.11, −0.05)]
  - Formality
    - higher **crime rate** [beta = 0.03 (0.01, 0.05)]
    - **older** [beta = 0.05 (0.03, 0.07)]
    - **female** [beta = 0.02 (0.00, 0.04)]

Are the racial disparities in the respectfulness of officer speech we observe driven by a small number of officers?

- calculated the officer-level difference between white and black stops for every officer (N = 90) in the dataset who had interactions with both blacks and whites
- find a roughly normal distribution of these deltas for officers of all races.

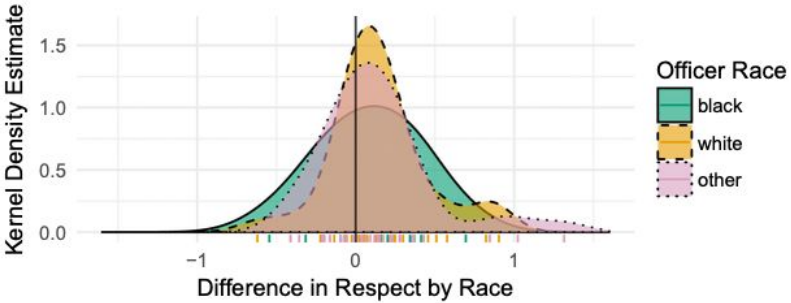


Fig. 4. Kernel density estimate of individual officer-level differences in Respect when talking to white as opposed to black community members, for the 90 officers in our dataset who have interactions with both blacks and whites. More positive numbers on the *x* axis represent a greater positive shift in Respect toward white community members.

Growth-curve analyses revealed that officers spoke with greater Respect [b = 0.35 (0.29, 0.40)] and reduced Formality [b = −0.57 (−0.62, −0.53)] as interactions progressed

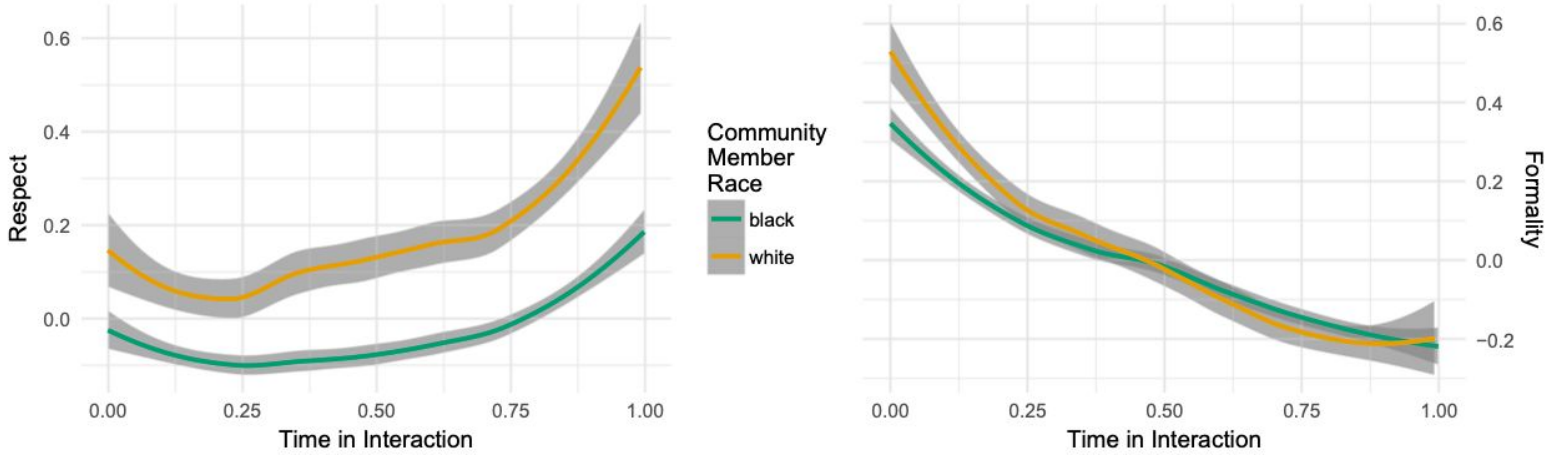
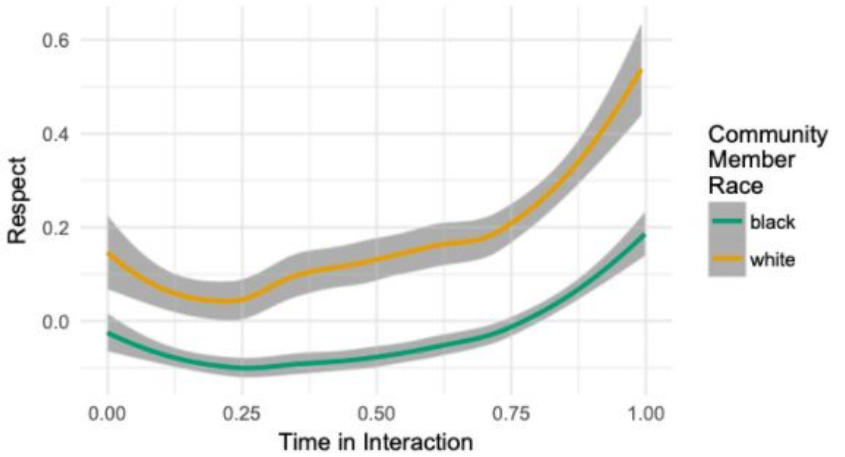


Fig. 5. Loess-smoothed estimates of the (*Left*) Respect and (*Right*) Formality of officers' utterances relative to the point in an interaction at which they occur. Respect tends to start low and increase over an interaction, whereas the opposite is true for Formality. The race discrepancy in Respect is consistent throughout the interactions in our dataset.

Growth-curve analyses revealed that officers spoke with greater Respect and reduced Formality as interactions progressed

- stops of white and black drivers converged in the Formality expressed during the interaction
- gap in officer Respect increased over time [b = 0.10 (0.05, 0.15)]
    - Respect increased more quickly in interactions with white drivers [b = 0.45 (0.38, 0.54)] than in interactions with black drivers [b = 0.24 (0.19, 0.29)]



Conclusion: Officers' language is less respectful when speaking to black community members.

- first showed that people make consistent judgments about such interactions from officers' language, and we identified two underlying, uncorrelated constructs perceived by participants: Respect and Formality
- then built computational linguistic models of these constructs, identifying crucial positive and negative politeness strategies in the police– community interactional context
- Applying these models to an entire month of vehicle stops, we showed strong evidence for racial disparities in Respect, but not in Formality

Future directions

- expand body camera analysis beyond text to include information from the audio such as speech intonation and emotional prosody, and video, such as the citizen’s facial expressions and body movement
- footage analysis could help us better understand what linguistic acts lead interactions to go well, which can inform police training and quantify its impacts over time
- More complicated models can be adopted for the linguistic language analysis

APPENDIX

2.4 Full Regression Model Output

	Respect			*Formality*		
	β	CI	p	β	CI	p
Fixed Parts						
Arrest Occurred	-0.08	-0.20 – 0.04	.210	0.04	-0.09 – 0.17	.532
Citation Issued	0.05	-0.06 – 0.16	.387	0.13	0.02 – 0.25	**.023**
Search Conducted	-0.23	-0.34 – -0.11	**<.001**	0.04	-0.08 – 0.17	.470
Age	0.05	-0.05 – 0.15	.321	0.11	0.01 – 0.21	**.036**
Gender (F)	-0.03	-0.12 – 0.07	.608	0.09	-0.01 – 0.19	.089
Race (W)	0.17	0.00 – 0.33	**.046**	-0.01	-0.19 – 0.16	.873
Officer Race (B)	-0.03	-0.18 – 0.11	.646	0.04	-0.11 – 0.20	.565
Officer Race (O)	0.00	-0.15 – 0.14	.966	-0.08	-0.23 – 0.07	.291
Officer Race (B) : Race (W)	0.02	-0.12 – 0.16	.799	-0.03	-0.18 – 0.11	.658
Officer Race (O) : Race (W)	-0.07	-0.22 – 0.09	.405	0.01	-0.15 – 0.18	.869
Random Parts						
σ^2		0.751			0.870	
$\tau_{00,\text{Stop:Officer}}$		0.010			0.000	
$\tau_{00,\text{Officer}}$		0.115			0.107	
$N_{\text{Stop:Officer}}$		254			254	
N_{Officer}		118			118	
$ICC_{\text{Stop:Officer}}$		0.011			0.000	
ICC_{Officer}		0.132			0.110	
Observations		414			414	
R^2 / Ω_0^2		.358 / .335			.255 / .213	

Table 7: Mixed-effects regression outputs on observed ratings from participants in Study 1 for models with *Respect* and *Formality* (PC1 and PC2) as dependent variables; fixed effects for the community member's race, age, and gender; and random effects at the officer and interaction level. Reference levels are black male community members, a white officer, and a warning issued with no citation, arrest, or search. Standardized coefficients are reported. P-values computed via the Wald-statistics approximation with the sjPlot R Package [2].

Total Vehicle Stops in April 2014	2159	
	Black	White
Race of Community Member	998	422
Unsuccessful Matches		
Officer Body-Worn Camera Not Activated	1	1
Video File Could Not be Opened	3	2
No Body-Worn Camera Issued to Officer	48	35
Could Not Locate File	63	32
Stops Matched	883	352
Proportion of Total Stops Matched	0.884	0.834
Stops Marked Ineligible for Transcription		
Single Video Does Not Capture Entire Duration of Stop	22	3
Recording Officer Not Primary Interlocutor	160	41
Stops Transcribed	701	308
Proportion of Total Stops Transcribed	0.702	0.729
Transcribed Stops Excluded From Analysis		
Fewer than 3 Turns	19	9
Stops in Dataset	682	299
Proportion of Total Stops in Dataset	0.683	0.709

Table 1: Accounting of all vehicle stops conducted by the Oakland Police Department in April 2014, and the sampling process by which they were included in the final dataset. We attempted to obtain as clean and complete a full sample of all vehicle stops of black and white community members conducted as possible.