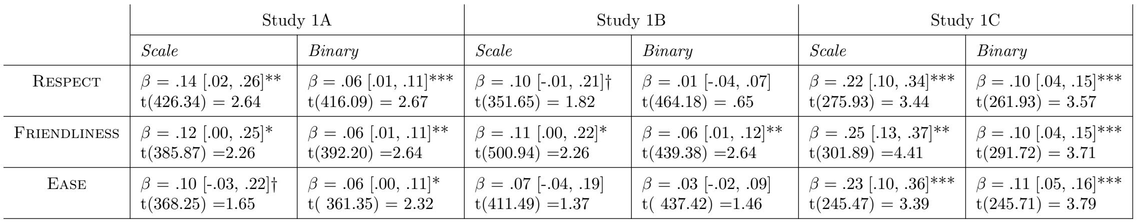
Supplemental Materials for The Thin Blue Waveform:

Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police



	Study 1A		Study 1B		Study 1C	
	Scale	*Binary*	*Scale*	*Binary*	*Scale*	*Binary*
Respect	β = .14 [.02, .26]** t(426.34) = 2.64	β = .06 [.01, .11]*** t(416.09) = 2.67	β = .10 [-.01, .21]† t(351.65) = 1.82	β = .01 [-.04, .07] t(464.18) = .65	β = .22 [.10, .34]*** t(275.93) = 3.44	β = .10 [.04, .15]*** t(261.93) = 3.57
Friendliness	β = .12 [.00, .25]* t(385.87) =2.26	β = .06 [.01, .11]** t(392.20) =2.64	β = .11 [.00, .22]* t(500.94) =2.26	β = .06 [.01, .12]** t(439.38) =2.64	β = .25 [.13, .37]** t(301.89) =4.41	β = .10 [.04, .15]*** t(291.72) = 3.71
Ease	β = .10 [-.03, .22]† t(368.25) =1.65	β = .06 [.00, .11]* t(361.35) = 2.32	β = .07 [-.04, .19] t(411.49) =1.37	β = .03 [-.02, .09] t(437.42) =1.46	β = .23 [.10, .36]*** t(245.47) = 3.39	β = .11 [.05, .16]*** t(245.71) = 3.79

Figure S1. Fixed effect estimates for racial disparities in officer speech in Studies 1A-1C, controlling for random intercepts for participants and stimuli.

Additional Analyses for Studies 1A-1C

In the main text, we consider interpersonal treatment as a composite judgment of officer prosody: the extent to which an officer's tone communicates respect, ease, and friendliness towards their interlocutor. Figure S1 displays the statistics for each dimension and each study individually, as well as the results for participants' categorical judgments of officer prosody (derived from a binary logistic mixed-effects model). However, we urge caution in interpreting binary measures for Study 1B in light of a large number of missing responses (16.7% of scale responses, 26.4% of categorical responses). This data appeared to be missing at random.

We used multiple imputation via chained equations to impute values for missing cases for our scale items (MICE; Graham, 2009). This technique replaces missing values from a distribution of plausible values for each missing entry. The analyses using different iterations of this process are then pooled, generating parameter estimates that preserve characteristics of the larger dataset (Graham, 2009). We conducted 20 permutations of the data with 10 maximum iterations using the mice R package (Buuren & Groothuis-Oudshoorn, 2010). In each permutation, we estimated missing values for our three scale variables and computed the treatment composite score. Aggregating these 20 permutations confirmed that the main effect of driver race was significant after imputing missing values, β= .20 [.11, .29], t(13274.80)=4.29, p<.001.

Supplemental Experiment

Studies 1A-1C reveal racial disparities in officer prosody present in thin slices of officer speech of approximately ten seconds. These clips preserved the cadence and tone of natural conversation by extending across several utterances. However, in order to blind participants to the driver's race, we removed a speaker from the two-party conversation between officer and citizen. Since an officer's communication is, in part, a response to their interlocutor, disparities in Studies 1A-1C could have reflected an officer's reaction to the driver. Further, these omissions may have created artifacts, such as long pauses between officer utterances or silenced responses to officers' questions, that affected participants' judgments.

In a supplemental experiment, we conducted a conservative test of racial disparities with a highly constrained set of content-filtered stimuli: single sentences communicating the reason the officer stopped the driver and utterances requesting the driver's documents. These are not only they most common speech acts in police stops (Bayley, 1986; Prabhakaran et al., 2018)(Bayley, 1986; Prabhakaran et al., 2018), but also represent two distinct linguistic sentence types with regards to intonation and discourse functions in this context: declarative statements assigning blame (stop justification) and interrogative requests (asking for documents).

Stimuli Generation

We sampled stimuli from the corpus of speech used in Studies 1A-1C. Rather than random windows of officer speech, we used transcriptions of the stops in our corpus to identify officer turns that contained a set of required lemmas for each of two acts. For document requests, we searched for utterances which contained the words “license”, “registration”, and “insurance” (e.g., “Can I see your license, uh, proof of insurance, and registration, please?”; “Do you have your license, insurance, registration on you?”). We selected utterances for stop justification that contained either the lemmas “reason”, “I”, and “stop” (e.g., “The reason I stopped you is you were talking on your cellphone.”), or “reason”, “pull”, and “you” (e.g. “Um, like I said, the reason why I pulled you over was because of it’s distracted driving month.”).

After manually checking that the identified utterances were from either document requests or stop explanations, we randomly selected fifty Black-directed and fifty White-directed utterances for each act. As anticipated, these stimuli were similar in their content: 63.1% of words in document request clips and 34.9% of words in the stop justification clips. These stimuli were further controlled in that the Black-directed and White directed speech acts occurred in approximately the same position in the larger interactions from which they were sampled.; We then isolated the portion of the body camera recordings that contained each utterance and applied the same content-filter used in Studies 1A-1C. overlapped.

Procedure

51 undergraduates participated for course credit or payment (32 women, 19 men) in Study 2. The racial composition of the sample was 39% white/Caucasian, 8% Latinx, 6% black/African-American, 24% Asian, and 23% multiracial/some other group, and the

mean age was 19.28 (SD=1.37). Participants were randomly assigned to one of two replications, each containing half of the stimuli presented in a random order. Thus, each participant rated 25 Black-directed and 25 White-directed clips for each act, or 100 clips in total.

The supplemental study followed the same rating procedure as Study 1, with two exceptions. First, prior to each stimulus being presented, participants saw a screen that identified the speech act they would be hearing. Second, given the constrained nature of our stimuli, we adjusted the six-point scale used in Study 1 to a three-point scale (e.g. *Cold-Friendly-Neutral*).

Results

We recoded participants' categorical judgments on each dimension on a -1 (*Talking Down, Unfriendly, Tense*) to 1 (*Respectful, Friendly, At Ease*) scale, then took the average of these measure to form a composite measure of interpersonal treatment (α= .79). As in Study 1, we analyzed our results with a linear mixed-effects model with cross-specified random effects of participants and stimuli. We further added a fixed effect term for act (effects coded, -1= Requesting Documents, 1=Providing Reason).

Consistent with the results reported in the body of the paper, officer prosody towards White drivers was judged as communicating more positive interpersonal treatment, β= .09 [.00, .17], t(197.1)=1.94, p=.05. This disparity was not moderated by speech act, although participants viewed officer's prosody more positively for document requests than stop explanations β= .13 [.04, .22], t(197.1)=2.97, p<.001.

Comparison of Condition-level Classification Images in Study 2

In addition to the classification image rating study described the main text, we conducted an additional comparison of CIs that were generated by condition, rather than by participant. Since condition-level CIs aggregate across participants, they produce less noisy images. For the same reason, however, they increase the likelihood of Type I errors since variation among participant representation is ignored (Cone et al., 2020). While the study presented in the main text is a more stringent test of prosody's influence on representations of the police, we describe the condition-level CI comparisons here by way of convergent evidence for our hypotheses.

The procedure of this study was identical to the image-rating phase in the main text, except that all participants compared the same two CIs: a CI generated from the positive prosody condition of the image-generation phase, and a CI generated from the negative prosody condition. In order to have at least 80% power to detect a small-to-moderate difference between the two images (d=.35), we set a recruitment goal of 70 participants. 85 participants were recruited via Amazon Mechanical Turk (MTurk), but 18 failed an attention check and were excluded prior to analysis, resulting in a final sample size of 69 (M_{age}=44.0, SD=11.24; N=32 Female). This sample was 12.9% Asian, 5.9% Black/African-American, 4.7% Latinx, 2.4% Native American, 65.9% White/Caucasian, and 8.2% Multiracial or some other race.

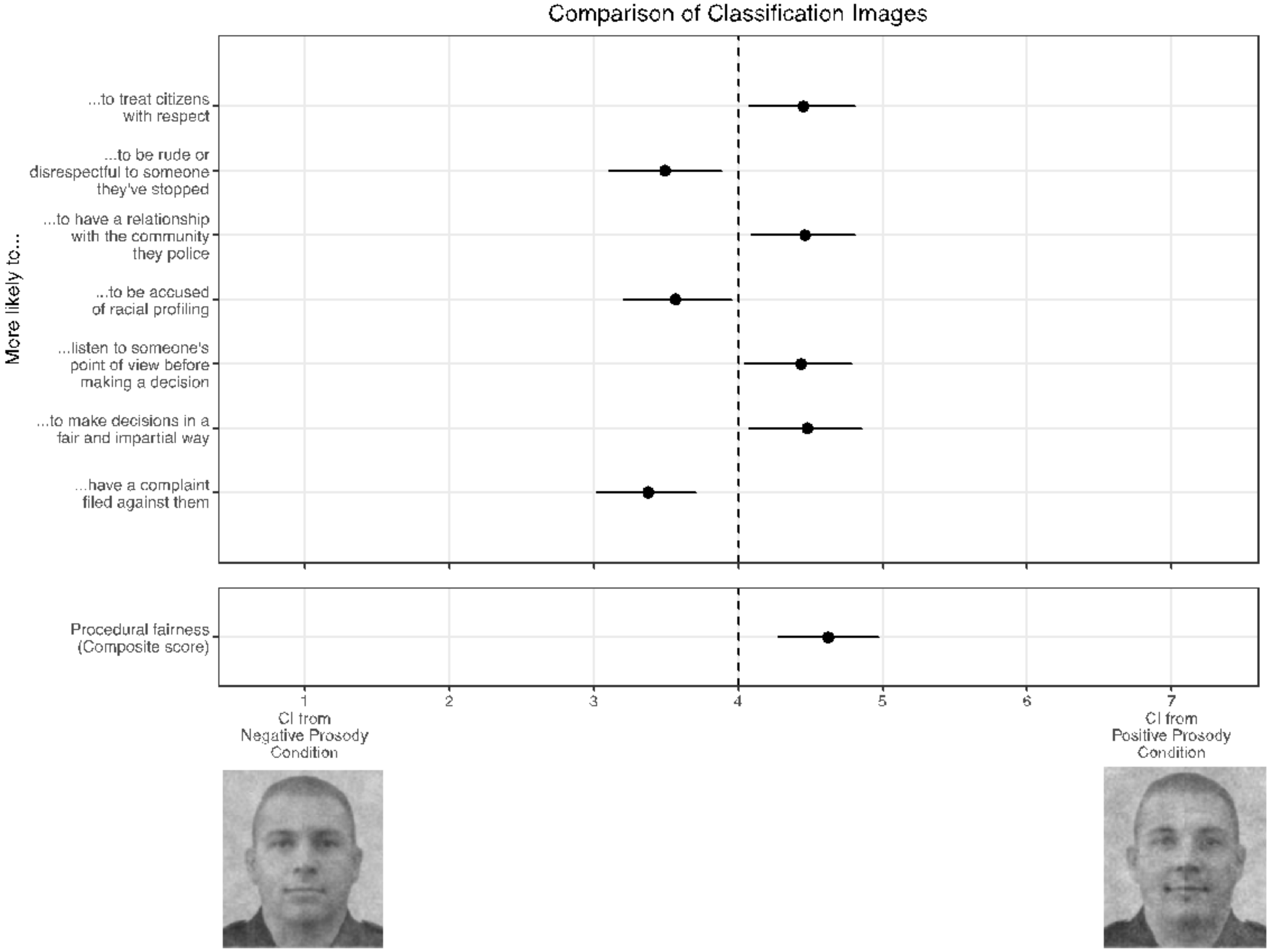


Figure SEQ Figure 2. Classification images created by participants during the image-generation phase of Study 2 alongside rating-phase participants' judgments. Bars represent 95% confidence intervals, and the dotted line represents the point of indifference between the classification images.

Results

Figure 2 displays the results for all items, along with the classification images. As in Study 2 in the main text, we combined these items into a single index of procedural fairness (α= .97), then tested whether participants' ratings differed from the midpoint of the scale (i.e. indifference between the CIs). Since there were only two CIs to compare, we conducted a one-sample t test rather than a linear mixed-effects regression. Consistent with the findings in the main text, the CI from the positive prosody condition generated was judged as more procedurally fair than the CI from the negative prosody condition (M=2.38 [2.00, 2.75], t(68)=3.29, p<.01, d=.37).

References

Bayley, D. H. (1986). The tactical choices of police patrol officers. *Journal of Criminal Justice*, *14*(4), 329–348.

Buuren, S. van, & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2020). Type I Error Is Inflated in the Two-Phase Reverse Correlation Procedure. *Social Psychological and Personality Science*, 1948550620938616.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Prabhakaran, V., Griffiths, C., Su, H., Verma, P., Morgan, N., Eberhardt, J. L., & Jurafsky, D. (2018). Detecting Institutional Dialog Acts in Police Traffic Stops. *Transactions of the Association for Computational Linguistics*, *6*, 467–481.