Supplementary Information for

Race-based disparities in academic disciplinary actions are associated with county-level rates of racial bias

Travis Riddle and Stacey Sinclair

Travis Riddle.
E-mail: triddle@princeton.edu

This PDF file includes:

Supplementary text
Figs. S1 to S7
Tables S1 to S3
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1808307116

Supporting Information Text

The OSF page for this project (https://osf.io/pu79a/) hosts many additional details. These include interactive maps for all disciplinary outcomes, interactive parameter plots for models fit in the course of this work, information for a preregistered version of these analyses, explanations for departures from these plans, code, and aggregated data. As of this writing, live versions of the maps and parameter plots can be found at http://www.travisriddle.com and https://triddle.shinyapps.io/riddle_sinclair/, respectively, though the OSF will serve as a permanent repository for this information.

Analytic Approach. Because Project Implicit is a nonrandom sample, we used multilevel regression and post-stratification (MRP) to obtain accurate geographical population-based estimates of implicit and explicit bias. This procedure corrects for biased sampling and regularizes extreme observations with little data to support them (e.g. a county with only a handful of respondents with especially high or low scores) (1, 2). Following past work (3), we identified age as one dimension along which IAT respondents differed from the general population in ways that could bias our conclusions (4). Our post-stratification weighting scheme is as follows: We first grouped respondents into five age group categories (15-24, 25-34, 35-54, 55-75, and 75+). We next fit multilevel models estimating bias (implicit and explicit biases separately) as a function of our state-level covariates (the "fixed" effects*), and allowed the estimates to vary by age bin, county, and state (the "random" effects). We fit separate models for each of the two years of analysis (2014 and 2016) Next, we determined the population of whites in each county in these age groups using the American Community Survey's 5-year estimates ending in 2014 and 2016. Finally, we used our estimated models to predict the expected response for each age bin, in each county, for the corresponding year. Our final county-level estimates are the average of the values predicted for the 5 age bins, weighted by the population size of that bin in that county. As a result of this procedure, we can be confident that our county-level estimates should more closely approximate what our estimates would look like if the Project Implicit data were truly representative along the age dimension in all counties.

We analyzed these data using a series of bayesian multilevel logistic regressions. We modeled the probability that a student would be disciplined as a function of a set of effects that are constant across observations (i.e. fixed effects: student race (dummy coded), implicit bias, explicit bias, an interaction term between student race and implicit bias, an interaction term between student race and explicit bias, and all covariates described below, plus each of the covariates interacting with student race) and a set of effects that vary across counties (i.e. random effects: overall intercept & student race). We fit separate models for each of the outcomes, and for each of the two years of analysis (2014 & 2016). All numerical predictors were standardized at the appropriate level (county, state) before model estimation (for post-stratification as well as for final inference) to help with estimation efficiency and interpretability. We also set priors for the intercept and coefficients in the bayesian model to be weakly informative normal distributions centered on zero with a standard deviation of five. This corresponds to a prior belief that all parameters take values between -15 and 15 with a probability of more than .99. Realistically, values outside this range are extremely unlikely given that all variables were standardized prior to estimation. Priors on all other parameters (e.g. variance of the county-level intercepts) were left to default settings from the software package used to fit the models

Because of the computational demands of fitting such a high-dimensional model to such a large dataset (the full model for each metric would consist of over 6k parameters to approximately 170k observations), we used a consensus monte carlo algorithm to obtain approximate posterior distributions for the parameters of interest (5). The approximate posteriors derived from this algorithm have been shown to be nearly indistinguishable from the true posterior, a result we verified using a small subset of our own data.

Implicit-Explicit Association. As described in the main text, there are two reasons for the high degree of county-level associations between implicit and explicit bias. First, these measures are correlated at the individual level (see main text). Second, the state-level predictors in the MRP models are similarly associated with each of the outcomes. In figure S1, we show the coefficient estimates for the state-level predictors in the post-stratification models. Comparing estimates for implicit and explicit bias within racial and sexuality bias, it's apparent that the coefficients are nearly always directionally similar, with the coefficients of the largest magnitude having estimates that are close in absolute value. Although we attempted to estimate models which could account for the individual-level correlation by including both explicit and implicit measures in the same model, we found that these models would not converge.

2013-2014 Data Sources and Description.

Disciplinary Actions. The 2013-2014 CRDC data is similar to that collected in 2015-2016 reported in the main text. In the 2013-2014 data, there are 95507 institutions enrolling approximately 50 million students, of which approximately 25.2 million are white and 7.8 million are Black. Using the same exclusions applied to the 2015-2016 data, the final sample used for modeling consists of 93493 institutions, enrolling 49.8 million students of which 25.1 million are white and 7.7 million are black. The percentage of students receiving each of these disciplinary actions in the 2013-2014 school year is shown in table S1

Racial Bias. The racial bias measures are derived from the same data as reported in the main text. For the 2013-2014 analysis, we restricted the sample to only those individuals who responded through the end of 2014. This consisted of approximately 1.3 million total respondents from 3099 counties. 1.19 million respondents provided data for the IAT and 1.15 million provided explicit bias ratings.

*We used a subset of the variables included as covariates in the final model. All covariates are displayed in S1

Sexuality Bias. The sexuality bias measures are derived from the same data as reported in the main text. For the 2013-2014 analysis, we restricted the sample to only those individuals who responded through the end of 2014. This consisted of approximately 997K total respondents from 3044 counties. Of these respondents, 890K respondents provided data for the IAT, and 948K respondents provided explicit bias ratings.

Covariates. Each county-level variable used as a covariate in the final model and the corresponding state-level variable used as a predictor in the post-stratification scheme (described below) were taken from the same source. Population size and proportions, socioeconomic indicators, mobility, and segregation indices were all taken from the American Community Survey (ACS) 5-year estimates for the time period ending in 2014. Urban-rural indicators were taken from the 2010 US Census, and crime rates were taken from the FBI Uniform Crime Reporting program, as made available through the National Archive of Criminal Justice Data for each year from 2010-2014. Each of these variables is described below.

Population size and proportions We obtained the total population, the proportion of the population that is white, the proportion of the population that is black, and the ratio of black-to-white people in the population from ACS table B02001.

Socioeconomic indicators For the state-level socioeconomic predictors for the post-stratified estimates, we obtained estimates for the percentage of population with a Bachelor's degree or higher, the percentage of the population aged 16 or over in the labor force that is unemployed, the median household income, and the percentage of families and people whose income in the last year was below the poverty line from the ACS table DP03. For the county-level data, we used the same set of variables described above and also included estimates of the *difference* between blacks and whites in each county on these socioeconomic indicators, using multiple imputation to fill in observations for some counties where estimates were unavailable.

Urban-rural indicator We obtained estimates of housing density per square mile of land area from Census table GCT-PH1.

Mobility We obtained estimates of population mobility by summing the percentage of Black Americans who moved from a different county, state, or country into the county of interest (county-level covariate) or who moved from a different state or country into the state of interest (state-level covariate). We took these metrics from the ACS table S0701.

Crime We computed estimates of the number of violent crimes per person by taking the number of crimes reported divided by the population size for each year, and averaging the resulting proportions across the 5 years of data.

Segregation We computed a dissimiliarity index as described by (6) metric reflects the proportion of a racial group within the county that would have to move in order for all census tracts to have group distributions that matched the overall distribution of the county. These computations were done using data from the ACS table B02001.

Implicit and Explicit bias in the same model. In the main text, we presented results from analyses with the associations of implicit and explicit bias were modeled separately. This was on account of the high degree of collinearity in these predictors. Our findings are not substantively changed when examining results from models where both measures were entered as predictors.

Figure S2 presents parameter estimates from the full model and from the separated models from the main text. Estimates are displayed for both the 2014 and 2016 data collections. As in the model with implicit and explicit bias entered separately, the full model is generally consistent with a reliable association between disciplinary disparities and explicit bias. However, implicit bias shows a less consistent relationship.

Sexuality Bias. Figure S3 shows the estimate of primary interest for each of the sexuality models. The estimates displayed are the coefficients for the interaction between race and each of the two bias measurements. This figure illustrates that in general, county-level explicit and implicit biases in favor of straight individuals are not as consistently associated with racial disciplinary disparities.

Moderation Analyses. To test whether the associations between bias and disciplinary disparities varied as a function of other meaningful variables, we estimated a series of models examining whether these effects were moderated by a subset of covariates that were included as controls in the model. These variables were selected because we believed that the degree of moderation along these dimensions would be theoretically interesting.

Analytic details. As reported in the main text, we examined the proportion of the population that is white, the dissimilarity index, and a variable that captured the gap in socioeconomic status between blacks and whites in a county. We created the SES gap variable by taking the difference between whites and blacks on the county-level measures of median income, percentage of college graduates, poverty rate, and unemployment rate after filling in missing values with multiple imputation. Using these gap variables, we performed a principal component analysis and used the first principal component as the variable to test as a moderator.

The parameters of interest for each of these models are the three-way interactions between bias, student racial group, and the covariate. This parameter reflects further deflections from the two-way interaction between bias and student racial groups. As described in the main text of the manuscript, negative values for this two way interaction indicate that as the level of bias in a county increases, the gap between black students (the baseline group) and white students also increases. Therefore, for any cases in which the value of the three-way parameter is reliably negative, this would indicate that as one moved into counties that were higher in the covariate of interest (i.e. a larger proportion of whites, a larger SES gap, or more segregation), the association between bias and the disciplinary gap also grows. These three-way parameters are plotted in figures S4 and S5 for the 2013-2014 and 2015-2016 data collections, respectively. As reported in the main text, there does not appear to be any wholly consistent moderation across years and outcomes for a given moderator.

Alternative modeling strategies. As with any set of analyses, other researchers might have opted to analyze these data differently. To alleviate concerns that our analyses are showing relationships that would not exist under other analytic regimes, we fit several other models to show that the results are qualitatively similar when different choices are made. We limited these analyses to the 2013-2014 data only. We proceed to show two sets of results. In each set of results, we estimate models using 1) frequentist and bayesian techniques, 2) county-level estimates based on the MRP and based on the naiive county means, and 3) with and without covariates. The first set of these results is based on an approach in which data is first aggregated to the county level before estimating a more commonplace linear regression. The second set of results repeats the analysis presented in the main text of our paper, but with all of the above variants included for comparison.

Aggregate and linear regression. The goal of this analysis was to examine relevant relationships in the simplest manner possible. So we did not trim outliers or small counties from the data as others who have used linear regression in similar circumstances have (7), thereby potentially yielding relatively imprecise parameter estimates.

Because the data come to us in the form of counts of students enrolled and counts of students disciplined, one way to simplify the model is to aggregate this information to the county-level such that the disciplinary disparity is represented by one number at the county level. Doing so then allows one to estimate a standard linear regression. To accomplish this, we computed the school-level proportion of white and black students disciplined, computed the difference between the proportion of black and white students disciplined, and then computed an average difference within each county. We then proceeded with the set of analyses described above.

Figure S6 shows, from each model, the parameter that corresponds to the association between bias and the black-white difference in proportion of students disciplined. For this outcome, negative values indicate greater disciplining rates for black students, and so negative values of this parameter correspond to the hypothesis that greater levels of bias are associated with larger disciplinary disparities. Looking specifically at the values for explicit bias, there are just two model specifications for which the estimate is not directionally consistent with our conclusions in the main paper (the two models with explicit bias computed with MRP as associated with law enforcement referral without covariates).

Multilevel model. Figure S7 shows the estimates derived from the above described variants of the multilevel model. As in the main text of the paper, the parameter we focus on here represents the change in the log-odds slope for white students in comparison to black students. More negative values of this parameter thus reflect a pattern of data in which as one moves into counties that have higher levels of bias, the gap between the rate at which black students and white students are disciplined grows. In this figure, the estimates for the bayesian explicit and implicit MRP models with covariates are the ones presented in the main text of the paper. As before, looking at the values for explicit bias, there is just one model specifications for which the estimate is not directionally consistent with our conclusions in the main paper (the bayesian MRP model without covariates).

Tables and Figures

Table S1. Percentage of students of each race receiving each type of disciplinary action in 2013-2014

metric	black	white
school arrests	0.29%	0.09%
expulsions	0.48%	0.22%
law enforcement referral	0.75%	0.34%
in-school suspension	11.11%	4.23%
out-of-school suspension	13.87%	3.63%

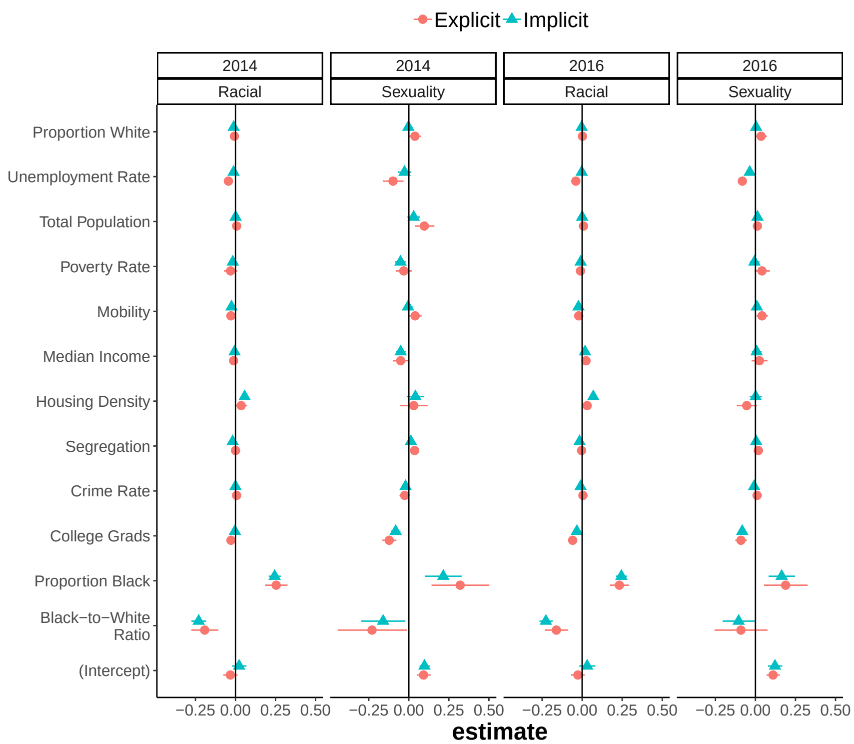

Fig. S1. Coefficient estimates for the state-level predictors in the models used for post-stratification. All variables were standardized prior to estimation to facilitate comparison across models and terms. Bars represent +/- 1 standard error. Plot is faceted by bias target and year.

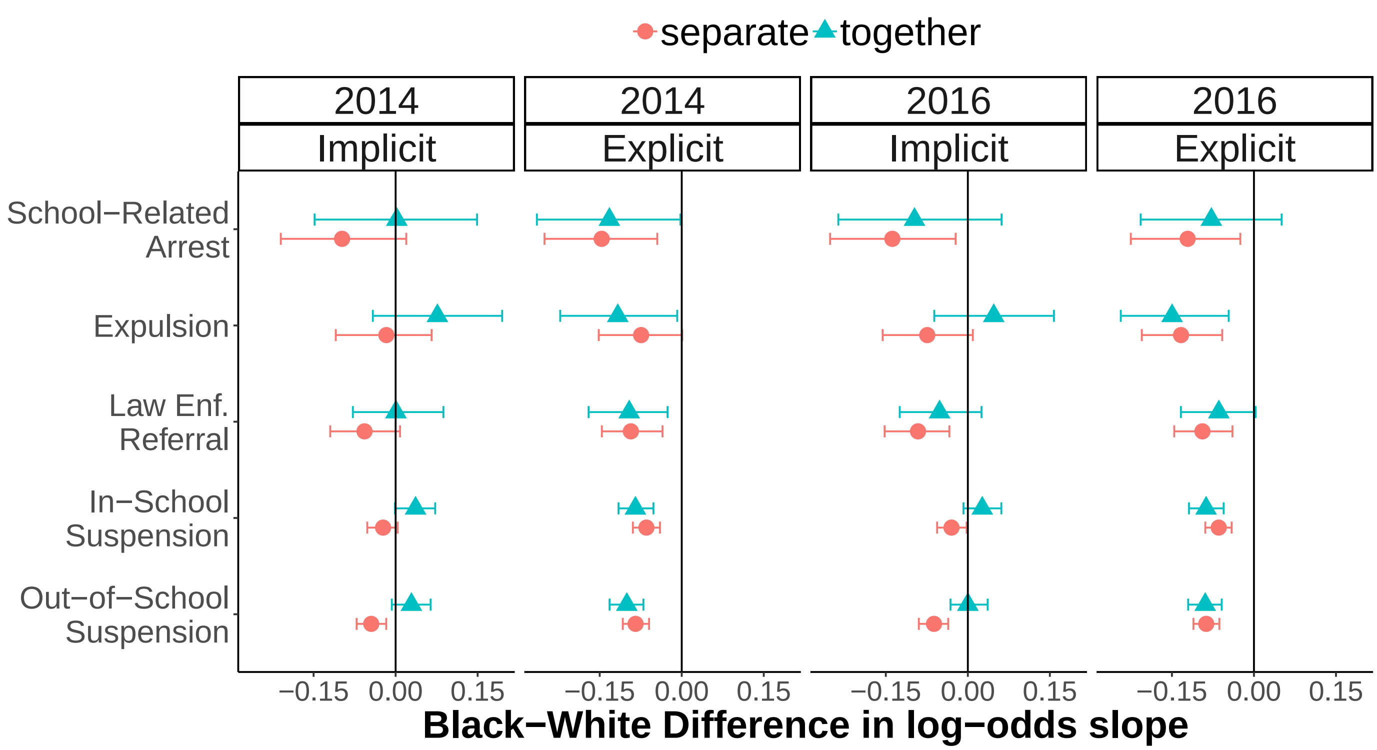

Fig. S2. Association between each metric and county-level estimates of explicit and implicit racial bias for statistical models that include both variables (i.e. full) and with them separated. Estimates are displayed for both data collection periods. Negative values indicate that the rate of increase (or decrease) for blacks is faster (or slower) than for whites. Point is the mean of the posterior and error bars represent 95% highest density intervals.

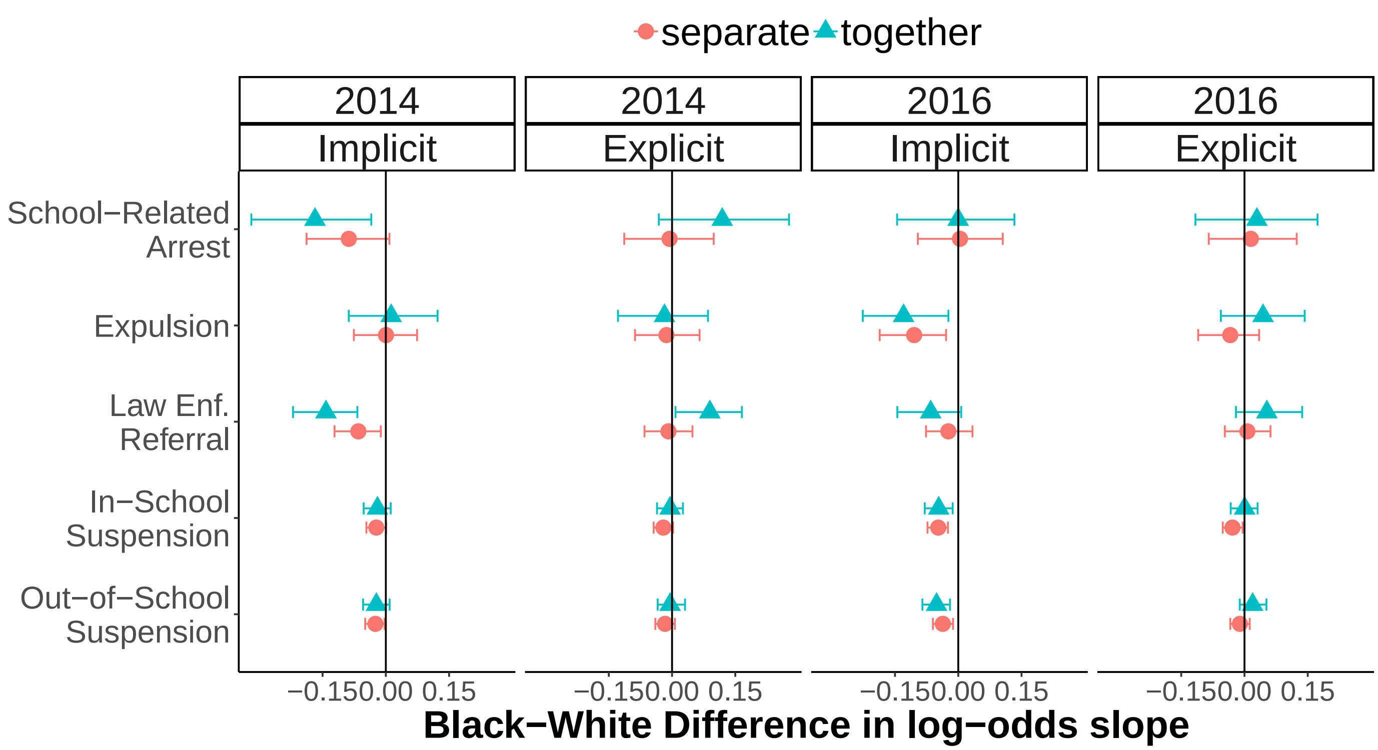

Fig. S3. Association between each metric and county-level estimates of explicit and implicit sexuality bias for statistical models that include both variables (i.e. full) and with them separated. Estimates are displayed for both data collection periods. Negative values indicate that the rate of increase (or decrease) for blacks is faster (or slower) than for whites. Point is the mean of the posterior and error bars represent 95% highest density intervals.

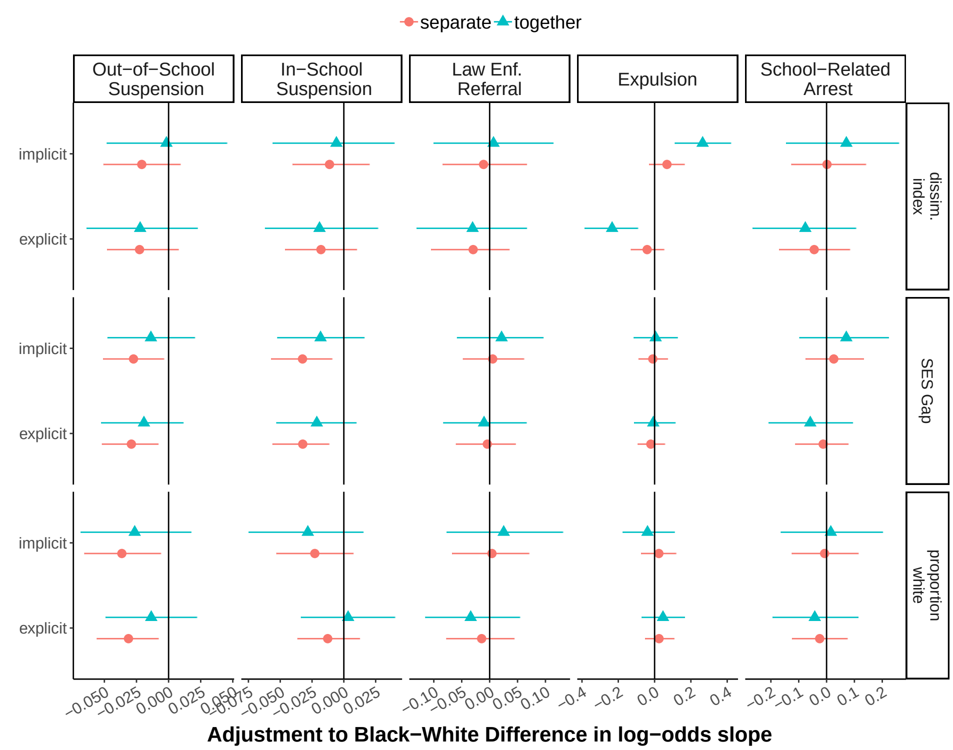

Fig. S4. Parameter estimates for moderators of the associations between bias and disciplinary actions in the 2013-2014 data. Point is the mean of the posterior. Error bars indicate 95% highest density intervals. Negative values indicate that the association between bias and disciplinary disparities becomes stronger as one moves into counties with higher levels of the indicated moderator

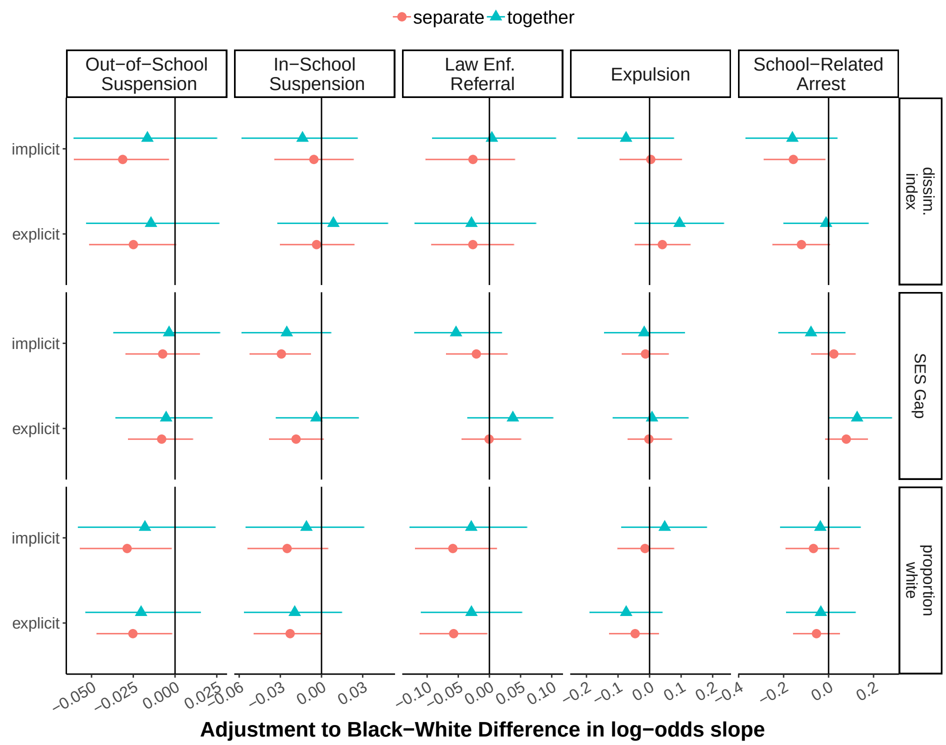

Fig. S5. Parameter estimates for moderators of the associations between bias and disciplinary actions in the 2015-2016 data. Point is the mean of the posterior. Error bars indicate 95% highest density intervals. Negative values indicate that the association between bias and disciplinary disparities becomes stronger as one moves into counties with higher levels of the indicated moderator

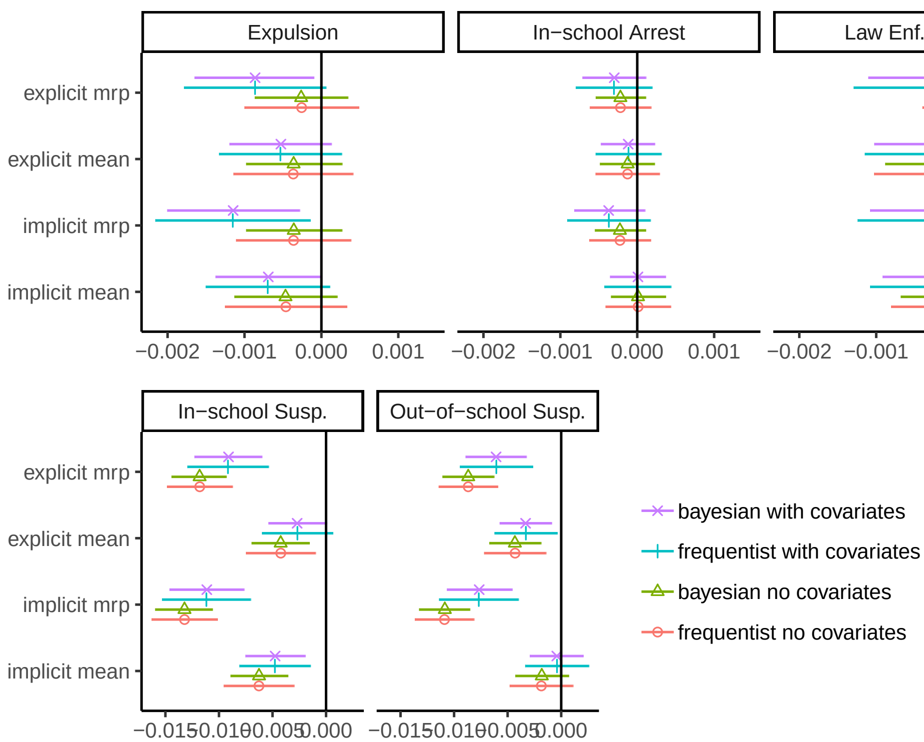

Fig. S6. Parameter estimates for a set of analyses in which data was first aggregated to the county level before fitting linear regressions. Point estimates correspond to either mean of the posterior distribution (bayesian models) or the maximum likelihood parameter estimate (frequentist models) for the association between bias and the black-white difference in discipline proportions. Error bars correspond to either 95% highest density intervals (bayesian models) or 95% confidence intervals (frequentist models).

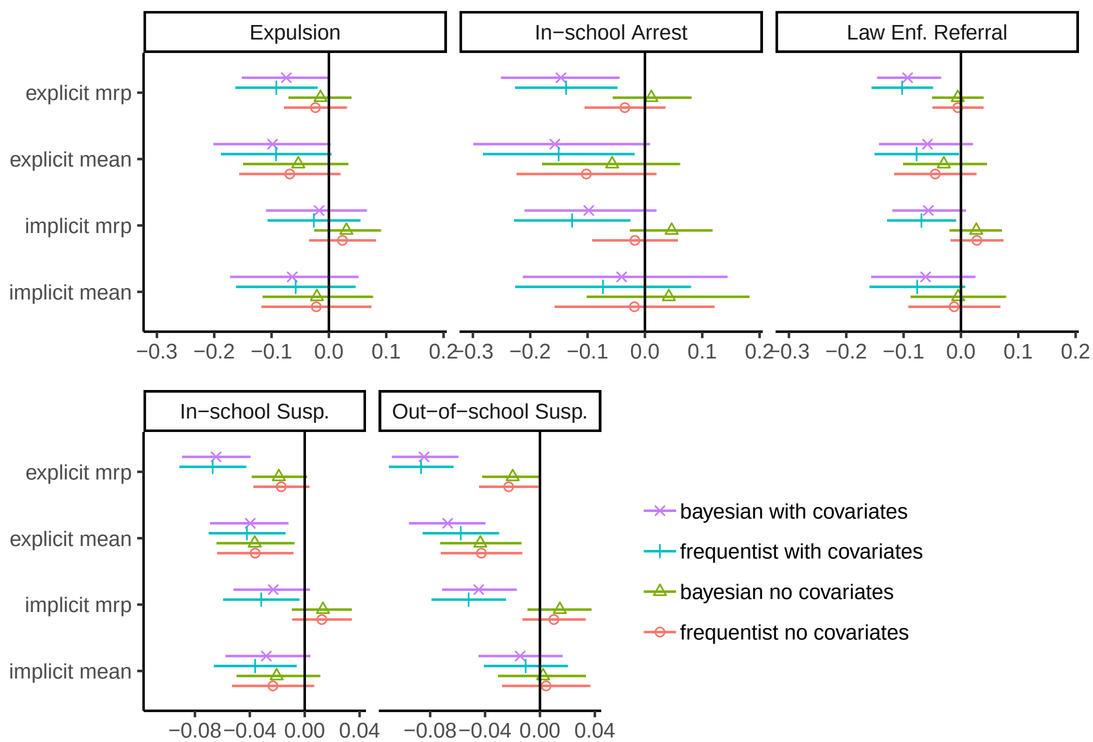

Fig. S7. Parameter estimates for a set of analyses in which a variety of multilevel models were fit to the raw data. Point estimates correspond to either mean of the posterior distribution (bayesian models) or the maximum likelihood parameter estimate (frequentist models) for the parameter that reflects the change in the log-odds slope for white students in comparison to black students. Error bars correspond to either 95% highest density intervals (bayesian models) or 95% confidence intervals (frequentist models).

Table S2. Correlation matrix for all 2014 county-level predictor variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1) Implicit Race																								
2) Explicit Race	0.79																							
3) Implicit Sex	0.51	0.58																						
4) Explicit Sex	0.31	0.44	0.76																					
5) Population Size	0.02	-0.04	-0.23	-0.24																				
6) White Population	0.01	-0.05	-0.25	-0.26	0.98																			
7) Black Population	0.11	0.06	-0.13	-0.16	0.78	0.71																		
8) Crime	0.12	0.05	-0.08	-0.08	0.13	0.11	0.15																	
9) Housing Density	0	-0.06	-0.18	-0.16	0.3	0.25	0.39	0.1																
10) Mobility	-0.25	-0.19	-0.1	-0.06	-0.1	-0.1	-0.11	-0.09	-0.05															
11) College Grads	-0.11	-0.09	-0.33	-0.29	0.22	0.24	0.19	-0.07	0.23	-0.03														
12) Unemployment	0.28	0.13	0.07	-0.02	0.13	0.12	0.16	0.31	0.06	-0.14	-0.18													
13) Median Income	-0.21	-0.18	-0.38	-0.36	0.23	0.27	0.12	-0.1	0.13	0.03	0.41	-0.3												
14) Poverty Rate	0.34	0.26	0.33	0.29	-0.05	-0.08	0.02	0.26	0.01	-0.12	-0.28	0.52	-0.73											
15) Proportion White	-0.36	-0.32	-0.13	-0.07	-0.18	-0.13	-0.3	-0.38	-0.15	0.21	0.08	-0.49	0.16	-0.45										
16) Proportion Black	0.58	0.52	0.32	0.2	0.08	0.04	0.26	0.28	0.09	-0.24	-0.08	0.37	-0.23	0.35	-0.79									
17) Bl-Wh Ratio	0.42	0.38	0.27	0.17	0.04	0	0.21	0.24	0.07	-0.17	-0.07	0.34	-0.23	0.34	-0.72	0.89								
18) Segregation	-0.04	-0.07	-0.13	-0.15	0.18	0.2	0.17	0.06	0.1	0.09	0.05	0.11	0.11	-0.1	0.1	-0.1	-0.09							
19) Wh-Bl Unemp. Diff	0	-0.02	-0.01	0	0.01	0.02	0	0.01	0.01	-0.02	0.02	-0.07	0.06	-0.02	0	-0.03	-0.04	-0.04						
20) Wh-Bl Income Diff.	0.11	0.16	-0.05	-0.05	0.1	0.11	0.12	0.04	0.11	-0.1	0.21	0	0.14	-0.03	-0.19	0.2	0.14	0.03	-0.15					
21) Wh-Bl Poverty Diff	-0.03	-0.08	-0.01	0	0.06	0.06	0.04	0.01	0.03	-0.17	0.02	0.01	0.13	-0.04	-0.04	-0.03	-0.04	-0.04	0.23	-0.39				
22) Wh-Bl College Diff	0.05	0.03	-0.09	-0.08	0.08	0.08	0.12	0.08	0.11	0.04	0.12	0.02	0.07	0.01	-0.13	0.11	0.08	0.05	-0.08	0.35	-0.15			
23) Overall SES	0.31	0.22	0.37	0.32	-0.13	-0.16	-0.04	0.24	-0.09	-0.09	-0.57	0.65	-0.85	0.89	-0.4	0.35	0.33	-0.06	-0.06	-0.12	-0.07	-0.05		
24) SES Gap	0.09	0.12	-0.04	-0.06	0.05	0.05	0.07	0.04	0.07	0.08	0.12	0.01	0.03	-0.01	-0.09	0.15	0.12	0.07	-0.48	0.83	-0.7	0.63	-0.04	
Mean	0.4	0.83	0.38	1.65	98.68	72.79	12.38	1.1	113.4	14.89	6.07	5.02	45.86	13.15	0.83	0.09	0.17	0.38	-8.3	15.97	-18.52	0.08	0	0
SD	0.02	0.17	0.05	0.49	316.23	198.52	54.12	0.94	823.71	17.69	4.83	2.09	12.73	7.66	0.17	0.14	0.42	0.19	17.2	16.26	21.72	0.15	1.5	1.28

[1] Crime rate is scaled to rate per 1000 people
[2] Total population, black population, and white population are in 1000's of people
[3] Income is in 1000's of dollars

Table S3. Correlation matrix for all 2016 county-level predictor variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1) Implicit Race																								
2) Explicit Race	0.79																							
3) Implicit Sex	0.53	0.6																						
4) Explicit Sex	0.32	0.46	0.76																					
5) Population Size	0	-0.05	-0.23	-0.24																				
6) White Population	0	-0.06	-0.25	-0.26	0.98																			
7) Black Population	0.09	0.05	-0.14	-0.16	0.78	0.71																		
8) Crime	0.06	0.01	-0.09	-0.08	0.11	0.09	0.13																	
9) Housing Density	-0.02	-0.07	-0.18	-0.16	0.3	0.25	0.39	0.07																
10) Mobility	-0.22	-0.16	-0.08	-0.06	-0.1	-0.1	-0.12	-0.07	-0.05															
11) College Grads	-0.23	-0.22	-0.52	-0.45	0.32	0.36	0.25	-0.05	0.23	0.02														
12) Unemployment	0.27	0.14	0.08	-0.02	0.12	0.11	0.16	0.29	0.06	-0.15	-0.15													
13) Median Income	-0.2	-0.18	-0.38	-0.35	0.24	0.28	0.12	-0.1	0.13	0.05	0.69	-0.27												
14) Poverty Rate	0.32	0.25	0.31	0.27	-0.04	-0.08	0.05	0.23	0.01	-0.13	-0.48	0.55	-0.71											
15) Proportion White	-0.33	-0.3	-0.11	-0.06	-0.19	-0.14	-0.3	-0.35	-0.15	0.21	0.01	-0.52	0.15	-0.52										
16) Proportion Black	0.55	0.5	0.3	0.19	0.08	0.04	0.26	0.24	0.09	-0.25	-0.09	0.41	-0.27	0.49	-0.8									
17) Bl-Wh Ratio	0.39	0.36	0.25	0.17	0.04	0	0.21	0.2	0.07	-0.18	-0.09	0.37	-0.26	0.46	-0.71	0.88								
18) Segregation	-0.03	-0.06	-0.11	-0.16	0.18	0.2	0.17	0.05	0.1	0.09	0.09	0.14	0.07	-0.01	0.09	-0.1	-0.09							
19) Wh-Bl Unemp. Diff	-0.03	-0.04	-0.03	-0.01	0.01	0.01	-0.01	0.02	0	-0.05	0.05	-0.12	0.08	-0.07	0.03	-0.04	-0.05	-0.01						
20) Wh-Bl Income Diff.	0.11	0.17	-0.08	-0.06	0.12	0.12	0.14	0.06	0.13	-0.16	0.22	0.01	0.17	-0.05	-0.21	0.2	0.14	0.04	-0.17					
21) Wh-Bl Poverty Diff	-0.04	-0.09	-0.02	0	0.05	0.05	0.03	0	0.02	-0.14	0.06	0.01	0.11	-0.04	-0.01	-0.04	-0.05	-0.01	0.23	-0.4				
22) Wh-Bl College Diff	0.04	0.02	-0.12	-0.1	0.08	0.08	0.11	0.06	0.1	0.01	0.2	0.01	0.08	-0.01	-0.13	0.11	0.07	0.01	-0.07	0.39	-0.13			
23) Overall SES	-0.32	-0.25	-0.43	-0.36	0.18	0.21	0.06	-0.2	0.1	0.1	0.76	-0.57	0.89	-0.89	0.36	-0.4	-0.37	0.02	0.09	0.15	0.07	0.09		
24) SES Gap	0.08	0.12	-0.05	-0.06	0.05	0.05	0.08	0.03	0.07	0.06	0.12	0.02	0.04	-0.01	-0.13	0.14	0.11	-0.01	-0.47	0.87	-0.75	0.57	0.05	
Mean	0.39	0.78	0.37	1.53	101.39	74.37	12.81	1.08	113.47	14.93	20.8	4.03	47.97	11.95	0.83	0.09	0.17	0.38	-6.75	16.83	-16.79	0.08	0	0
SD	0.02	0.17	0.05	0.48	325.2	202.67	55.37	0.94	823.96	17.26	9.14	1.69	12.61	5.76	0.17	0.15	0.44	0.19	15.11	15.88	21.15	0.16	1.57	1.34

[1] Crime rate is scaled to rate per 1000 people
[2] Total population, black population, and white population are in 1000's of people
[3] Income is in 1000's of dollars

References

1. Park DK, Gelman A, Bafumi J (2004) Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis* 12(4):375–385.
2. Gelman A, Little TC (1997) Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* 23(2):127–35.
3. Leitner JB, Hehman E, Ayduk O, Mendoza-Denton R (2016) Blacks' death rate due to circulatory diseases is positively related to whites' explicit racial bias: A nationwide investigation using project implicit. *Psychological science* 27(10):1299–1311.
4. Gonsalkorale K, Sherman JW, Klauer KC (2009) Aging and prejudice: Diminished regulation of automatic race bias among older adults. *Journal of Experimental Social Psychology* 45(2):410–414.
5. Scott SL, et al. (2016) Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management* 11(2):78–88.
6. Massey DS, Denton NA (1988) The dimensions of residential segregation. *Social forces* 67(2):281–315.
7. Hehman E, Flake JK, Calanchini J (2017) Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science* p. 1948550617711229.