Supplementary Materials for
Race, Voice, and Authority in Discussion Groups

October 3, 2024

Contents

S1 Race and Corporate Harm

Jury studies often focus on cases involving violent crime, but race is also relevant in deliberation about civil matters, especially those related to corporate behavior. Americans from different racial backgrounds own businesses and accumulate wealth at vastly different rates (Herring and Henderson 2016). In addition, patterns of residential segregation and zoning disproportionately expose people of color to environmental hazards from factories (Bolin, Grineski, and Collins 2005, p. 157). These experiences shape their views about corporations and their responsibilities to the public.

It is not surprising, then, that people of color are more supportive of regulating sources of risk, even when accounting for political attitudes and demographics (Kahan et al. 2007). In fact, racial gaps in views of corporate harm are even larger than gaps based on income (Unnever, Benson, and Cullen 2008). Because opinions about corporate power diverge so noticeably by race, and because civil lawsuits are a key mechanism for corporate accountability, the ability of jurors of color to exercise voice and wield influence within the jury is an important element of democratic equality. Gifford and Jones (Gifford and Jones 2016) quote an attorney stating that "Expressing their perceptions of justice through awards in serious personal injury cases is one of the few opportunities that most people of color have to ... send a message to corporations and other powerful social players" (p. 589). In civil contexts, the discursive dynamics of the jury are thus a key factor in whether the distinct perspectives and life experiences of jurors of color are incorporated into the jury's verdict. These decision-making groups are therefore a highly relevant setting in which to examine racial gaps in participation.

S2 Participant and Group Characteristics

The tables in this section describe the demographic characteristics of the sample (Table S1) and the demographic composition of the assembled groups (Table S2).

Table S1: Key Variable Summary Statistics (Individual-Level, All Groups)

Statistic	N	Mean	St. Dev.	Min	Max
Race: White	2,442	0.85	0.35	0	1
Female	2,439	0.60	0.49	0	1
Younger (18-29)	2,439	0.36	0.48	0	1
Middle Age (30-59)	2,439	0.43	0.50	0	1
Older (60+)	2,439	0.21	0.40	0	1
High School	2,442	0.23	0.42	0	1
Some College	2,442	0.40	0.49	0	1
College Graduate	2,442	0.38	0.48	0	1
Low Income (¡ 30K)	2,402	0.31	0.46	0	1
Middle Income (30-50K)	2,402	0.34	0.47	0	1
High Income (¿ 50K)	2,402	0.35	0.48	0	1
Pre-Deliberation Punishment Preference	2,421	3.51	2.57	0	8

The sample is fairly similar to the composition of the United States in 2020 on race (85% vs. 75% white non-Hispanic), age (20% vs. 21% older than 60), and education (38% college graduates vs. 35%), though the sample is more female. Compared to the demographics of the study site, Maricopa County, in the year

Table S2: Key Variable Summary Statistics (Group-Level, All Groups)

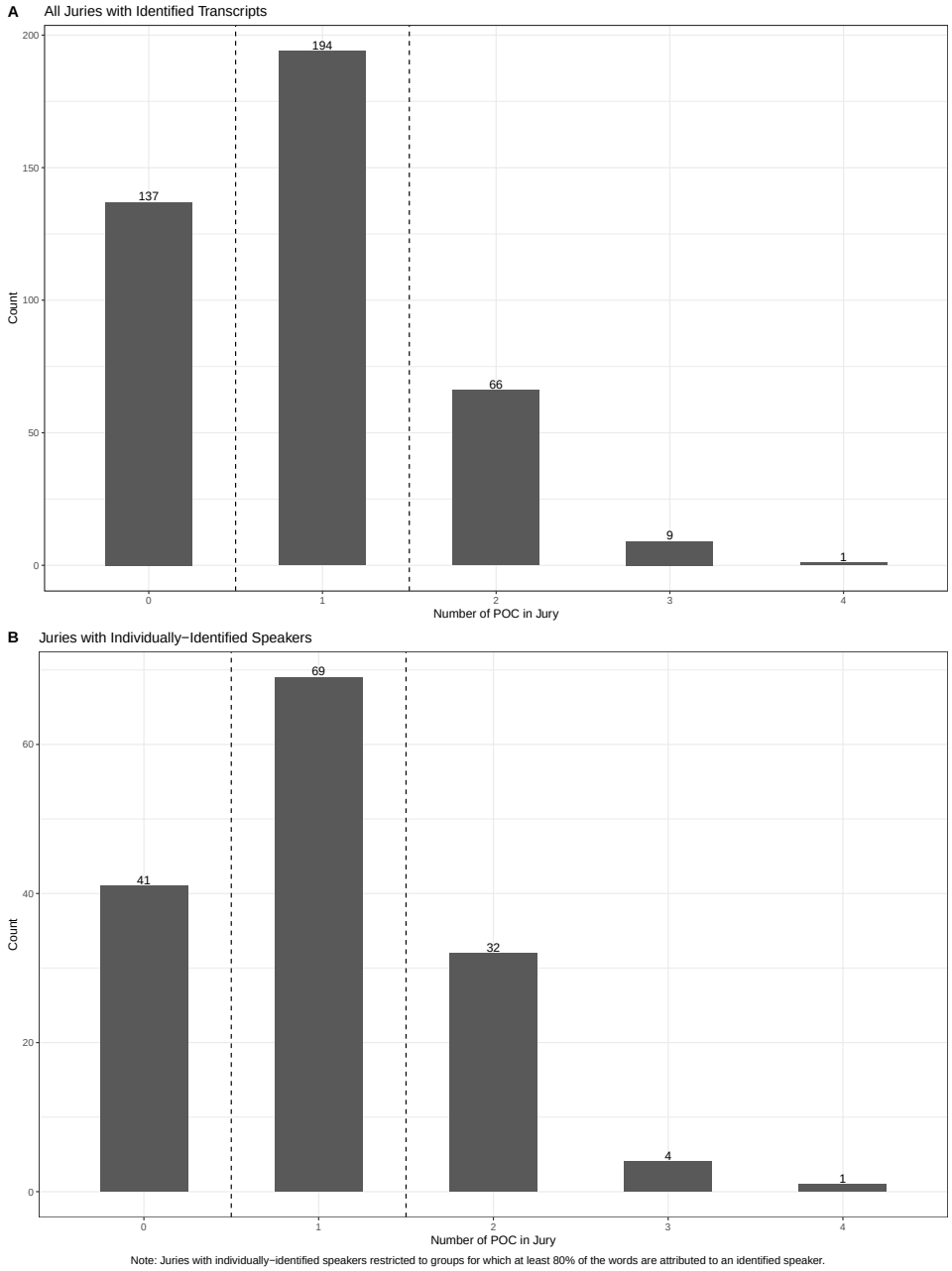
Statistic	N	Mean	St. Dev.	Min	Max
Number of Women	404	3.58	0.75	1	6
Number of Whites	407	5.12	0.77	2	6
Number Younger	404	2.17	1.08	0	5
Number Mid-Age	404	2.59	1.18	0	5
Number Older	404	1.24	0.97	0	5
Number HS Grad or Less	407	1.35	0.99	0	5
Number Some Coll.	407	2.39	1.11	0	5
Number College Grads	407	2.26	1.16	0	5
Number Low Income	368	1.90	1.07	0	5
Number Mid-Income	368	2.01	1.16	0	5
Number High Income	368	2.08	1.19	0	5
Median Pre-Deliberation Prefs (Round 1)	407	3.45	1.98	0.00	8.00
SD Pre-Delib Prefs (Round 1)	407	2.10	0.73	0.00	3.85

2000[1], the sample is more white (85% white non-Hispanic vs. 66%) more female (59% vs. 50%), and more educated (22% with a high school diploma or less vs. 41%). These differences could be attributed to the fact that the sample was drawn from registered voters, not the population as a whole.

[1]These figures were calculated from the 2000 Decennial Census, drawn from https://www.azcensus.com/maricopa-county/.

Figure S1 reports the group racial composition in the set of all juries for which we have linked transcripts (Panel A) and the set of juries with individually-identified speakers that are linked to the quantitative data (Panel B).

Figure S1: Jury Racial Composition



Note: Juries with individually−identified speakers restricted to groups for which at least 80% of the words are attributed to an identified speaker.

S3 Linkage of Transcripts and Data

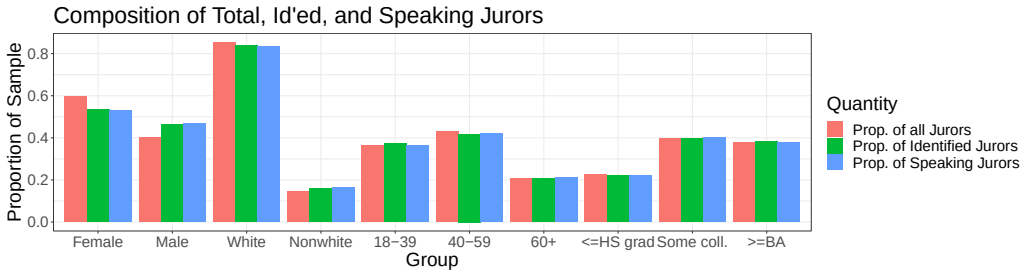
The analyses in the main text draw on two primary versions of the data: the first is the full set of 2,442 people in the 407 groups with available transcripts, and the second is the subset of 956 people who are individually linked to some speech in the transcripts. Our analysis of the individual data is restricted to the 767 participants in the 147 groups for which at least 80% of the words are attributed to an identified speaker. We thus linked transcripts to individual participants, but we did not succeed in linking all the participants. We describe the basic process of linking the transcripts to the quantitative data in the main text. This section offers further detail on the linkage and discusses differences between the two versions of the data. We conducted extensive checks and validations to ensure that the resulting missing transcript data did not bias the results.

The linking process was facilitated by the fact that at the beginning of their deliberation, jurors typically went around the table and stated their individual pre-deliberation preferences for punitive damages. Jurors sometimes mentioned their names or were called on by name as well. For this linkage, both rounds of deliberation were informative.

Because participants needed to have a unique preference within their gender in their group to be identified, males (who are comparatively less common in this data) and people with unique preferences are overrepresented in the identified sample. We do not, however, see evidence that the identified people differed from the full sample in other dimensions: those whose speech was identified do not differ from the full sample in their race, age, or education. In other words, we are reassured that missingness as a result of the linking process is not a threat to causal inference.

Figure S2 shows the demographic composition of three versions of the data: the full set of 2,442 participants in groups with linked transcripts, the participants who were identifiable in their group's transcript, and the participants who were linked to some speech in their group's transcript. These participants are broken down by their gender, race, age, and education level.

Figure S2: Juror Attributes



Importantly, there are no clear differences in the composition of identified and unidentified participants based on race. There are also no clear differences by age or education. However, males are noticeably more likely to be linked to their speech in the transcripts than are females.

What are the causes of this gender gap? As discussed above, participants were linked to text in their group's transcript by matching their recorded gender and punishment preferences to the gender and stated punishment preferences of speakers in the transcript. This means that linking to text was more often possible for people whose preferences were unique within their gender in their group. Combined with the fact that there were typically fewer men in each group than women, since men were only 40% of the sample, this means that men were easier to link to speech in the transcript. Coders also reported having greater difficulty

individually identifying the speech of women than of men.

The key for purposes of this study, however, is that there is no difference in our ability to link white and POC participants to their speech, assuaging possible concerns about the accuracy of the racial participation gaps we found.

S3.1 Intercoder Reliability

We conducted multiple tests of intercoder reliability, addressing two types of questions. First, do research assistants assign the same lines of text to the same speakers? In other words, when an RA assigns a set of speech turns to a single speaker, how often does another RA assign those same speech turns to the same speaker? We tested this using four training transcripts coded by all six RAs who worked on the project; that is, each RA assigned each line of speech in the transcript to a speaker. We first compiled the full set of speech lines — the "speech profile" — each RA attributed to a single speaker. For each transcript, then, there are 6 speech profiles (one for each juror) coded by each RA. We then compared every speech profile to all the speech profiles in the jury coded by other RAs. This allows us to identify the most similar pairs for each speech profile — that is, for any given speech profile assembled by one coder, we can find the most similar speech profiles assembled by each other coder. This represents the set of profiles from each other coder that are most likely to be referring to the same speaker. Once each profile was paired with the profile from each other coder that is most likely to be referring to the same speaker, we calculate the median similarity between the speech profile and each of its closest neighbors. This represents how similar the speech profile is to the profile a typical other coder attributed to the same speaker. The distance metric used to capture the similarity between two speech profiles is based on the restricted Damerau-Levenshtein distance, which counts the number of text insertions, deletions, substitutions, or transpositions that would be necessary to make the speech profiles identical. Such distance functions are commonly used "to quantify the similarity between two strings" or texts (von der Loo 2014, 111). This count is then rescaled to make these numbers comparable across speech profiles of different length, so that 0 indicates complete dissimilarity and 1 indicates complete similarity. The median and mean similarity of two speech profiles are both .67, suggesting that for a typical speech profile from one RA, about 33% of the characters would need to be changed to perfectly match the other RAs' speech profiles for the same speaker. When we incorporate length weights, to reflect the fact that some speech profiles include many thousands of characters while others include a hundred or fewer, the mean is .73.[2]

This measure of intercoder agreement likely represents a lower bound. These tests were conducted early in the data cleaning process, and RAs became more confident and consistent in coding as they gained experience.

Our intercoder reliability tests also addressed a second question: Do RAs link speakers to the same observations in the data? That is, when an RA identifies a single speaker in the transcript and links them to a row in the juror-level dataset, how often does another RA link the same speaker to the same row? For this check, we assigned pairs of research assistants to examine a total of 45 randomly chosen jury transcripts — a little over 10 percent of all available transcripts. Of the 270 jurors in these transcripts, 83 percent were linked to the same rows in the data by both RAs who analyzed the transcript. On average, Krippendorff's alpha for the 45 juries is 0.80, which reaches the standard threshold for very high levels of intercoder agreement (Hughes 2024). The median Krippendorff's alpha across the 45 test transcripts is 1.0 — perfect agreement.

[2]Omitting speech turns with 3 or fewer words raises the unadjusted mean to .69 and the weighted mean to .74.

S4 Preference Mention Extraction Code

To identify mentions of specific punishment preferences in the deliberation transcripts, we wrote a program that processes the raw text from the transcripts and outputs numeric representations of dollar amounts and scale points preferred by participants. Its core function takes phrases like "one hundred thousand" and transforms them into numbers. The program uses the following steps:

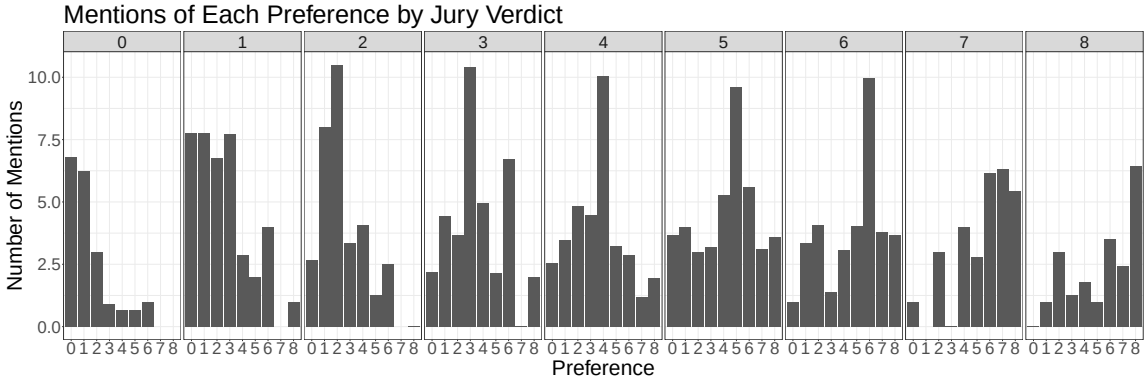
1. Replace verbal statements of scale points with numbers; e.g. "extremely severe" is replaced with an 8
2. Replace common but non-pattern-conforming verbal versions of numbers with pattern-conforming ones (e.g. "one half" with .5; "a hundred" with 100; "grand" with thousand)
3. Exclude phrases in which numbers appear but are not expressions of preferences (e.g. anything with pattern (number):(number)(number), which is a time; phrases like "(number) of us" and "(number) people")
4. Expand phrases like "10 or 20 million" or "1 to 2 thousand" to apply the magnitude to both numbers in the range (i.e. "10 million or 20 million")
5. Identify remaining multiword numbers by grouping together words that express numbers without other words in between
6. Run the core function to transform multiword numbers to numeric:
   (a) Identify and replace basic number words (e.g. one, ten, thirty to 1, 10, 30)
   (b) When applicable, identify words expressing magnitude, like hundred, thousand, and million
   (c) When applicable, multiply relevant numbers by their orders of magnitude (e.g. in "nine thousand", multiply the number 9 by the magnitude 1000)
   (d) Add the remaining numbers together for a final result (e.g. for "one thousand and seven", add 1000 and 7 to get 1007)
7. In scale deliberations, remove numbers outside the range of 0-8
8. In dollar deliberations, remove the numbers 200,000, 100 million, and 200 million
9. In dollar deliberations, locate and transform preference mentions that are likely to be of a different magnitude than the one detected[3]: if a preference mention multiplied by one thousand or one million is mentioned elsewhere in the deliberation, assume the preference is that higher multiple
10. Remove preference mentions that are not recorded as initial preferences by at least one person in the group

To validate this measure, we checked whether the number of times a group mentioned a potential decision correlates with the likelihood of the group settling on that decision. Figure S3 divides groups by their final decision, then shows how often groups with each decision mentioned each scale point. The figure is restricted to deliberations on the ratings scale.

[3]In our pilot testing of this code, we found that people often abbreviate their preferences by dropping an order of magnitude. For example, if people are debating between 100,000 and 200,000 as punishment amounts, they will shift to saying simply "I think 100 is better" or "I could go with 200." This section of the code is intended to address this practice.

In almost every case, the eventual decision is the scale point mentioned most by the group, and it is always discussed often. (Note, however, that the measure only captures scale points that were at least one person's initial preference. For example, a group might have had 3 members preferring 6 and 3 preferring 8; mentions of 7 will not be captured, even if that were the eventual decision, if 7 was no person's starting preference.)

Figure S3: Preference Mentions and Group Verdicts



S5 Descriptive Statistics

Table S3 presents descriptive statistics for key variables used in models in the main text.

Table S3: Descriptive Statistics

Measure	Min	1st Quart.	Mean	Median	3rd Quart.	Max
Speech length (words)	0	121.0	452.8	313	618.0	2842.0
Total Group Words (attr.)	42	1180.0	2523.8	2241	3555.0	7325.0
Indiv. Mentions of Any Pref	0	1.0	4.9	3	6.5	66.0
First Turn: Distance from Beginning	1	2.0	10.2	6	12.0	279.0
Last Turn: Distance from End	1	4.0	17.8	9	20.0	342.0
Total Length in Turns	16	105.0	207.7	178	270.0	813.0
Own Mentions of Pref	0	1.0	2.2	1	3.0	24.0
Others' Mentions of Pref	0	0.0	4.5	2	7.0	40.0
Total Others' Mentions	0	10.0	22.4	19	31.0	114.0
Foreperson Mentions of Pref	0	0.0	2.7	2	4.0	26.0
Forperson Total Mentions	0	5.0	10.3	7	14.0	66.0
Absolute Distance from Group Median	0	0.5	1.6	1	2.5	7.5
N Sharing Pref	0	0.0	0.5	0	1.0	4.0
Group Mentions of Indiv Pref	0	2.0	7.2	5	10.0	47.0
Total Mentions, Any Pref (attr.)	0	5.0	21.1	17	32.0	118.0

S6 Interaction Models with Full Controls

Table S4 contains models of the outcomes in Table 1 through Table 5, in order: speech length, the total number of preferences a person mentions, the timing of participants' first and last turns, and the number of times the person's preference is mentioned by their group in deliberation. The models below test for an interaction between individual race and group racial composition, but unlike those in the main text, they include the full set of preference and endogenous controls. Note that the "foreperson" designation is coded from the linked data, so it is unavailable for the total mentions analyses.

As in the main text, none of the interaction terms are significant; the effect of individual race does not seem to vary based on the group's racial composition.

Table S4: Interactions with Controls

	Speech Length	Prefs Mentioned	First Turn	Last Turn	Total Mentions of Pref
Indiv. Race: White	0.556*	0.121	0.090	0.285*	0.083
	(0.250)	(0.073)	(0.128)	(0.125)	(0.124)
5 Whites	0.247	−0.036	0.116	0.126	−0.167
	(0.261)	(0.105)	(0.146)	(0.162)	(0.121)
6 Whites	−0.090	0.025	0.106	−0.069	−0.005
	(0.191)	(0.087)	(0.099)	(0.116)	(0.081)
White Indiv. x 5 Whites	−0.138	−0.027	0.033	−0.144	0.113
	(0.297)	(0.101)	(0.154)	(0.173)	(0.144)
Preference	0.042	0.029**	0.001	0.003	0.061***
	(0.022)	(0.010)	(0.016)	(0.015)	(0.012)
Pref. Distance	0.112**	−0.036*	0.008	0.040	−0.037
	(0.034)	(0.017)	(0.025)	(0.025)	(0.022)
Others Sharing Pref.	−0.194*	−0.083*	0.018	−0.085	0.284***
	(0.089)	(0.041)	(0.041)	(0.054)	(0.040)
Foreperson	1.370***	0.461***	1.275***	1.230***	
	(0.130)	(0.064)	(0.068)	(0.070)	
Total Group Words	0.536***				
	(0.099)				
Total Indiv. Words		0.295***			
		(0.018)			
Total Turns			−0.123*	−0.119*	
			(0.050)	(0.057)	
Total Group Pref. Mentions					0.218***
					(0.049)
Num.Obs.	743	743	729	729	1758
SE Clusters	Group	Group	Group	Group	Group
N Clusters	143	143	143	143	311
R2	0.325	0.521	0.285	0.266	0.163
R2 Adj.	0.303	0.505	0.260	0.241	0.152

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

S7 Race and Preference Sharing

In predicting full groups', other people's, and foreperson's mentions of a person's preference, the number of other people who share that preference is statistically and substantively significant. When a person's preference is unique in their group, it is mentioned a median of 3 times in a round of deliberation; when one other person in the group has the same preference, the median increases to 6. This is an almost "mechanical" effect — a simple result of twice as many people voicing the preference — that, at first blush, is unrelated to race.

However, Table S5 shows that race is in fact implicated: the first model regresses the number of people in a group sharing the focal person's preference on individual race, deliberation scale, and scenario fixed effects, and it suggests white people on average have more peers in their group sharing their preference. POC participants, though, have different (on average, more punitive) preferences than white participants, and these different preferences could explain these people's different likelihoods of having allies in their group. To test whether this gap can be explained by white and POC participants holding different preferences, the second model in Table S5 adds an indicator variable for the person having a preference greater than 0 (that is, 0 on the rating scale or exactly \$0 on the dollar scale).[4] Indeed, with the controls for individual preference, the coefficient on race disappears. This pattern suggests that because they tend to hold different preferences than white participants, POC are less likely to have "allies" in their group that exactly share their preference. Race can thus matter in multiple ways: individual race directly affects the expression of many aspects of voice. But it also operates "indirectly," by shaping the distribution of preferences present within the group as deliberation begins — a distribution that itself affects many aspects of the subsequent deliberative exchange, with implications for the voice and authority of POC deliberators within their groups.

Table S5: Number of Others Sharing Preference

	Model 1	Model 2
Indiv. Race: White	0.151*	0.036
	(0.061)	(0.052)
Preference \$¿\$ 0		−1.209***
		(0.102)
Scale: Ratings	−0.055	0.086
	(0.070)	(0.059)
Scenario FEs	X	X
SE Clusters	Group	Group
N Clusters	386	386
R2 Adj.	0.209	0.355

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table S6 further explores the relationship between preference sharing and uptake. Its models predict mentions by the full group, the focal person, and others in their group, interacting individual race with the distance of their preference from the group median (col. 1) or with the number of others in their group who share their preference (col. 2).

[4] The baseline category is the 0 scale point.

Table S6: Models with Preference Distance Interactions

	Other Mens	Other Mens
Omitted: POC		
Indiv. Race: White	−0.307 (0.180)	−0.020 (0.101)
Distance: .5 to 1	−0.350 (0.245)	
Distance: 1.5 to 2.5	−0.220 (0.227)	
Distance: 3+	−0.624 (0.276)*	
N. Sharing Pref.	0.365 (0.063)***	
White x Distance .5 to 1	0.389 (0.241)	
White x Distance 1.5 to 2.5	0.232 (0.226)	
White x Distance 3+	0.579 (0.279)*	
N Sharing Pref = 1		0.335 (0.155)*
N Sharing Pref = 2		1.012 (0.307)**
White x 1 Sharing Pref.		0.247 (0.170)
White x 2+ Sharing Pref.		−0.260 (0.284)
SE Clusters	Group	Group
N Clusters	143	143
R2 Adj.	0.268	0.273

S8 Dissenters

In the main text, we show that POC members experience less uptake when their preferences are far from the group's pre-deliberation median. Consistent with our pre-analysis plan, we focus here on a subsample of deliberators: dissenters.

We define dissenters as people whose pre-deliberation preferences are at least 1.5 points from their group's mean pre-deliberation preference. About half of participants qualify as a dissenter in at least one round of deliberations. A majority of group decisions fall within 1.5 points of their pre-deliberation mean, so dissenters must exercise voice and influence to convince their fellow deliberators to move away from their initial preferences. About half of participants qualify as a dissenter in at least one round of deliberations..

The median dissenter speaks about twice as much as the median non-dissenter, and they tend to first speak earlier. Thus, dissenters appear to exercise voice. However, they experience less uptake, with their preferences mentioned less in deliberation.

One stringent test of racial differences in voice and uptake is to compare "lone" POC dissenters surrounded by white peers to white dissenters in all-white groups, accounting for preferences and allies.[5] We used coarsened exact matching to construct weights so that the distributions of POC and white dissenters are similar in their preferences relative to their group's mean and in the number of people who share their preference in their group.

Unfortunately, when we subset the sample in this way, this most direct test is underpowered. Calculations of the minimum detectable effect (MDE) suggest that for each outcome, the gap among dissenters would need to be at least 30% larger than the effect for the full sample in order for us to have detected an effect.

Table S7: Matched Dissenter Analyses

	Speech Length	Total Prefs	First Turn	Last Turn	Total Mens	Own Mens	Other Mens	Fore Mens
Indiv. Race: Nonwhite	−0.294	−0.248	0.065	−0.164	−0.190	−0.124	−0.065	0.025
	(0.290)	(0.170)	(0.193)	(0.205)	(0.125)	(0.168)	(0.179)	(0.198)
Scale: Rating	0.521*	0.589***	0.015	−0.203	0.240*	0.043	0.426*	−0.077
	(0.251)	(0.161)	(0.186)	(0.196)	(0.117)	(0.158)	(0.164)	(0.186)
Num.Obs.	163	163	157	157	375	156	163	137
R2 Adj.	0.117	0.121	−0.028	0.035	0.039	−0.002	0.123	0.079
SE Clusters	Group	Group	Group	Group	Group	Group	Group	Group
N Clusters	91	91	90	90	170	88	91	72

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To remain consistent with our pre-registration plan, we present the results of this under-powered analysis. Table S7 presents the results of regressing each measure of voice and uptake on individual race, scale, and case, for the matched dissenter sample. This is the same as the basic model in the main text. The results suggest that POC dissenters participate less and experience less uptake than white dissenters. However, the racial gaps are smaller among dissenters than among all group members. The lower power available due to the smaller sample size increases the uncertainty in our estimates, and the coefficient on race is insignificant for all the analyses relying on the smaller linked-individual sample. Only the racial gap on total mentions, where our sample size is much larger, exceeds standard levels of statistical significance. Given the limited statistical power in most models, we cannot conclude that the gap does or does not persist with this exacting test.

[5] While white and POC participants have similar preferences on dollar punishments, POC dissenters are 1.5 times more likely than white dissenters to be more punitive than their group in ratings deliberations.

S9 Additional Measures of Voice and Uptake

Another possible measure of influence concerns whether a group spends its time discussing preferences far from the focal deliberator's. Table S8 shows the relationship between race and the number of times a preference more than three points (of 8) away from the participant's is mentioned.

In the base model, there is no significant race gap. When accounting for the number of people who share the focal participant's preference, and for the number whose preferences are more than 3 points away, a race gap emerges. The gap remains when adding controls for the group's pre-deliberation median preference and for the total number of mentions in the group. As in other outcomes, the racial gap is no smaller in more diverse groups.

Table S8: Mentions of Prefs. >3pts from Focal Member's

	Base Model	Preference Control	Endog. Controls	Interaction
Indiv. Race: White	−0.186*	−0.208**	−0.076	−0.126
	(0.077)	(0.069)	(0.048)	(0.116)
Preference		−0.079***	−0.065***	
		(0.013)	(0.011)	
Pref. Distance		0.056	0.057	
		(0.042)	(0.032)	
Others Sharing Pref		−0.106*	0.233***	
		(0.053)	(0.043)	
Others w. Far Prefs		0.224***	0.381***	
		(0.063)	(0.040)	
Total Group Mentions			0.767***	
			(0.033)	
Predelib. Median			−0.062*	
Omitted: Less than 5 whites				
			(0.024)	
5 Whites				−0.067
				(0.170)
6 Whites				−0.146
Omitted: White x Less than 5 whites				
				(0.166)
White x 5-White Group				−0.023
				(0.138)
Scenario and Scale FEs	X	X	X	X
SE Clusters	Group	Group	Group	Group
N Clusters	337	319	255	337
R2 Adj.	0.052	0.165	0.622	0.051

* p < 0.05, ** p < 0.01, *** p < 0.001

Finally, we measure how often a group member's preference is mentioned by the white members of their group. If white deliberators take up the preferences of their POC peers less than those of their white peers, this effect could be dampened in models looking at mentions by all group members. Table S9 shows that there is no race gap in the number of times a person's preference is mentioned by white people in their group.

Table S9: White Group Members' Mentions of Participant Preference

	Base Model	Preference Control	Endog. Control	Interaction
Indiv. Race: White	−0.013	−0.011	0.051	−0.199
	(0.085)	(0.081)	(0.074)	(0.127)
Preference		0.053**	0.062***	
		(0.016)	(0.014)	
Pref. Distance		−0.009	−0.007	
		(0.024)	(0.022)	
Others Sharing Pref.		0.398***	0.428***	
		(0.054)	(0.041)	
Foreperson			−0.001	
			(0.068)	
Omitted: Less than 5 whites				
5 Whites				0.126
				(0.168)
Omitted: White x Less than 5 whites				
6 Whites				0.374**
				(0.118)
Scenario and Scale FEs	X	X	X	X
SE Clusters	Group	Group	Group	Group
N Clusters	147	143	143	147
White x 5-White Group				0.147
				(0.175)
R2 Adj.	0.172	0.271	0.487	0.184

* p < 0.05, ** p < 0.01, *** p < 0.001

S10 Models with Demographic Controls

While the main text focuses on the effects of race on voice and uptake, other demographic characteristics of deliberators could be related to their participation. For example, educational attainment is correlated with race and with participation in group discussions. Might the race gaps in participation we observe be attributable to educational attainment? To test this, we add controls for educational attainment, income, age, and sex to models of five key outcomes from the main text.

The results suggest that education is highly predictive of participation (especially holding a college degree). Income and sex[6] are also implicated. However, a sizeable race gap in participation remains for each outcome after including these controls (though the race gap in total preference mentions falls just short of significance at p=.054). We conclude that while education is important in shaping participation, it does not account for the effect of race in our data.

A related possibility is that education moderates the effect of race; perhaps the race gap is more, or less, pronounced for participants with high levels of education. We do not find evidence of this: in models interacting individual race with education, there are no significant interaction effects between race and education on any of these key measures.

[6]Interpretation of these models is complicated by the fact that women were less likely to be linked to transcripts than men; any gap in participation could be due to gendered measurement error, not gendered behavior.

Table S10: Models with Demographic Controls

	Speech Length	Prefs Mentioned	First Turn	Last Turn	Mens. of Pref
Omitted: POC					
Indiv. Race: White	0.468**	0.234***	0.225*	0.298**	0.081
	(0.144)	(0.070)	(0.087)	(0.096)	(0.066)
Omitted: HS grad or less					
Some college	0.333*	0.238**	0.154	0.139	0.019
	(0.143)	(0.076)	(0.091)	(0.089)	(0.052)
BA or more	0.630***	0.382***	0.232*	0.380***	0.120*
	(0.135)	(0.079)	(0.100)	(0.092)	(0.057)
Omitted: Income textless 30k					
Income 30-50k	0.164	0.105	−0.028	−0.152	0.132*
	(0.136)	(0.072)	(0.082)	(0.089)	(0.056)
Income 50k+	0.352*	0.137	0.218*	0.061	0.176**
	(0.140)	(0.080)	(0.091)	(0.104)	(0.059)
Omitted: Age 18-39					
Age 40-59	0.101	0.075	−0.042	−0.085	0.021
	(0.110)	(0.063)	(0.078)	(0.083)	(0.049)
Age 60+	−0.269	−0.013	−0.124	−0.078	−0.021
	(0.159)	(0.070)	(0.087)	(0.090)	(0.054)
Omitted: Sex = F					
Sex: M	0.529***	0.235***	0.036	0.183*	0.066
	(0.098)	(0.058)	(0.076)	(0.074)	(0.039)
SE Clusters	Group	Group	Group	Group	Group
N Clusters	147	147	147	147	327
R2 Adj.	0.135	0.176	0.029	0.053	0.073

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

S11 Models of Hispanic Participants

Throughout the main text, we measure participants' racial identities using only two categories: white participants and people of color. Because of the size and diversity of the sample, there are not sufficient numbers to divide participants of color further into individual racial and ethnic groups. Only the largest group of POC, Hispanic participants, has more than 50 members in the linked sample. We replicate the basic models' key results in Table **??**, replacing the white (vs. POC) measure of race with an indicator for whether a participant identifies as Hispanic. Though the standard errors are larger, estimates are consistent with those using the POC indicator: like POC as a whole, Hispanic participants speak less, mention their preferences less, speak at less influential times, and have their preferences mentioned less overall.

Table S11: Models with Hispanic Indicator labelhisp

	Speech Length	Prefs Mentioned	First Turn	Last Turn	Mens. of Pref
Indiv. Race: Hispanic	−0.760***	−0.362***	−0.298*	−0.243	−0.152
	(0.215)	(0.106)	(0.130)	(0.140)	(0.093)
SE Clusters	Group	Group	Group	Group	Group
N Clusters	147	147	147	147	327
R2 Adj.	0.049	0.111	0.002	0.006	0.063

* p < 0.05, ** p < 0.01, *** p < 0.001

S12 Negative Binomial Models

This section reports a robustness check using negative binoamial models instead of count models. Count models in the main text account for skewness by log-transforming the dependent variable. The models in Table S12 show key findings when we instead employ non-transformed dependent variables and account for skewness with a negative binomial model. In these models, First Turn and Last Turn are coded such that lower values reflect an advantage for whites. All key results hold with this alternative specification.

Table S12: Negative Binomial Regressions

	Speech Length	Prefs. Mentioned	First Turn	Last Turn	Mens. of Pref.
Indiv. Race: White	0.383**	0.471***	−0.294*	−0.275*	0.120
	(0.120)	(0.108)	(0.150)	(0.113)	(0.082)
Scenario and Scale FEs	X	X	X	X	X
SE Clusters	Group	Group	Group	Group	Group
N Clusters	147	147	147	147	407
Log.Lik.	−5427.403	−2000.049	−2512.801	−2919.867	−5506.199
RMSE	442.16	5.44	16.24	27.76	6.97

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

S13 Word Frequency Analysis

We identified words distinctive to white and POC speakers with three measures. The first combines two dimensions of the frequency with which participants use words: the number of times a word is used, and the number of distinct people who use it. The second two measures use machine learning classification algorithms to identify words that are most predictive of a person's race.

To measure how frequently white and POC participants use specific words in deliberations, we use two dimensions: term frequency (TF) and proportion using (PU). TF is the total number of times a word is used by people in a racial group, divided by the total number of words used by people in that racial group. For example, the word "harm" is used by white participants 166 times, and white participants used a total of 467,841 words, leading to a TF of .00035. PU is the proportion of people in a group who use a word at least

once. For example, of the 154 POC participants, 9 use the word "harm" at least once, so the PU of "harm" for POC participants is .058.

To compare these measures between white and POC participants, we then convert the raw TF and PU values into percentiles within racial groups. This accounts for differences in the overall distribution of word frequency across people of different racial backgrounds, allowing us to focus on differences in usage of individual words. Finally, we identified the words with the largest difference in frequency of use between white and POC participants. We focused all words that fell at least 5 percentile points higher in their usage by POC (white) participants than for white (POC) participants for both the TF and PU measures.

We use two different classification models to identify the words which are most distinctive of white and POC speakers. Both treat this task as a "fictitious prediction problem" Grimmer, Roberts, and Stewart (2022), in which they identify the words which would be most useful in trying to predict a person's race based on their use of words. The first is a general method that fits a generalized linear model (GLM) using regularization to identify the strongest predictors Friedman, Hastie, and Tibshirani (2010). The second is Multinomial Inverse Regression (MNIR), developed in Taddy (2013) specifically for text analysis, which fits models regressing word use on individual race, also returning the most-predictive words. From each model, we take the 50 most-predictive words for each racial group.

We take several standard steps to process the raw text data before calculating these measures, including removing special characters, stemming words, converting all terms to lowercase, removing proper names, and removing terms used very commonly in spoken English (lightly adapted from Leech, Rayson, and Wilson (2001)). To ensure the lists represented systematic differences across the body of deliberations, we also required the words to be used by at least 10 separate people in each racial group and to appear multiple times across at least 3 different legal cases.

The words on the lists below appear among the most-distinctive for at least 2 of the 3 measures described above (percentile difference, GLM, and MNIR). Note that the percentile difference measure and each of the prediction model measures are correlated with one another at between .65 and .75.

Words used more often by white participants[7]**:** punit than zero percent point case juri talk between compensatori award part differ standard start bit high anoth after busi never believ possibl question felt read defend whole pretti basic keep next end everi live assum word middl call matter show less idea sit ever larg pick

Words used more often by POC participants: hundr work knew same wasnt done thousand shes exact warn car through wrong fault someon everyth base long shouldnt cover correct buy close rest came manufactur total public seem suffer sound ahead anyway day fair stay act chanc hold onc hire nobodi consum each sell hour yet serious continu order across die realiz proper receiv

S14 Topic Models

This section describes the process of running the topic models discussed in the main text in further detail. We followed the recommendations of (Roberts et al. 2014) and used the related R package `stm` for all computations.

To process the text for analysis, we removed stopwords, proper names, punctuation and symbols, then stemmed the remaining words and removed those used less than five times (which were typically misspellings or garbled words). We then repeated the process described below for both individual-level speech (i.e. each document was a single person's speech in one round of deliberation) and for group-level speech (i.e. each document was a group's full transcript for one round of deliberation).

[7]We list stemmed versions of words here; that is, for example, the word stem "larg" includes uses of the words large, larger, and largest. The models use word stems to calculate distinctiveness.

In each model, we allowed the prevalence of topics to vary by legal scenario, round number, scale (ratings or dollars), and individual race/group racial composition (for individual-/group-level models respectively). To determine the number of topics to identify, we repeated models using 20, 30, and 40 topics and examined the results for coherence. We settled on 30 topics; models of this size generally identified topics for most of the 15 scenarios separately, but did not identify many topics without a legible interpretation.

We then used the `selectModel` function in R's `stm` package to identify potentially high-performing models. This function generates an initial list of 50 potential models, selects the highest-performing 15% based on semantic coherence, exclusivity, and sparsity, then runs these models fully, identifying topics and assigning topic proportions to each text. We then extracted differences in the proportion usage of each topic between white and POC speakers, and by racial composition of the group. This approach allows us to compare the use of each topic by individual race (for individual-level analysis).

The group-level models uncovered no differences in topic usage by racial composition. Figure S4 shows the distribution in t-statistics for the difference in topic proportion for all 30 topics in each of the 10 identified models, testing for differences in usage between juries with Less than 5 whites and those with 5 or 6 whites. Only a few topics appear to be used more often by groups of one composition or another, and these appear to be no more different than would be expected by chance. The individual-level models also uncovered

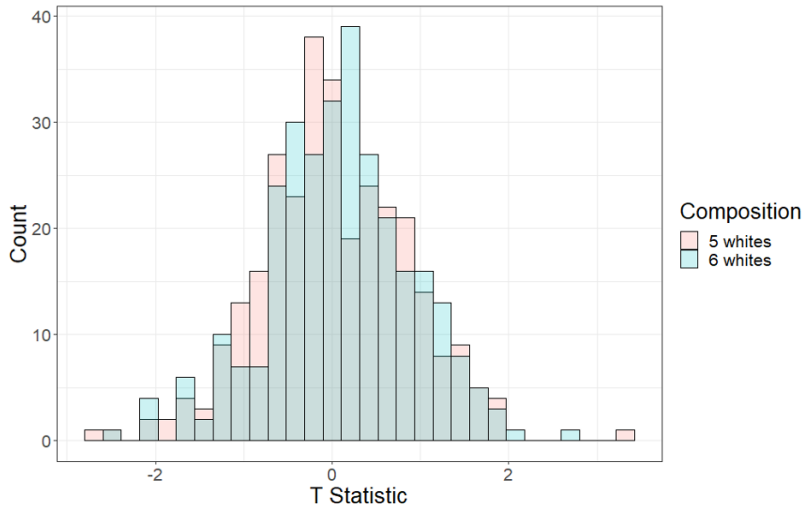


Figure S4: T-statistics for differences in topic usage by group racial composition (omitted category: Less than 5 whites).

few differences, though one did appear in 3 of the 10 models: a topic that appears to relate to foreperson speech was more common among white speakers than POC. This is sensible given the higher rate at which white participants were selected as forepersons. Among topics related to specific scenarios, a few were more common among white or POC speakers, but these topics were not systematically more commonly used by participants in either group—that is, as many scenario topics were more-used by white speakers as were more-used by POC speakers.

S15 Experimental Stimuli

This section provides examples of the experimental stimuli received by the study participants. Each jury was assigned to one case. Table S13 includes brief overviews of the 15 cases employed in the study and shows that the number of juries in the full sample assigned to each case is balanced.

Summary information about a sample case (*Glover v. General Assistance*) can be found in S14.1 (drawn from Appendix A in Schkade, Sunstein, and Kahneman (2000)).

The judge's instructions are provided in S14.2 (see Sunstein et al. (2002)'s appendix, pages 259-260). Instructions about dollars or severity rating decisions are presented in S14.3 (see Figure 3.1 of Sunstein et al. (2002)).

Table S13: Cases Considered by Juries

Number	Case	Description	# Juries
1	*Williams v. National Motors*	Motorcycle driver injured when brakes fail	35
2	*Smith v. Public Entertainment*	Circus patron shot in arm by drunk security guard	34
3	*Douglas v. Coastal Industries*	Auto air bag opens unexpectedly, injuring driver	33
4	*Sanders v. A&G Cosmetics*	Man suffers skin damage from using baldness cure	34
5	*Stanley v. Gersten Productions*	Elderly woman suffers back injuries from using exercise video	32
6	*Glover v. General Assistance*	Child ingests large quantity of allergy medicine, needs hospital stay	33
7	*Lawson v. TGI International*	Employee suffers anemia due to benzene exposure on the job	34
8	*Newton v. Novel Clothing*	Small child playing with matches burned when pajamas catch on fire	34
9	*West v. MedTech*	Disabled man injured when wheelchair lift malfunctions	32
10	*Windsor v. Int. Computers*	Secretary chronically ill due to radiation from computer monitor	32
11	*Reynolds v. Marine Sulphur*	Seaman injured when molten sulfur container fails	33
12	*Crandall v. C&S Railroad*	Train hits car at crossing, injuring driver	34
13	*Dulworth v. Global Elevator*	Shopper injured in fall when escalator suddenly stops	33
14	*Huges v. Jardel*	Store employee raped in mall parking lot	33
15	*Nelson v. Trojan Yachts*	Man nearly drowns when defective boat sinks	33

Descriptions provided by original research team and found in Sunstein et al. (2002), Table 3.1.

S15.1 Glover v. General Assistance

Joan Glover, a five-year-old child, ingested a large amount of a non- prescription allergy medicine called Allerfree, and required a three-week hospital stay. The Allerfree bottle used a faulty childproof safety cap. The Glovers sued the manufacturer of Allerfree, the General Assistance company. The trial jury ordered General Assistance to pay the Glovers $200,000 in compensatory damages.

Facts of the Case Established at Trial. — Joan's parents testified that after her birth they had "child-proofed" their house and ensured that all of their medications had childproof safety caps. The Allerfree bottle carries a label reading "Childproof Cap." Joan found the pills in a kitchen drawer and ingested most of the bottle. The overdose permanently weakened her respiratory system, which will make her more susceptible to breathing-related diseases such as asthma and emphysema for the rest of her life.

General Assistance is a large company (with profits of $100-200 million per year) that manufactures a variety of non-prescription medicines. The company has sold tens of thousands of bottles of medicines with childproof safety caps that were generally effective, but had a failure rate much higher than any others in the industry. Internal company documents showed that General Assistance had chosen to ignore federal regulations requiring more effective safety caps. An internal memo presented at trial says that "this stupid, unnecessary federal regulation is a waste of our money"; it acknowledges the risk that Allerfree might be punished for violating the regulation but says "the punishments are extremely mild; basically we'd be asked to improve the safety caps in the future." An official at the Food and Drug Administration had previously warned a General Assistance executive that the company was "on shaky ground on this one."

Closing Argument by Glovers' Attorney. — The attorney for the Glovers argued that General Assistance's disregard for children's safety and for the law was abhorrent and represented exactly the kind of reckless corporate greed deserving of a high award of punitive damages. He concluded that General Assistance's shocking profit-mongering should be punished so that the company would not feel itself at liberty to put children at risk in the future.

Closing Argument by General Assistance's Attorney. — The attorney for General Assistance emphasized that while the cap had a high failure rate relative to others on the market, it had nonetheless been conceded at trial that the cap was effective in most cases. She argued that, given that the FDA official had only communicated with General Assistance verbally, and had not required the company to take any action, it was not at all clear that the cap was actually in violation of the regulation at all.

S15.2 Judge's Instructions

Taken from *Jardel Company, Inc., et al. v. K. Hughes* (Supreme Court of Delaware, 1987).

The judge has given you the following instructions that you are required by law to use in deciding whether or not to award punitive damages.

The purposes of punitive damages are to punish a defendant and to deter a defendant and others from committing similar acts in the future.

Plaintiff has the burden of proving that punitive damages should be awarded by a preponderance of the evidence. You may award punitive damages only if you find that the defendant's conduct

1. was malicious; or

2. manifested reckless or callous disregard for the rights of others.

Conduct is malicious if it is accompanied by ill will, or spite, or if it is for the purpose of injuring another.

In order for conduct to be reckless or callous disregard of the rights of others, four factors must be present. First, a defendant must be subjectively conscious of a particular grave danger or risk of harm, and the danger or risk must be a foreseeable and probably effect of the conduct. Second, the particular danger or risk of which the defendant was subjectively conscious must in fact have eventuated. Third, a defendant must have disregarded the risk in deciding how to act. Fourth, a defendant's conduct in ignoring the danger or risk must have involved a gross deviation from the level of care which an ordinary person would use, having due regard to all the circumstances.

Reckless conduct is not the same as negligence. Negligence is the failure to use such care as a reasonable, prudent, and careful person would use under similar circumstances. Reckless conduct differs from negligence in that it requires a conscious choice of action, either with knowledge of serious danger to others or with knowledge of facts which would disclose the danger to any reasonable person.

To "establish by a preponderance of the evidence" means to prove that something is more likely so than not so. In other words, a preponderance of the evidence in the case means such evidence, when considered and compared with that opposed to it, has more convincing force, and produces in your minds belief that what is sought to be proved is more likely true than not true.

In your decisions on issues of fact, a corporation is entitled to the same fair trial at your hands as a private individual. All persons, including corporations, partnerships, and other organizations, stand equal before the law, and are to be dealt with by the judge and jury as equals in a court of justice.

The verdict must represent the considered judgment of each juror. In order to return a verdict, it is necessary that each juror agree thereto. Your verdict must be unanimous.

Upon retiring to the jury room, you will select one of your number to act as your presiding juror. The presiding juror will preside over your deliberations.

S15.3 Verdict Task Instructions

Severity Rating

How much should the defendant be punished because of their actions and to deter the defendant and others from similar actions in the future? Note that the compensatory damages that the defendant must pay do not count as part of the punishment. Please circle the number that best expresses the *jury's* judgment of the *appropriate level of punishment*.

None		Mild		Substantial		Severe		Extremely Severe
0	1	2	3	4	5	6	7	8

Dollars

What amount of punitive damages (if any) should the defendant be required to pay as punishment an to deter the defendant and others from similar actions in the future? Note that the compensatory damages that the defendant must pay do not count as part of the punishment. Please write the *amount of punitive damages* that the *jury* agreed on in the blank below.

$ _____

S16 Ethics Statement

This research relies on secondary analysis of data collected by a different group of scholars for a different purpose (Sunstein et al. 2002; Schkade, Sunstein, and Kahneman 2000). Based on all the information we have received from the original researchers, data collection followed all relevant guidelines and contemporaneous best practices for the ethical treatment of human subjects.

Supplemental Information References

Bolin, Robert, Sara Grineski, and Timothy Collins. 2005. "The Geography of Despair: Environmental Racism in the Making of South Phoenix, Arizona, USA." *Human Ecology Review* 12(2): 156–168.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1): 1–22.

Gifford, Donald G., and Brian Jones. 2016. "Keeping Cases from Black Juries: An Empirical Analysis of How Race, Income Inequality, and Regional History Affect Tort Law." *Washington and Lee Law Review* 73(2): 557–651.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Herring, Cedric, and Loren Henderson. 2016. "Wealth Inequality in Black and White: Cultural and Structural Sources of the Racial Wealth Gap." *Race and Social Problems* 8(1): 4–17.

Hughes, John. 2024. "Toward Improved Inference for Krippendorff's Alpha Agreement Coefficient." *Journal of Statistical Planning and Inference* 233: 106170.

Kahan, Dan M., Donald Braman, John Gastil, Paul Slovic, and C. K. Mertz. 2007. "Culture and Identity-Protective Cognition: Explaining the White-Male Effect in Risk Perception." *Journal of Empirical Legal Studies* 4(3): 465–505.

Leech, Geoffrey, Paul Rayson, and Andrew Wilson. 2001. "Companion website for: Word frequencies in written and spoken English: based on the British National Corpus." Longman: London. http://www. comp. lancs. ac. uk/ucrel/bncfreq/lists ....

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–1082.

Schkade, David, Cass Sunstein, and Daniel Kahneman. 2000. "Deliberating about Dollars: The Severity Shift Empirical Study." *Columbia Law Review* 100: 1139–1176.

Sunstein, Cass R., Reid Hastie, John W. Payne, David A. Schkade, and W. Kip Viscusi. 2002. *Punitive Damages: How Juries Decide*. Chicago: University of Chicago Press.

Taddy, Matt. 2013. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association* 108(503): 755–770.

Unnever, James D., Michael L. Benson, and Francis T. Cullen. 2008. "Public Support for Getting Tough on Corporate Crime." *Journal of Research in Crime and Delinquency* 45(2): 163–190.

von der Loo, Mark P. J. 2014. "The stringdist Package for Approximate String Matching." *The R Journal* 6(1): 111–122.