Studying the Role of Emotional Arousal on Supreme Court Voting: Replication of Dietrich, Enos, and Sen (2018)

Diego Arias

Abstract

Dietrich,Enos, and Sen (2018) show that U.S Supreme Court Justices reveal their leanings through their vocal pitch (a measure of their emotional arousal) during oral arguments. This replication looks at how the gender and age of the Justices mediates this effect. This extension was done by adding the predictors of the gender and age of the Justice at the trial to generalized linear models already looking at vocal pitch difference as a predictor of voting behavior. Ultimately, this replication found that the relationship between vocal pitch and voting behavior is much greater among women justices than male justices. This effect was not seen with age, meaning the age of the justice has no effect on the extent to which their vocal pitch differences predicted their vote.

Introduction

Previous studies have shown that metrics like how many questions a Supreme Court justice asks a petitioner and the pleasantness of their word choice during oral arguments can be used to predict their vote (Roberts, 2005; Black et al., 2011). In their paper, "Emotional Arousal Predicts Voting on the U.S. Supreme Court", Dietrich, Enos, anf Sen (2018) extend this finding by analyzing over 3,000 hours of Supreme Court audio recordings to see if justices implicitly reveal their leaning during the oral arguments in their vocal pitch. Their main analysis is a generalized linear mixed-effects model which uses the difference in the vocal pitch that justices speak to each attorney to operationalize their emotional arousal and ultimately predict their eventual vote. Vocal pitch was chosen as it is an important subconscious cue about people's emotion that they many times have little control over. It is important to note that the judges' vocal pitch at different instances was quantified in the number of standard deviations above or below his/her average vocal pitch in the trial. The benefit of using a glmer model is that it uses random intercepts for each justice, which accounts for inherent vocal pitch and emotional expression differences. Ultimately, the authors found that the higher the emotional arousal directed an at attorney (compared to his/her opponent), the less likely they will win the Justice's vote. This model predicted 57.5% of Justice votes accurately and 66.5% of case outcomes accurately. This is as predictive as one of best models in the literature called {Marshall}+ which uses 95 predictors, compared to the authors' using just one. Further analysis showed the vocal pitch measure has unique predictive power, with models 2,3, and 4 using controls in the aforementioned Black et al. (2011) paper as well differences in the use of "pleasant" and "unpleasant" words and finding that these additional variables only increased the predictive power of the model by around seven percentage points.

This present paper attempts to first replicate the main findings of the 2018 paper. Fortunately, the data and code are made easily accessible by the authors on dataverse. This replication, alike the analysis originally done for the paper, is in R. Fortunately, I was able to fully replicate all of the author's code and run all of their models, obtaining the same findings that are reported in the paper. My full code can be found on my github account (diego-arias).

While glad to be able to fully replicate the authors' paper, I was left with a desire to try to extend some of their main models to see if there were any underlying trends in the data that might be able to help create a more predictive model. First, I used a Bayesian generalized linear mixed-effects model using the Stan package to check the initial results of the main model of the paper. This model showed very similar results to those reported in the original frequentist model, and ultimately further showed the robustness of the initially reported effects.

I also decided to see if new predictors could make their models even more accurate. To do this, I added columns in the data frame which coded the gender and age of each justice during each case. A generalized linear mixed effects model looking at the effect of pitch difference and age on voting outcome showed no significant effect of the age coefficient or the interaction of age and vocal pitch difference. This shows that the age of a Justice during the case does not influence their ultimate vote, and more interestingly does not affect the extent to which vocal pitch differences predict their ultimate vote. More interestingly, a generalized linear mixed effects model looking at the effect of pitch difference and gender on voting outcome showed a slight interaction between gender and vocal pitch difference. Notably, a female justice's vocal pitch was found to be more predictive of their ultimate vote than a male's. This new model, however, was not more predictive of judge's ultimate voting behavior than the author's original one.

Literature Review

The findings of Dietrich, Enos, and Sen (2018) provides a whole new approach to modeling Supreme Court voting than what is in the previous literature. Previously, research had found that the lawyer who is asked more questions during the oral arguments is more likely to lose the case (Roberts, 2005). This was replicated by Epstein and colleagues in 2010, who in addition found that a lawyer who is asked longer questions is also more likely to lose. Building off of these concepts and also previous psychology literature on word choice, Black and colleagues (2011) used data from cases going back to 1979 and found that the more a Justice uses unpleasant words toward an attorney, the less likely the attorney is to win the case. This makes intuitive sense, as you would expect a justice to use more unpleasant words, whether that be on purpose or subconsciously, toward an attorney who he currently disagrees with more.

The Dietrich, Enos, and Sen (2018) paper build on this previous work to instead argue that how the justices speak toward attorneys may have more predictive power than the words that they use. Vocal pitch was chosen as a measure of the justices true emotional responses is that that many times inflections can occur unbeknown to the speaker (Ekman et al., 1991). For example, many times emotional arousal can dictate a higher vocal pitch, as the vocal cords become tighter, without conscious awareness (Mauss and Robinson, 2009).

For this reason, it is safe to assume that studying vocal pitch can give cues about a speaker's internal emotions that he himself may choose to not disclose in the content of his actual speaking. This is especially valuable when analyzing Supreme Court cases, where Justices might try to conceal their emotions but their vocal pitch might by giving away which attorney they are most in disagreement with, and thus are more likely to vote against.

Replication

As I previously stated, I was able to replicate all of the major findings from the original paper. This includes both the generalized linear mixed-effects models and also the main table. In the appendix, I provide my replication of the main table of the original paper

Extension

In an effort to get have the nos accurate data points for my extension the first thing I did was get rid of extreme outliers. Through visualizing the vocal pitch variable with a boxplot, I found and removed an outlier that was very distance from even all other outliers. While this did not end up changing any of the coefficients of the original model, I decided to use this filtered data set for all of my analysis in an effort to create the best fit possible.

The second exploration I did was checking the authors' original main model with a Bayesian generalized linear mixed effects model (achieved using the rstanarm packages). On the whole, the coefficients of both models very similar. More specifically, both the intercept and vocal pitch effect coefficients are well between each other's standard errors in both models. This helps solidify the initial results from the model that higher pitch difference toward a layer leads to a lower likelihood of voting in their favor, which is best seen in Figure 1.

After checking their original model with a Bayesian approach, I tried to see if I could, by adding predictors, create a new model that would better predict the voting outcome. While certain things like political leaning might have really added to the predictive power of the model, I wanted to look for things that might specifically interact with the vocal pitch effect that the authors found. More specifically, I wanted to see if there was any difference between how age and gender interact with this effect, and if accounting for this interaction could create a better model.

I first created a generalized linear mixed effects model that included vocal pitch, the age of the justice, and the interaction of those two variables as predictors for the ultimate vote. Ultimately, age has a nonsignificant on voting outcome (p=0.59) and is also a negligible mediator on the effect of voice pitch difference on voting outcome (with an interaction p value of 0.80). This can best seen in Figure 2, which shows how there are virtually no differences in the effect that pitch difference has on voting across different age brackets. This means that the effect that the authors originally discovered, that the larger the pitch difference directed at the petitioner, the less of a chance the judge will vote in favor of him, is not strongly affected by the age of the judge. A possible reason for this is that justices tend to be older (with this sample ranging from 50-90), and thus might present a sample with less vocal variability. Future studies that use vocal pitch measures to predict behavior with different demographics should consider studying the interraction with age.

My next attempt to improve the authors' original model was creating a generalized mixed effects model which included vocal pitch difference, gender, and the interaction between the two variable's as predictors. The model ultimately finds no effect of the gender of the justice on the final vote (p = .91) and a slight interaction between the gender and voice pitch (p=.066). Although this effect is not quite significant, it proposes that female justices pitch differences might better predicts their voting behavior (because there is a stronger relationship between pitch differences and voting behavior among female justices). To better understand this effect, I reduced the model to a linear regression. Once we get rid of the random intercept per judge, the gender and vocal pitch interaction is significant (p = .049). This demonstrates how in the real world there seems to be a difference in how vocal pitch predicts voting among female and male justices, but this difference is made insignificant when using a generalized linear mixed effects model that already accounts for random variation within a small number of justices. This is best seen in Figure 3, where the red lines show that voting behavior in females is more strongly predicted by vocal pitch differences.

Unfortunately, this generalized linear mixed effects model ultimately does not perform better than the authors' original model at predicting voting (it is still correct 57% of the time). Nevertheless, if you divide the data into two different data sets, one female and one male, and run the original model on each, the female vocal pitch model predicts 58.4% of votes correctly while the male vocal pitch model only predicts 56.2% of votes correctly. This further shows that this predictor is stronger in women.

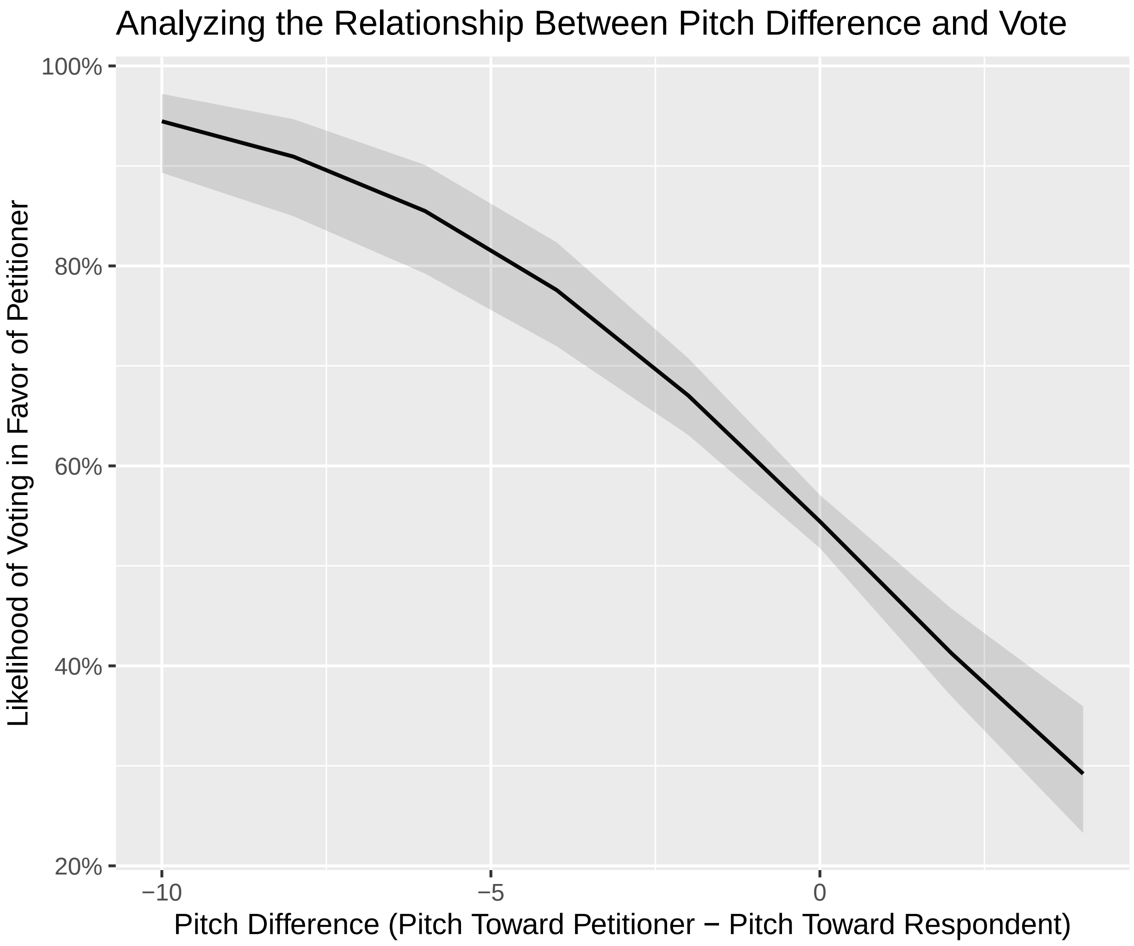

Figure 1

This figure presents the main result found by Dietrich, Enos, and Sen (2018): that if a justice speaks to an attorny in a higher pitch,there is a smaller probability that the Justice will vote in favor of them. This can be seen in the plot because as the pitch difference toward the petioner relative to the respondent increases, the likelihood of the Justice voting in favor of the petitioner decreases.

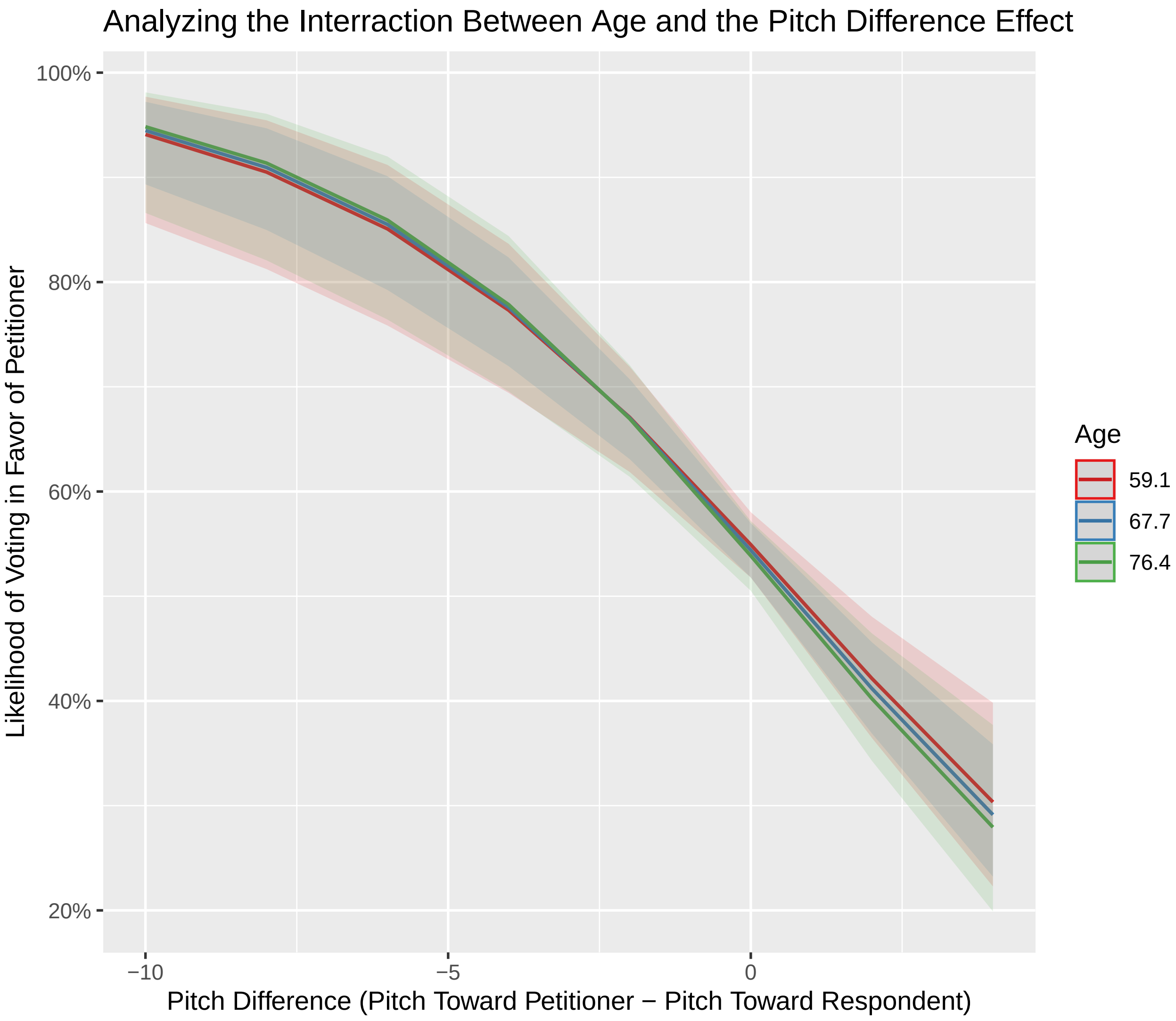

Figure 2

This plot shows the lack of an interraction between the age of the Supreme Court Justice and vocal pitch difference predictors when analyzing the Supreme Court Justices' ultimate vote. This means that the effect reported by the original authors' does not depend on how old the Justice is during the oral argument. Still, it is possible that no interaction was found do to the small range of ages (50-90) since vocal pitch might be a better predictor for younger populations.

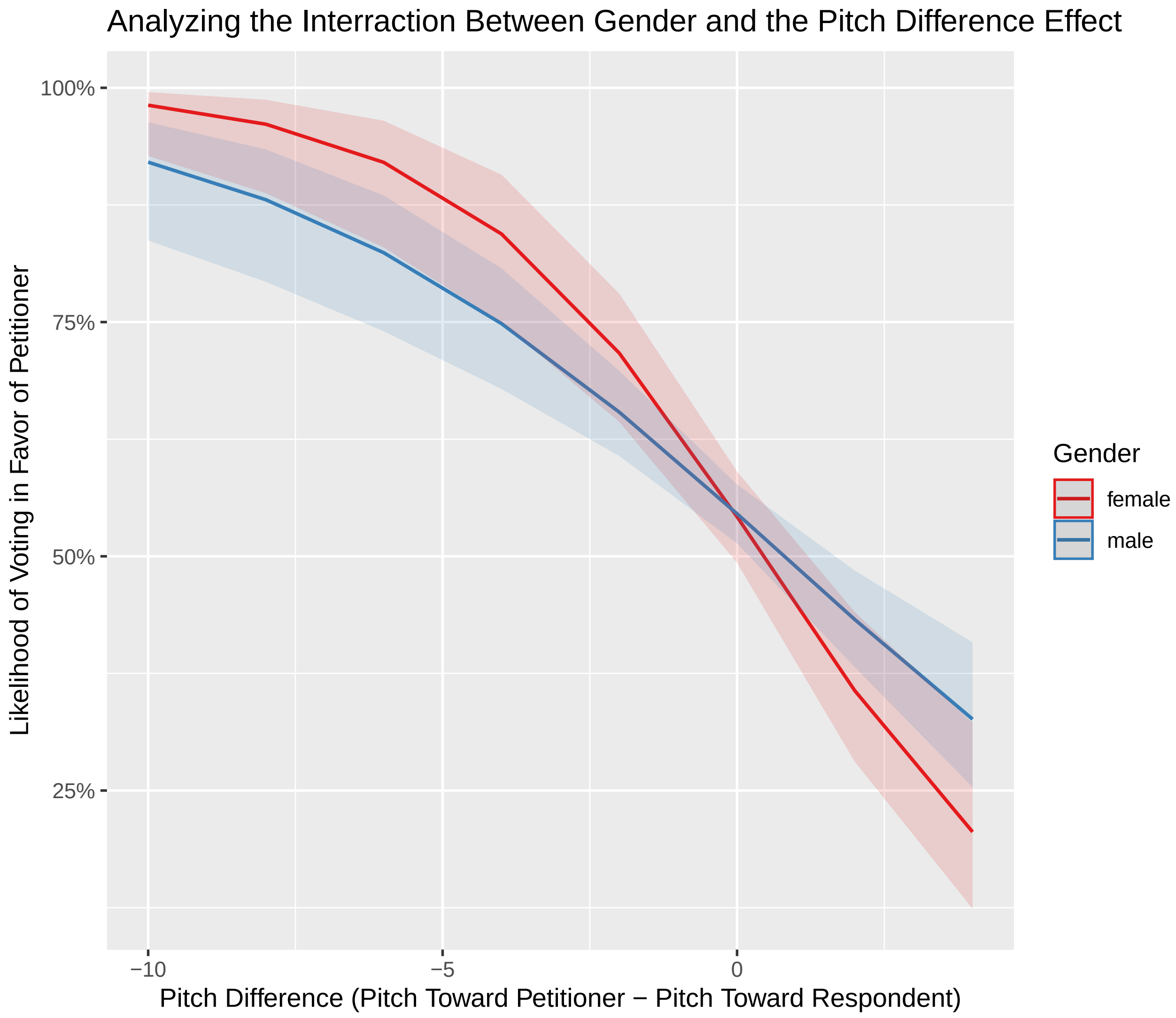

Figure 2

This plot shows the interraction between the gender of a Justice and their vocal pitch difference measure when analyzing their ultimate vote (p = 0.59) Ultimately, the pitch difference measure has a larger effect in predicting the vote among female justices than male ones (as seen with how the slope of the line being much steeper for females). When the model is divided into two subsets, one with the female justices and one with the male ones, the female vocal pitch model predicts 58.4% of votes correctly, while the male vocal pitchmodel predicts 56.2% of votes correctly, further supporting the point that this predictor is stronger in women.

Supplemental Table

This table combines the results from the main models of my extension. Models 1-3 are all generalized linear mixed-effects models (model 1 is the bayesian version of the original model, model 2 accounts for the age interraction, and model 3 accounts for the gender interraction). Model 4 is a logistic regression model that more specifically looks at the interraction of gender and vocal pitch

Table 1:

	generalized linear mixed-effects			*logistic*
	(1)	(2)	(3)	(4)
Constant	0.183*** p = 0.004	0.344 p = 0.270	0.169* p = 0.093	0.141*** p = 0.009
pitch_diff	−0.227*** p = 0.00000	−0.194 p = 0.501	−0.380*** p = 0.00000	−0.388*** p = 0.00000
age		−0.002 p = 0.590		
pitch_diff:age		−0.001 p = 0.803		
femalemale			0.013 p = 0.915	0.034 p = 0.585
pitch_diff:femalemale			0.153* p = 0.067	0.163** p = 0.049
Observations	3,784	5,208	5,208	5,208
Log Likelihood	−2,585.171	−3,551.551	−3,550.016	−3,557.388
Akaike Inf. Crit.	5,176.342	7,113.101	7,110.032	7,122.777
Note:			*p<0.1; **p<0.05; ***p<0.01	

Conclusion

This paper set out to replicate and extend the results of Dietrich, Enos, and Sen (2018) in their paper, "Emotional Arousal Predicts Voting on the U.S. Supreme Court". This paper was able to replicate all of the main results from the original study, supporting the authors' main finding that the higher the emotional arousal directed an at attorney (measured through vocal pitch), the less likely they will win the Justice's vote. Similarly to the authors, I found this model to predict 57.5% of Justice votes accurately and 66.5% of case outcomes accurately.

My attempt to extend this research started by removing an outlier in the data among the pitch difference vairable. I then checked the validity of the author's original model with a Bayesian generalized liner mixed effects model in Bayesian. Since the model yielded similar results, this further supported the validity of the original claim.Moreover, this extension created two new models, which used the age and gender of the justice in addition to their pitch difference to predict voting. While neither model found significant results, a closer look shows that there is a significant interaction between gender and the effect of vocal pitch difference on voting behavior when not taking into account the random variation per Justice. This, and the fact that the pitch difference model is much more predictive in the female subset of the data (predicting 58.4% of votes correctly) than the male (predicting only 56.2% of votes correctly), shows that the the original effect that the authors' found is much stronger among female justices than male justices. While this sort of interraction was not found with the age variable, it is important to remember how Supreme Court Justices in the dataset are never younger than 50 and an effect might have been found in a different sample.

This paper shows how there are slight differences in how people present subconscious cues such as vocal pitch differences. Future research should follow as understanding these differences might allow us to better predict outcomes like voting behavior in the future.

References

Black, Ryan C., Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. “Emotions, oral arguments, and Supreme Court decision making.” The Journal of Politics 73, no. 2 (2011): 572-581.

Dietrich, Bryce J., Ryan D. Enos, and Maya Sen. “Emotional arousal predicts voting on the US supreme court.” Political Analysis 27.2 (2019): 237-243

Ekman, Paul, Maureen O’Sullivan, Wallace V. Friesen, and Klaus R. Scherer. “Invited article: Face, voice, and body in detecting deceit.” Journal of nonverbal behavior 15, no. 2 (1991): 125-135.

Mauss, Iris B., and Michael D. Robinson. “Measures of emotion: A review.” Cognition and emotion 23, no. 2 (2009): 209-237.

Roberts Jr, John G. “Oral Advocacy and the Re-emergence of a Supreme Court Bar.” Journal of Supreme Court History 30, no. 1 (2005): 68-81.

Appendix

Results from Dietrich et al. (2018) were successfully replicated. As an example, this is a replicated version of Table 1 of the original paper

Table 2:

	intercept only	no controls	dal	harvard	liwc
	(1)	(2)	(3)	(4)	(5)
Constant	0.201***	0.178***	−0.025	−0.027	−0.026
	(0.055)	(0.055)	(0.160)	(0.160)	(0.160)
pitch_diff		−0.266***	−0.214***	−0.214***	−0.214***
		(0.036)	(0.038)	(0.038)	(0.038)
I((petitioner_neg_words/petitioner_wc) - (respondent_neg_words/respondent_wc))			−1.973	0.071	−2.132
			(1.467)	(0.845)	(1.310)
I((petitioner_pos_words/petitioner_wc) - (respondent_pos_words/respondent_wc))			−1.650	0.272	−1.676
			(1.084)	(0.685)	(1.047)
I(petitioner_count - respondent_count)			−0.057***	−0.057***	−0.057***
			(0.008)	(0.008)	(0.008)
lagged_ideology			0.158***	0.158***	0.158***
			(0.033)	(0.033)	(0.033)
conservative_lc			0.011	0.012	0.012
			(0.073)	(0.073)	(0.073)
I(lagged_ideology ∗conservative_lc)			−0.263***	−0.264***	−0.263***
			(0.034)	(0.034)	(0.034)
sgpetac			0.540***	0.543***	0.540***
			(0.079)	(0.079)	(0.079)
sgrespac			−0.672***	−0.666***	−0.673***
			(0.104)	(0.104)	(0.104)
petac			0.039***	0.039***	0.039***
			(0.008)	(0.008)	(0.008)
respac			−0.058***	−0.058***	−0.058***
			(0.007)	(0.007)	(0.007)
petNumStat			0.045***	0.046***	0.045***
			(0.014)	(0.014)	(0.014)
respNumStat			−0.003	−0.003	−0.003
			(0.014)	(0.014)	(0.014)
Observations	5,208	5,208	4,976	4,976	4,976
Log Likelihood	−3,580.577	−3,551.717	−3,201.603	−3,203.514	−3,201.121
Akaike Inf. Crit.	7,165.153	7,109.434	6,433.206	6,437.028	6,432.241
Note:					*p<0.1; **p<0.05; ***p<0.01