Emotional Arousal Predicts Voting on the U.S. Supreme Court

First Draft: August 24, 2015
This Draft: August 2, 2017

Abstract

Do judges telegraph their preferences during oral arguments? Using the U.S. Supreme Court as our example, we demonstrate that Justices implicitly reveal their leanings during oral arguments, even before arguments and deliberations have concluded. Specifically, we extract the emotional content of over 3,000 hours of audio recordings spanning 30 years of oral arguments before the Court. We then use the level of emotional arousal, as measured by vocal pitch, in each of the Justices' voices during these arguments to accurately predict many of their eventual votes on these cases. Our approach yields predictions that are statistically and practically significant and robust to including a range of controls; in turn, this suggests that subconscious vocal inflections carry information that legal, political, and textual information do not.

Introduction

In February of 2017, the Trump Administration issued an executive order that banned the entry of people from seven majority-Muslim countries into the U.S. and unleashed chaos in airports all over the world. Legal challenges were swift and, within two weeks, the 9th Circuit Court of Appeals had scheduled oral arguments concerning the order. Public interest in how the court would rule was significant. More than 130,000 people listened to the arguments, and hundreds of experts weighed in on how the judges would vote. Many of these predictions relied on the three judges' emotional reactions and vocal expressions during oral argument. For example, the *New York Times* provided a live analysis of the judges' reactions, assessing whether they were "pretty skeptical" or "friendly."[1] Ultimately, the three-judge panel ruled unanimously against the travel ban, but it was not without days of uncertainty for those affected.

In this paper, we do what observers of the 9th Circuit oral argument were attempting by asking whether we can systematically detect how judges will vote based on emotional responses at oral argument. We address this by examining the U.S. Supreme Court, which has decades of audio data. Specifically, we explore audio from nearly 3,000 hours of oral arguments from the last 30 years and find that *vocal pitch alone* is strongly predictive of Supreme Court Justices' votes. The results are robust to the inclusion of other factors and predict outcomes at least as well as more complex models accounting for substantive features of the actors and cases, suggesting that vocal pitch predicts decisions in ways that characteristics like ideology or legal issue areas do not. In results presented in the Supplemental Information, moreover, we extend our findings to the 9th Circuit's travel ban argument, suggesting that our approach has good external validity.

As we discuss below, our findings are consistent with several causal narratives. For example, it could be the case that Justices actively rely on their emotions in reaching important

[1] `https://www.nytimes.com/interactive/2017/02/07/us/ninth-circuit-oral-arguments-trump-immigration.html`.

decisions. It may also be that judges experience aroused responses during oral arguments because they receive information conflicting with previously made decisions. Although we cannot disaggregate these explanations, our results clearly show that non-substantive and implicit signals, even among elite actors such as federal judges and Supreme Court justices, can provide additional meaningful information on their attitudes beyond what can be found in their textual pronouncements.

Description of Supreme Court Oral Arguments Audio Data, Emotional Arousal, and Vocal Pitch

We are not the first to suggest that emotion plays an important role in judicial oral arguments (e.g., Johnson et al., 2009; Shullman, 2004). For example, Black et al. (2011) argue that the "tenor" of oral arguments can be used as a barometer of how Justices will rule in a given case. While others have utilized the number of questions directed towards each side (Roberts, Jr., 2005; Epstein, Landes and Posner, 2010), Black et al. (2011) captured the emotions expressed during oral arguments using the number of "pleasant" and "unpleasant" words. They find that the more unpleasant words Justices use towards an attorney, the less likely that attorney will prevail in the case. Even though we acknowledge the importance of what Justices say, we suggest *how* they say those words may be of equal, if not greater, predictive importance.

Why might vocal pitch predict the behavior of even elite actors like Supreme Court Justices and why might the emotion signaled by vocal pitch be more important than the substantive content? First, changes in vocal inflections, like pitch, often occur unbeknownst to the speaker (Ekman et al., 1991). For the Justices, emotional arousal may be more likely when interacting with someone with whom they disagree. When this occurs, the heart begins to race, palms begin to sweat, and all muscles, including the vocal cords, tighten (Posner, Russell and Peterson, 2005). This is the primary reason why "the most consistent association reported in the literature is between arousal and vocal pitch, such that higher levels of arousal

have been linked to higher-pitched vocal samples" (Mauss and Robinson, 2009, 222). For example, Laukka, Juslin and Bresin (2005) asked actors to portray "weak" and "strong" versions of a variety of emotions that were later judged by amateur and expert judges on the degree to which the actors were displaying an "activated" or "intense" emotional state. Not only was vocal pitch correlated with both activation and intensity, but the "patterns of vocal cues for activation and emotion intensity showed numerous similarities," suggesting both may be capturing the same "physiological reaction" (648).[2]

Moreover, due to the automatic nature of this response, a speaker's vocal pitch will often provide insights into a speaker's level of activation beyond their conscious communication. Indeed, "several studies have shown, that like the body, the tone of a person's voice leaks information that is not revealed by the verbal content or facial expressions associated with the message" (Zuckerman and Driver, 1985, 129). (Additional discussion of this literature can be found in the Supplemental Information.) For these reasons, verbal and nonverbal behavior should be thought of in terms of a "leakage hierarchy" with "verbal content" located in the "controllable end of the continuum, whereas the body and tone of voice may be classified as less controllable and more leaky channels" (Zuckerman and Driver, 1985, 130). This suggests that Justices may subconsciously indicate their ultimate preferences towards a case by raising their vocal pitch towards either the petitioner or respondent. This is the case regardless of whether the Justice formulates her response contemporaneously or whether she is reacting on the basis of predispositions about the case.

Second, some Justices, like Antonin Scalia, may be more willing to express emotion as compared to others. Such differences are problematic for text-based measures since they only capture emotion that is verbalized. According to Russell (2003), this ignores a range of emotions which occur prior to conscious awareness. As analogy, consider felt body temperature.

[2]In a similar study, Bänziger and Scherer (2005) found actors portraying emotions with "high" activation not only spoke with a higher average vocal pitch (mean F_0), but they also found little variation in the F_0 contours, suggesting that "simple summaries"—such as F0 mean or F0 range—were sufficient to account for the most important variations observed between emotion categories" (265).

Even though our body's temperature changes all the time, we do not always identify those changes as being *hot* or *cold*. For some, a small decrease in temperature may be enough to say, "I am cold!" For others, that same decrease may not even be recognized. Emotional expression on the Supreme Court functions in a similar way—some is verbal, some is non-verbal. For some Justices, an attorney's error may be egregious enough to warrant calling it "idiotic," while for others that same error may not even raise an eyebrow, verbal or otherwise. Text-based measures are extraordinarily useful when one is interested in understanding the former, but struggle with latter.[3]

To achieve this end, we collected audio recordings from oral arguments in 1,773 cases, beginning in 1982 and ending in 2014. Using the timestamps provided by the Oyez Project,[4] we further parsed these cases into discrete segments of audio uttered by (1) the Justices themselves, (2) the lawyer/s representing the petitioner, and (3) the lawyer/s representing the respondent.[5] Lawyers spoke for 2,137 hours. Justices spoke for 502 hours. For the Justices, this represented 146,335 discrete utterances. Additional descriptive statistics are provided in the Supplemental Information, Tables S1–S2.

Results: How Emotion Arousal Predicts Supreme Court Justices' Voting

We expect that a Justice who is more emotionally activated when speaking towards an attorney will be more likely to vote against that attorney. If this is correct, a higher vocal pitch will predict a stronger negative response. To analyze this, we code whether a Jus-

[3]Note that some non-verbal responses could still be consciously made—for example, a Justice explicitly raising the tone of her voice to talk over another Justice. Our approach actually gauges the predictive power of both conscious and subconscious non-verbal expression. However, as we note in the SI, we have reason to think that much (though not certainly all) variation in vocal pitch operates in a subconscious level and is un-controllable.

[4]https://www.oyez.org. More information on the audio data acquisition is in the Supplemental Information.

[5]The "petitioner" is the party bringing the case to the Court; the "respondent" is the party responding to the petitioner's claim. Both parties are represented by separate counsel, often experienced attorneys from the Supreme Court bar.

tice votes in favor of the petitioner, a 1 or 0 variable.[6] We construct a measure of "Pitch Difference" by subtracting vocal pitch in questions directed towards petitioners from vocal pitch in questions directed toward respondents. (Vocal pitch was measured using *Praat*, a speech synthesis program that estimates the fundamental frequency by dividing the autocorrelation of a windowed signal by the autocorrelation of the window itself. To estimate the fundamental frequency we only use voiced speech. More details can be found in the Supplemental Information.) For each Justice, we converted vocal pitch to standard deviations above and below his or her average vocal pitch, which accounts for systematic differences between Justices (for example between male and female Justices) as well as any measurement error associated with extracting the fundamental frequency.[7]

The main results are presented in Table 1, Model 2. All models are multilevel logistic regressions with random intercepts for each Justice.[8] These results show *that the higher emotional arousal or excitement directed at an attorney compared to his or her opponent, the less likely that attorney is to win the Justice's vote* ($p < 0.001$).[9] From Model 1, when the vocal pitch of questions directed to both sides is the same, the predicted probability of a Justice voting for the petitioner is 0.55. However, the probability of a Justice voting for the petitioner drops by 7 percentage points if the difference between the vocal pitch directed to the petitioner is one standard deviation higher than the vocal pitch directed at the respondent. The overall prediction rate is also reported.[10] Here, we are able to predict

[6]Since the outcome is 1 (Justice votes for the petitioner) or 0 (Justice votes against the petitioner), we use a multilevel logistic regression, which we implemented here via the `lme4` package in `R` statistical software language.

[7]We explore these issues in the "Estimating the Fundamental Frequency" section of the Supplemental Information. Since male cords are typically longer (17.5–25mm) than female cords (12.5–17.5mm), men tend to talk at a lower vocal pitch—making standardization essential for comparing male and female Justices. Such standardization also helps account for errors that may occur when estimating the fundamental frequency, although we show this is unlikely to be consequential when estimating mean vocal pitch.

[8]Due to space limitations, we do not describe all the control variables in the main text. Full descriptions and additional model specifications can be found in the Supplemental Information.

[9]Since we cannot assume that Justices' votes within cases are independent, we include a randomly varying intercept for each Justice. This additional parameter also helps account for other within-Justice differences.

[10]We used a .50 threshold for these calculations, meaning when the model returned a predicted probability greater than .50, we predicted the Justice would vote for the petitioner.

Table 1: Does Vocal Pitch Predict Votes in Favor of the Petitioner?

	Intercept Only	No Controls	DAL	Harvard IV	LIWC
	(1)	(2)	(3)	(4)	(5)
Fixed Effects					
Constant	0.202^{***}	0.178^{***}	-0.025	-0.027	-0.026
	(0.056)	(0.055)	(0.160)	(0.160)	(0.160)
Pitch Difference		-0.266^{***}	-0.214^{***}	-0.215^{***}	-0.214^{***}
		(0.036)	(0.038)	(0.038)	(0.038)
% More Unpleasant Words Directed at Petitioner			-1.971	0.071	-2.129
			(1.471)	(0.846)	(1.308)
% More Pleasant Words Directed at Petitioner			-1.647	0.272	-1.673
			(1.086)	(0.685)	(1.046)
# More Questions Directed at Petitioner			-0.057^{***}	-0.057^{***}	-0.057^{***}
			(0.008)	(0.008)	(0.008)
Political Ideology$_{t-1}$			0.158^{***}	0.158^{***}	0.158^{***}
			(0.033)	(0.033)	(0.033)
LC Decision Was Conservative			0.011	0.012	0.012
			(0.073)	(0.073)	(0.073)
Political Ideology$_{t-1}$ × LC Decision Was Conservative			-0.263^{***}	-0.264^{***}	-0.263^{***}
			(0.034)	(0.034)	(0.034)
Solicitor General as Amicus Supporting Petitioner			0.540^{***}	0.543^{***}	0.540^{***}
			(0.079)	(0.079)	(0.079)
Solicitor General as Amicus Supporting Respondent			-0.672^{***}	-0.666^{***}	-0.673^{***}
			(0.104)	(0.104)	(0.104)
# of Amicus Briefs Supporting Petitioner			0.039^{***}	0.039^{***}	0.039^{***}
			(0.008)	(0.008)	(0.008)
# of Amicus Briefs Supporting Respondent			-0.058^{***}	-0.058^{***}	-0.058^{***}
			(0.007)	(0.007)	(0.007)
Petitioner's Level of Resources			0.045^{***}	0.046^{***}	0.045^{***}
			(0.014)	(0.014)	(0.014)
Respondent's Level of Resources			-0.003	-0.003	-0.003
			(0.014)	(0.014)	(0.014)
Random Effects					
Intercept	0.03^{***}	0.03^{***}	0.02^{**}	0.02^{**}	0.02^{**}
	(0.16)	(0.16)	(0.13)	(0.12)	(0.12)
N_1	5,209	5,209	4,977	4,977	4,977
N_2	18	18	18	18	18
$\log L$	$-3{,}581.134$	$-3{,}551.721$	$-3{,}201.611$	$-3{,}203.521$	$-3{,}201.128$
AIC	7,166.268	7,109.441	6,433.222	6,437.041	6,432.256
Percent Correctly Predicted	54.92	57.50	63.91	63.49	63.67

Note: Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Models include statements with question marks. The average vocal pitch in questions directed towards the petitioner ("Petitioner Pitch") minus the average vocal pitch in questions directed towards the respondent ("Respondent Pitch") is captured in "Pitch Difference" (Petitioner Pitch - Respondent Pitch). Model 2 uses the Dictionary of Affect in Language (DAL). Model 3 uses the Harvard-IV dictionary. Model 4 uses the Linguistic Inquiry and Word Count (LIWC) dictionary. The rest of the controls are the same as Black et al. (2011). Please refer to the Supplemental Information for more details about each dictionary, the controls, and our cross-validation approach. Levels of significance are reported as follows: $^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$.

57.50 percent of Justices' votes accurately (see Table 1) and 66.55 percent of overall case outcomes accurately (see Table S4) using only pitch difference, suggesting that vocal pitch predicts not only how individual Justices vote but also the eventual disposition of the case.[11]

Models 2, 3, and 4 include the controls used by Black et al. (2011), as well as the differences in the use of "pleasant" and "unpleasant" words as defined by the Dictionary of Affect in Language (DAL), the Harvard IV dictionary (also known as the General Inquirer), and the Linguistic Inquiry and Word Count (LIWC) dictionary, respectively. Since the Harvard IV dictionary is publicly available, we provide the words used for Model 3 in the Supplemental Information. For Models 2 and 4, we provide some examples of "positive" and "negative" words. Unlike Black et al. (2011), we use the Martin-Quinn scores estimated in the previous term, as Martin-Quinn scores are dynamically estimated within each term using Justices' votes, which would introduce endogeneity if not lagged. These are continuous measures from liberal (-1.0) to conservative (+1.0) and vary from Justice to Justice and from term to term.

After accounting for vocal pitch, including these other variables only increases the predictive power of the model by around seven percentage points, suggesting that vocal pitch has unique predictive value. To further assess the substantive importance of vocal pitch, we compared the performance of vocal pitch (and of only vocal pitch) to a widely known algorithm developed by Katz, Bommarito and Blackman (2014), known as {Marshall}+. This algorithm uses 95 variables to predict Supreme Court Justices' voting and is known as one of the most predictive algorithms currently available.[12] As explained in the Supplemental Information, we restricted our analysis to the period from 1998 to 2012. In total, {Marshall}+ successfully predicts 64.76 percent of cases correctly, which is 1.79 percentage points *lower* than our prediction rate of 66.55 percent of cases. A simple χ^2 test reveals the

[11] In the Supplemental Information, we also report the true positive, true negative, false positive, and false negative rates – comparing each to a variety of previous studies. These statistics are included as part of a broader discussion of how best to assess predictive performance.

[12] These 95 variables include case information, ideological information, Supreme Court trends, Justice background characteristics, etc.). See `http://lexpredict.com/portfolio/predicting-the-supreme-court`.

models are similarly predictive ($\chi^2 = 0.52$, $df = 1$, $p > 0.05$), suggesting we are able to equal the predictive power of a model that uses 95 predictors using only one—vocal pitch.

Model 1 also outperforms traditional petitioner-based models in which a "plaintiff always wins" rule is applied. Although seemingly simple, such a rule is actually fairly sophisticated and takes into account a lot of what scholars know about strategic planning, the rule of four, and principles of precedent. Using only vocal pitch, Model 1 significantly ($p \leq 0.001$) outperforms this baseline by 2.58 percentage points. Even though Models 2, 3, and 4 all perform better, when one only uses the number of "positive" and "negative" words, the prediction rates are substantially worse. For example, when the only predictors are the percent more positive and negative words directed at the petitioner, the model successfully predicts 0.81 and 0.72 percentage points better than the "plaintiff always wins" model, depending on whether one uses the DAL or LIWC dictionaries, respectively. The Harvard-IV dictionary actually predicts 8.25 percentage points *less* than this baseline. As we show in the Supplemental Information, regardless of the text-based measure one uses, vocal pitch does substantially better at predicting both cases and votes. These results are not meant to suggest that vocal pitch is the *only* variable that should be incorporated into models of oral arguments. Nor do our results suggest text-based measures have no place in the study of emotional expression on the Supreme Court. Rather, our results demonstrate vocal pitch should be one of many variables that should be taken into consideration when assessing oral arguments.

Discussion

For scholars interested in predicting Justice votes weeks, if not months, before the Court's ruling is released, oral arguments "provide a barometer of how justices will rule in a given case" (Black et al., 2011, 574). While we are not the first to suggest emotional expressions are an important part of such prognostications, the vast majority of these studies have only considered text-based measures (for important exceptions, see Schubert et al., 1992; Knox

and Lucas, 2017). We show vocal pitch on its own is about as predictive of Justices' votes and overall case outcomes as models that use all publicly available quantitative legal and non-legal information, including additional textual information related to emotion. These comparisons are not meant to suggest that vocal pitch is the *only* variable scholars should use when assessing emotional expression on the Supreme Court. We argue the {Marshall}+ algorithm, text-based measures, and the "petitioner always wins" rule can (and should) be used to predict Justice votes. However, non-verbal signals, including changes in vocal pitch, also carry considerable weight. Justices choose their words carefully, but have far less control over how those words are spoken—and these subconscious vocal cues, our findings show, carry important information about eventual rulings.

References

Bänziger, Tanja and Klaus R Scherer. 2005. "The Role of Intonation in Emotional Expressions." *Speech Communication* 46(3):252–267.

Black, Ryan C., Sarah A. Treul, Timothy Johnson and Jerry Goldman. 2011. "Emotions, Oral Arguments, and Supreme Court Decision Making." *The Journal of Politics* 73(2):572–581.

Ekman, Paul, Maureen O'Sullivan, Wallace V. Friesen and Klaus R. Scherer. 1991. "Invited Article: Face, Voice, and Body in Detecting Deceit." *Journal of Nonverbal Behavior* 15(2):125–135.

Epstein, Lee, William M. Landes and Richard A. Posner. 2010. "Inferring the Winning Party in the Supreme Court from the Pattern of Questioning at Oral Argument." *The Journal of Legal Studies* 39(2):433–467.

Johnson, Timothy R., Ryan C. Black, Jerry Goldman and Sarah A. Treul. 2009. "Inquiring Minds Want to Know: Do Justices Tip Their Hands with Questions at Oral Argument in the U.S. Supreme Court?" *Washington University Journal of Law & Policy* 29:241–261.

Katz, Daniel Martin, Michael James Bommarito and Josh Blackman. 2014. "Predicting the Behavior of the Supreme Court of the United States: A General Approach." *Available at SSRN 2463244* .

Knox, Dean and Christopher Lucas. 2017. "A General Approach to Classifying Mode of Speech: The Speaker-Affect Model for Audio Data." Unpublished manuscript.

Laukka, Petri, Patrik Juslin and Roberto Bresin. 2005. "A Dimensional Approach to Vocal Expression of Emotion." *Cognition & Emotion* 19(5):633–653.

Mauss, Iris B. and Michael D. Robinson. 2009. "Measures of Emotion: A Review." *Cognition and Emotion* 23(2):209–237.

Posner, Jonathan, James A. Russell and Bradley S. Peterson. 2005. "The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive development, and Psychopathology." *Development and Psychopathology* 1y:715–734.

Roberts, Jr., John G. 2005. "Inquiring Minds Want to Know: Do Justices Tip Their Hands with Questions at Oral Argument in the U.S. Supreme Court?" *Journal of Supreme Court History* 30:68–81.

Russell, James A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110:145–172.

Schubert, James N., Steven A. Peterson, Glendon Schubert and Stephen Wasby. 1992. "Observing Supreme Court Oral Argument: A Biosocial Approach." *Politics and the Life Sciences* 11(1):35–51.

Shullman, Sarah Levien. 2004. "The Illusion of Devil's Advocacy: How the Justices of the Supreme Court Foreshadow Their Decisions during Oral Argument." *Journal of Appellate Practice and Process* 6:271–293.

Zuckerman, Miron and Robert E. Driver. 1985. Telling Lies: Verbal and Nonverbal Correlates of Deception. In *Multichannel Integrations of Nonverbal Behavior*, ed. Aaron W. Siegman and Stanley Feldstein. New York, NY: Lawrence Erlbaum chapter 3, pp. 129–147.

Supporting Information (to go online) for: Emotional Arousal Predicts Voting on the Supreme Court

Description of Supreme Court Audio Data

Although the U.S. National Archives maintain the official audio recordings, for this paper we accessed the audio files in .mp3 format via the Oyez Project at the Chicago-Kent College of Law, which is a public repository that aims "to be a complete and authoritative source for all audio recorded in the Court since the installation of a recording system in October 1955." To get additional information on the content of the files and to verify the identity of each speaker, we also obtained the transcript for each argument, which can be found in the Oyez Project player.

As explained on Oyez,[1] we do not have audio for every case. For example, some of the audio from the 1970s and 1990s resides in the National Archives in tape format. Even though Oyez is actively transferring this to a digital format, they still have not completed this process. There was also a problem with the Court's tape machine, especially during the early 1990s, which made the Justices talk slightly faster than normal. In order to synchronize these recordings with the transcripts, the Oyez team (especially Pat Ward) had to manually re-engineer the audio to make it "faithful to the speakers' voices."

Other problems arose in the 1980s. Unbeknownst to the Court, their tape manufacturer changed the formula they used to create the reel-to-reel tapes used for oral arguments. Instead of lasting 80 years, like typical reel-to-reel tapes, these new tapes only lasted a few years. Ultimately, they were plagued by what Oyez calls the "sticky shed syndrome" in which tape reels tended to stick together when stored. To solve this problem, the Oyez team had to bake the tapes in a slow oven for several hours, then mount the tapes and dub them to another real or digital storage device. Unfortunately, this was only partially successful and many of the recordings from the 1980s were lost.

These and other problems described in the "About the audio" portion of the Oyez website means the majority of our dataset is from 1998-2014. As these are essentially random errors,

[1] https://www.oyez.org/about

we use the full dataset in all of our models. For example, the Justices were unaware that the manufacturer had changed the composition of the reel-to-reel tapes. Similarly, we are also confident the Justices did not change their behavior based on a recording error that was uncovered decades later. Moreover, our results are robust to excluding years where we have a number of missing cases (i.e., 1980s and early 1990s). Indeed, our results remain the same regardless of whether our models are estimated using data from 1981-2014 or 1998-2014. Below we provide additional detail about the data collection process.[2]

Creating Statements from Utterances

Since the Oyez transcripts return a line for every utterance, we first collapsed all the utterances into "statements." For example, if Justice Roberts said, "I believe your argument is incorrect. Your client is guilty," Oyez splits this statement into two utterances. Specifically, they report one line for "I believe your argument is incorrect," and another line for "Your client is guilty." This is due to the nature of the "transcript," which is actually the scrolling text from the Oyez player. An example of this player can be found at this URL:

`https://web.archive.org/web/20150701205416/http://www.oyez.org/cases/2010-2019/2012/2012_12_399/argument`

In this example, the "statements" seem to represent complete thoughts. We contend these are more theoretically interesting. If we were to only include utterances, then we would essentially be chopping up Scalia's comment into something seemingly defined by the Oyez player. This doesn't make any sense. When Scalia interacts with an attorney, he is not only making one long-ish statement, but the "utterances" found on Oyez sometime do not even represent complete paragraphs, meaning they seem to be tied more to the nature of the player and less to any meaningful linguistic break.

Extracting the Audio

Once we obtained "statements," we extracted the audio. This was done using *ffmpeg*, a command line program for audio and video processing. Using the `R` statistical software language, we created a .csv file that can be fed directly into *ffmpeg* using a simple bash script. This script can be found here:

`author_website/code/split_transcript.sh`

Ultimately, this script will return one .wav file for each statement. These were saved in a separate folder called "utterances." This step is, by far, the most time consuming. Even

[2]Our data collection took place between June 1, 2015 and August 1, 2015. Sometime between September 6, 2015 and October 12, 2105 Oyez changed their website dramatically. The new version of the website contains more aggressive web scraping barriers. We provide links to a version of the website captured on July 1, 2015, even though we acknowledge this is not reflective of the current site. We chose this version of the website because it is reflective of the website we worked with in order to collect our data.

using multiple processors on Amazon's Web Services (AWS), splitting the audio files took several weeks. Examples of the resulting output can be found here:

`author_website/audio/2012_11-626_argument_s53.885_s65.257.wav`

`author_website/audio/2012_11-626_argument_s71.494_s81.182.wav`

Extracting the Vocal Pitch

With these .wav files in hand, we used *Praat*[3] to extract vocal pitch.[4] To ensure reliability, we excluded any questions and non-questions that were less than one second. Examples of the resulting output can be found here:

`author_website/pitch/2012_11-626_argument_s53_885_s65_257.txt`

`author_website/pitch/2012_11-626_argument_s71_494_s81_182.txt`

For those interested in extracting vocal pitch using *Praat* in batch, you can find the script we used here:

`author_website/code/extract_pitch.txt`

To use this script, one has to create two directories, "audio" and "pitch." Once these directories are created, place the `.wav` files of interest in the "audio" directory, then open *Praat* and click on "Open Praat Script." More details on how to estimate the fundamental frequency using *Praat* can be found below.

Adding Vocal Pitch to the Transcripts

Using the `R` statistical software language, we added the pitch results to the transcripts, which were collapsed into statements. If we had vocal pitch for a given statement, it is recorded, otherwise there is an NA. Statements without vocal pitch were not included in any of our analyses. An example of the transcripts without vocal pitch can be found here:

`author_website/initial_results/transcript_example.csv`

An example of the final results (including vocal pitch) can be found here:

`author_website/initial_results/results_example.csv`

[3] `http://www.fon.hum.uva.nl/praat`

[4] It is important to note we extract vocal pitch only using voiced speech. Generally speaking, an utterance could be composed of (1) voiced speech, (2) unvoiced speech, and (3) silence. Although there is some debate over whether to use unvoiced speech when estimating the fundamental frequency (for review, see Hess 2007), for the most part scholars tend to only use voice speech. An interesting object of future study would be to examine the predictive power of Justices' silence—for example, whether Justices direct more or fewer pauses to the respondent or petitioner attorneys.

Adding Other Features from Oyez

Once the vocal pitch was added, we added a few features to the results in order to make later calculations possible. First, we added each speaker's "type." For each transcript, Oyez indicates whether the speaker is a "justice," "advocate," "unidentified," or "other." We used these classifications to determine whether a Justice or attorney ("advocate") was speaking. Second, we also added the description for each advocate. This helped us identify the petitioner and respondent. To understand this step, one should reference this URL:

`https://web.archive.org/web/20150701205416/http://www.oyez.org/cases/2010-2019/2012/2012_12_399`

Notice under "Lisa S. Blatt" it says "for the petitioner." Similarly, under "Paul D. Clement" it says "for respondent Guardian ad Litem in support of the petitioner." Using the following `R` code, we added these statements and speaker types to the results:

`author_website/code/get_advocates.R`

Getting Each Justices' Baseline

Next, we had to find each Justices' overall mean and standard deviation. Women typically speak at a higher vocal pitch than men. Given that, we created baseline measures for each Justice using all of the audio from their questions and non-questions. These can be found in Table S1. Using these measures, we re-scaled vocal pitch to standard deviations above and below each Justices' average vocal pitch. For example, Justice Kagan's mean vocal pitch is 171.60Hz with a standard deviation of 28.05Hz. If she asked a question with a vocal pitch of 200Hz, our re-scaled measure would be $\frac{200-171.60}{28.05} = 1.01$, suggesting for that question her vocal pitch was a little over one standard deviation higher than her baseline.

For this step, we had to do some adjustments since Oyez uses multiple names for some Justices. For example, John Roberts is recorded as either "john_g_roberts" or "john_g_roberts_jr." Once these adjustments were made, we created baseline measures using every audio file available for each Justice. To replicate this step, download the following .zip file:

`author_website/code/justice_baseline.zip`

Inside the .zip file, you should find a sample dataset and code. For those who wish to fully replicate this step, the full version of the dataset is available upon request.

Getting the Vocal Pitch Results

With these variables added, we compiled the vocal pitch results necessary for our paper. In essence, we cycle through each line of a transcript and ask whether the statement is longer than a second. When a statement met this criteria, we record whether it contained a question

mark, then we obtained the baseline measure for the Justice speaking. Using these baseline measures, we converted the given vocal pitch into standard deviations above or below the Justices' baseline. Once this is done, we look for an attorney in the lines following the question/statement. If one is found, we determine whether the attorney is a "respondent" or "petitioner," using the information outlined above. This process was repeated for questions and non-questions at both the case and Justice-level. To replicate this step, download the following .zip file:

`author_website/code/pitch_results.zip`

Inside the .zip file, you should find a sample dataset and code. For those who wish to fully replicate this step, the full version of the dataset is available upon request.

Finalizing the Results

With the audio data collected, we added the vocal pitch results to the Justice-level Supreme Court Database. After that, we added measures of Justice ideology. In the next section, we describe this data in more detail. The data needed to replicate Table 1 in the main text can be found here:

`author_website/final_results/justice_results.csv`

Please use the following code to estimate the models we used in the paper:

`author_website/code/final_models.R`

Description of Non-Audio Supreme Court Data

As described above, we linked the audio recordings from Oyez to data on Supreme Court rulings, specifically which party (the petitioner versus the respondent) won or lost each case and how each Justice eventually voted. For this, we relied on the existing Supreme Court Database (Spaeth et al. 2015), which provides case- and Justice-level data on the identities of the petitioner and respondent, what the case was about, and how the Justices voted (for or against the petitioner or respondent). We linked these to the oral arguments audio data by using the docket number of each case.

We have several analyses in which we control or analyze Justice ideology. Although there are a variety of ways to measure ideology on the Supreme Court (Schubert 1965, 1974; Rohde and Spaeth 1976; Segal and Cover 1989), we use scores developed by (Martin and Quinn 2002). The "Martin-Quinn" scores are estimated using dynamic item response theory, which allows Justice ideology to be estimated for every Justice serving from 1937 to 2014. Since their publication, these scores have been the primary way to measure ideology on the Supreme Court, despite a small number of scholars who question their utility in

some instances (Farnsworth 2007). For a reply, please consult (Epstein et al. 2007). The Martin-Quinn scores are a continuous measure from conservative to liberal and have been substantially cross-validated with other measures.

In addition to ideology, we included the controls used by Black et al. (2011). The `Number More Questions Directed at Petitioner` is the number of questions asked of the petitioner minus the number of questions asked of the respondent. `Lower Court Decision Was Conservative` is a binary variable where a 1 indicates the decision of the court being reviewed was ideologically conservative and 0 if it was liberal. Since litigants may benefit from the federal government's support (Segal 1988), we included `Solicitor General as Amicus Supports Petitioner` and `Solicitor General as Amicus Supports Respondent` which returns 1 when the federal government supports the petitioner and respondent, respectively. Similarly, when litigants garner more support from interest groups, they are more likely to win (Collins 2004, 2008; Songer and Sheehan 1993). To control for this effect, we included `Number of Amicus Briefs Supporting Petitioner` and `Number of Amicus Briefs Supporting Respondent` which are the total number of amicus briefs filed on behalf of the petitioner and respondent, respectively. Finally, we used the Supreme Court database party codes to capture the petitioner's and respondent's level of resources. Similar to Black et al. (2011) and Collins (2004, 2008) we placed the petitioners and respondents into the following categories: (1) poor individuals, (2) minorities, (3) individuals, (4) unions or interest groups, (5) small businesses, (6) businesses, (7) corporations, (8) local governments, (9) state governments, and the U.S. government (10). In these categories, we assume poor individuals (1) have the least resources and the U.S. government (10) have the most resources. For more details, please see Black et al. (2011, 576).

Estimating the Fundamental Frequency Using *Praat*

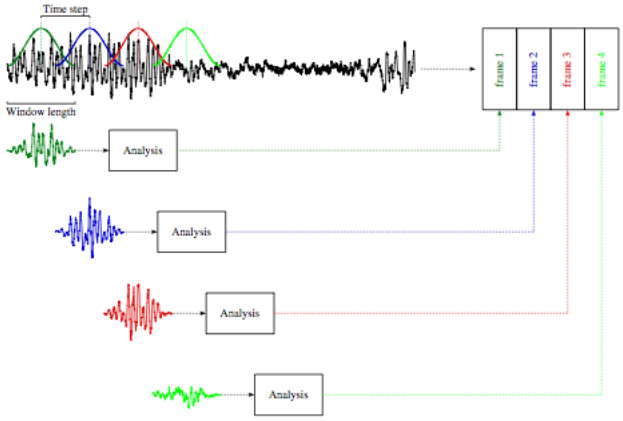
Similar to other algorithms which focus on time-domain periodicity, the Boersma (1993) algorithm estimates the fundamental frequency by dividing the autocorrelation of a windowed signal by the autocorrelation of the window itself. While there are a number of software implementations, we used *Praat*, which is by far the most popular. Indeed, a recent Google Scholar search showed *Praat* had been cited 12,527 times[5], which is substantially more than its next closest competitors *SoundScope* and *WinPitch*, which yielded only 410 and 367 citations, respectively.[6]

However, using *Praat* is not as straightforward as it seems. Specifically, three parameters must be specified (see Weenink 2012, 22-23). First, one has to set the "pitch range," which is simply the "pitch floor" and "pitch ceiling," where the former refers to the lowest frequency

[5]This search can be found here `http://scholar.google.com/scholar?cites=8104635033638065008&as_sdt=40000005&sciodt=0,22&hl=en`.

[6]For a complete list of audio analysis software see `http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html`

Figure S1: Extracting Pitch (borrowed from Weenink 2012, 22)



Praat will consider as a possible candidate for the signal's fundamental frequency, while the latter is the same in terms of the highest frequency (Weenink 2012, 98). While both the pitch floor and pitch ceiling are important, the former is more important than the latter because the former is also used to define the "window length," which is simply the length of the speech segment to be analyzed. By default, *Praat* derives the window length by dividing three seconds (or 3000ms) by the pitch floor, meaning if one were to set the pitch floor to 50Hz the resulting window length would be 60ms.

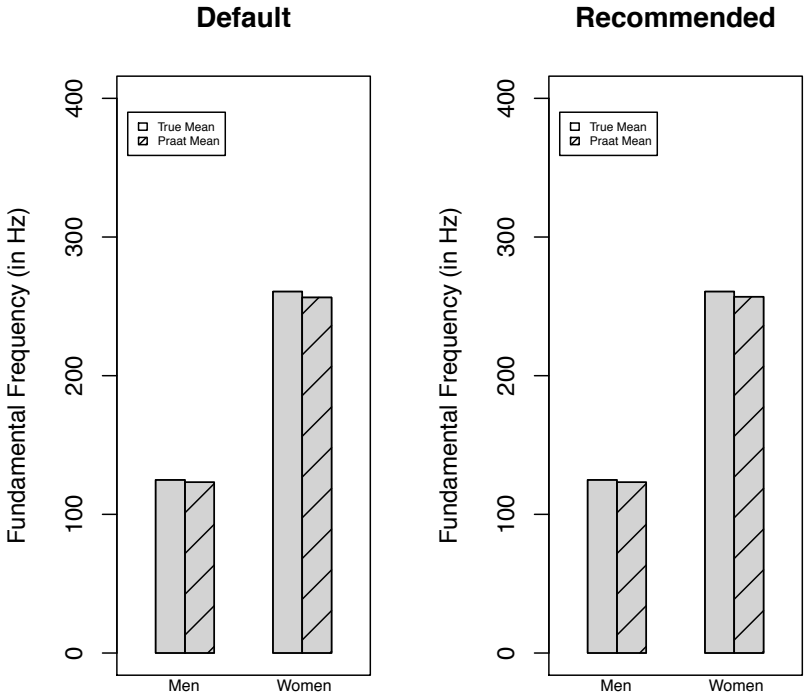
Second, one has to set the "time step" which determines the amount of overlap between successive speech segments. This is shown in Figure S1. Here, in order to analyze a speech signal *Praat* has to "cut up the sound into small segments and analyse each interval separately and pretend it has constant characteristics" (Weenink 2012, 21). This follows directly from the underlying autocorrelation function (ACF) which requires one to assume the signal is stationary within each window, which is why the algorithm divides the audio file into small segments, then takes the average.

Finally, one has to select an appropriate "window shape." This function determines how the segment will be extracted from the speech signal. Generally, one can think of the windowing function as fading the sound in and out, making transitions between speech segments relatively seamless. While one could use any number of windowing functions (Hanning, Bartlett, Guassian, etc.) to achieve this end, the Boersma (1993) algorithm utilizes a Hanning window, which happens to be the *Praat* default.[7]

We conducted a small validation exercise in order to demonstrate *Praat* can extract the fundamental frequency even when the pitch floor and ceiling are misspecified. The data we report is from one randomly selected sentence ("I'd like to leave this in your safe") from

[7]However, it is important to note that Boersma (1993) suggested using a Guassian window could achieve similar results

Figure S2: Estimating the Mean Fundamental Frequency Using *Praat*



Bagshaw, Hiller and Jack (1993)'s data which includes the audio and laryngograph from a man and a woman reading 50 English sentences.[8] The latter is obtained by placing small electrodes on the neck which record the actual vibrations created by the opening and closing of the glottis (Rothenberg 1992). The results are reported in Figure S2.

Here, we show the true mean (clear box) and the *Praat* estimate (stripped box) when one uses the default (first panel) and recommended settings (second panel). Specifically, the *Praat* default sets the pitch floor to 75Hz and the pitch ceiling to 600Hz for all speakers, while the recommended *Praat* settings are $75-300$Hz for men and $100-500$Hz for women.[9] As you can see in Figure S2, when it comes to estimating the mean fundamental frequency there is a negligible difference between the *Praat* default and recommended settings. In fact, if one were to randomly set the pitch floor and ceiling between 50 and 600Hz, then the estimate of the mean fundamental frequency would be insignificantly ($p > .05$) different from the actual mean fundamental frequency 33.97 percent of the time for men and 25.62 percent

[8]This data set is available online at `http://www.cstr.ed.ac.uk/research/projects/fda/`

[9]These suggested ranges can be found at: `http://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html`

of the time for women, meaning you have, at worst, around a 1 in 4 chance of getting an accurate estimate of the mean fundamental frequency.[10]

These results are consistent with Vogel et al. (2009). In this study, the authors tested whether *Praat* returned accurate F_0 estimates when using pitch floors of 50, 70, and, 100Hz and pitch ceilings of 250, 300, 500, 600, and 625Hz for both men and women. Looking to Table 2A in Vogel et al. (2009), we can see for men *Praat* returns an absolute error rate of 8.44Hz when using the default settings, whereas the error rate is 0.91Hz when using the recommended settings. A similar story is found for women. Here, *Praat* returns an absolute error rate of 4.21Hz when using the default settings, whereas the error rate is 5.05Hz when using the recommended settings.[11] Although there is a noticeable difference between the *Praat* default and recommended settings for men, for the most part these results are similar to those outlined above.

We set the pitch floor to 50Hz and the pitch ceiling to 600Hz. This was done because in the *Praat* manual,[12] it says the pitch floor may have to be lowered to 50Hz for "pathological male voices." These pathologies can range from an illness like acute or chronic laryngitis to simply being somewhat hoarse after a speaker either misuses (e.g., talking loudly instead of using a microphone when speaking to a large audience) or overuses (e.g., continuing to talk for long periods even after the voice is strained) their voice. Even though we are fairly certain none of the Justices had laryngitis, we were concerned about the Justices' voices being strained from time to time, ultimately making them sound more hoarse than normal. With that said, Figure S2 shows the mean fundamental frequency can be reliably estimated using *Praat*, even when there is variation in the pitch floor and pitch ceiling.

"For a number of reasons, the task of pitch determination must be counted among the most difficult problems in speech analysis" (Hess 2007, 184). For example, *Praat* allows researchers to specify a "voicing threshold," which determines whether a speech segment is considered "voiced" or "unvoiced." The default voicing threshold is 0.45, meaning any speech segment with a normalized autocorrelation maximum (outside of 0) higher than the 0.45 would be considered "voiced," whereas speech segments where this is not the case would be considered "unvoiced." Unfortunately, *Praat* does not provide an estimate of the fundamental frequency for "unvoiced" speech, arguably because it is does not make sense to do so. Instead, *Praat* returns zeros for the "unvoiced" sections of the pitch contour.

There are essentially two ways to deal with these "unvoiced" sections. Either the re-

[10] In this test we ran *Praat* using every possible combination of pitch floor and pitch ceiling within 50 and 600Hz. This yielded $12,321$ results for both the male and female speaker. Once this was done, we used a simple two-sample t-test (assuming unequal variance and sample size) to compare the means.

[11] Vogel reports the results from three data sources. These results are the averages of all three.

[12] This suggestion can be found here: `http://www.fon.hum.uva.nl/praat/manual/Voice.html`. A similar suggestion is also made here: `https://web.stanford.edu/dept/linguistics/corpora/material/PRAAT_workshop_manual_v421.pdf`

searcher simply disregards the "unvoiced" sections or the researcher interpolates the missing values using the estimated fundamental frequency from the "voiced" sections. *Praat* seems to advocate excluding the "unvoiced" sections which is why the software returns the "voiced" mean when users extract the average fundamental frequency from a pitch object. Unfortunately, there is no definitive source for those interested in understanding more about interpolating the pitch contour. Hess (2007) provides some examples, but for the most part interpolation is used on a case-by-case basis. For example, *Praat* includes several interpolation algorithms with little discussion of when they should be used. Although similar algorithms have been used in a variety of studies, we encourage the reader to use this approach cautiously.

Along these lines, although we are confident *Praat* can consistently estimate the mean fundamental frequency, we encourage readers to use a standardized variable. First, women tend to speak at a higher vocal pitch than men making it is difficult to compare the behavior of male and female speakers unless vocal pitch is standardized. Second, standardization helps account for any measurement error that may be produced when using *Praat*. For example, let's assume Vogel et al. (2009) is correct and *Praat* may return an absolute error rate of 8.44Hz for male speakers, depending on the settings. By subtracting the speaker's mean and dividing by the speaker's standard deviation, this constant is essentially cancelled out – making claims about the relative increase or decrease of vocal pitch more tenable. This is not to say scholars should not use *Praat* to estimate the fundamental frequency, but rather they should do so fully understanding the difficulties associated with pitch determination. We hope our study provides a framework for this future discussion.

Vocal Pitch and Emotional Arousal

According to Posner, Russell and Peterson (2005), reticular formation (RF) is thought to regulate arousal within the limbic system and thalamus (Heilman 2000; Jones 2003). When sensory stimuli are present, emotional arousal is likely relayed to the RF through the amygdalorecticular pathways (Koch and Ebert 1993; Rosen et al. 1991). This broadly increases activity in the cerebral cortex (Heilman, Watson and Valenstein 2011; Jones 2003), which triggers changes in muscle tone and in the sweat glands (Jones 2003), both of which are associated with subjective ratings of emotional arousal (Lang et al. 1993). This increased blood flow to the muscles also causes vocal cords to contract naturally, raising the fundamental frequency (F_0) of one's voice.[13] This is the primary reason why mean F_0 has been routinely been shown to be associated with "affect-related arousal" (Owren and Bachorowski 2007, 47).

[13]We use the following definition of fundamental frequency (Titze 2000):$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}$, where L is the vocal fold length, σ is the longitudinal stress on the vocal folds, and ρ is the vocal fold tissue density. Thus, "voice pitch is inversely proportional to vocal fold length and directly proportional to the square root of tension on the vocal folds" (Puts, Gaulin and Verdolini 2006, 284).

The concept of "arousal" is derived from the circumplex model of affect which posits that all affective states arise from two neurophysiological systems, one related to a pleasure-displeasure continuum (called "valence") and the other related to alertness (called "arousal" or "activation"). According to Russell (2003), at any given moment, one's emotional disposition is a single integral blend of these two dimensions. The horizontal dimension ranges from one extreme (e.g., agony) through a neutral point to its opposite extreme (e.g., ecstasy). But of particular interest is the vertical dimension, which ranges from a deactivated emotional state, such as being sleepy, to an activated emotional state, ultimately culminating in "frenetic excitement" (Russell 2003, 148). This dimension captures one's sense of mobilization and energy—that is, emotional arousal.

"The most consistent association reported in the literature is between arousal and vocal pitch, such that higher levels of arousal have been linked to higher-pitched vocal samples"(Mauss and Robinson 2009, 222). For example, Laukka, Juslin and Bresin (2005) asked actors to portray anger, disgust, fear, happiness, and sadness with "weak" and "strong" intensity. A mixture of amateur and expert judges were then asked the degree to which a speaker was "activated." Not only did they find actors portraying more intense emotions were judged to be more "activated," but they also found "the correlation between activation and valence was virtually zero, which supports the notion that these two dimensions are orthogonal" (643). Perhaps more importantly, the correlation between mean F_0 and a speaker's activation and intensity was 0.62 and 0.72, respectively.[14] Although the authors are quick to point out that activation and intensity are theoretically distinct concepts, they also note "the patterns of vocal cues for activation and emotion intensity showed numerous similarities, which could reflect that they share a similar relation to physiological arousal" (648).

Although the vast majority of studies have validated the use of vocal pitch as a measure of emotional activation using actors, others have attempted to demonstrate the same relationship by inducing specific emotional states. For example, Bachorowski and Owren (1995) asked respondents to complete a 210-trial word identification task on a computer. After each block of 10, the respondents received either positive ("Good Job") or negative ("Try Harder") feedback. After they received the feedback, subjects were asked to say the name of the next block and task aloud. The vocal pitch from these utterances was found to be associated with self-reported levels of emotional intensity, leading the authors to conclude "that vocal pitch can be used to assess the level of emotional arousal currently experienced by the individual" (Mauss and Robinson 2009, 222).

A more recent study conducted by Pisanski, Nowak and Sorokowski (2016) asked under-

[14] Bänziger and Scherer (2005) conducted a very similar study and found essentially the same result. Rather than having actors portray emotions with differing levels of intensity, they asked actors to portray fear, joy, anger, and sadness with either "low" or "high" arousal. Ultimately, they found that the mean fundamental frequency was significantly higher when actors were asked to portray "high" arousal emotions, suggesting not only is vocal pitch highly correlated with emotional arousal, but when asking individuals to portray higher emotional arousal they do so by increasing their vocal pitch.

graduate students to describe their studies and read aloud five sentences from the "Rainbow Passage" two weeks prior to an in-class oral exam. When they did this, the authors not only measured the mean F_0, but they also took saliva samples to cortisol levels – a measure typically associated with stress. They then repeated this procedure immediately before the oral exam when the undergraduate students were thought to be under more stress. Not only did they find "mean F0 and minimum F0 increased significantly under stress" (236), but "cortisol levels increased by an average of 74%" on the day of the exam. Unsurprisingly, they found "20% of the variation in mean F_0 under stress" could be attributed to changing cortisol levels (236). More specifically, "for every unit increase in free cortisol concentrations (1ng/ml), mean F0 increased by 1.1 Hz" – suggesting vocal pitch is highly correlated with the muscles tightening and getting ready for action.[15]

Vocal Pitch and Emotional Leakage

Freud once said "if his lips are silent, he chatters with his fingertips; betrayal oozes out of him at every pore" (Zuckerman and Driver 1985, 129). This is built on the assumption that nonverbal behaviors leak information that people try to hide – meaning nonverbal behavior is involuntary and unintended. However, not all nonverbal channels are alike since some are more controllable than others — in particular the "tone of voice" has often been offered as one channel that is particularly telling. Indeed, "several studies have shown, that like the body, the tone of a person's voice leaks information that is not revealed by the verbal content or facial expressions associated with the message" (Zuckerman and Driver 1985, 129). For these reasons, verbal and nonverbal behavior should be thought of in terms of a "leakage hierarchy" with "verbal content" located in the "controllable end of the continuum, whereas the body and tone of voice may be classified as less controllable and more leaky channels" (Zuckerman and Driver 1985, 130).

This hypothesis has also been extended to "self-presentation" (Elkins et al. 2014, 505). Regardless of the motivation, emotional deception is more deliberate. Whether it is a friend feigning laughter or a politician displaying anger for strategic purposes, more effort is required to convince others of false feelings. Not only do such efforts require more cognitive resources, but the constant thought of whether the fabricated performance is succeeding or failing increases the stress the individual feels. This often causes those involved in emotional deception to become overly concerned with their overt behaviors. For example, a friend trying to feign laughter might inadvertently laugh too much because he or she does not want to be exposed as a fraud. Indeed, "deliberate attempts by liars at controlling expressive behaviors, such as attempts to control thoughts and feelings, can be the seeds of their

[15]Indeed, even though stress and emotional activation are not the same thing, Pisanski, Nowak and Sorokowski (2016) consider the relationship between vocal pitch and stress because the fundamental frequency "is inversely related to the rate of vocal fold vibration, and thus increases as the vocal folds stretch and become tenser or when sub-glottal pressure and vocal intensity increase" (234). This highly suggests the two concepts are closely related.

own destruction" (DePaulo et al. 2003, 78). Consequently, often times when individuals attempt to control their vocal pitch, they may actually sound "more tense and less pleasant or compelling than someone speaking sincerely," which is associated with "increased vocal pitch" (Elkins et al. 2014, 505).

"Positive" and "Negative" Words

We used three dictionaries to assess the degree to which Justices used "positive" and "negative" words during oral arguments. The Linguistic Inquiry Word Count (LIWC) dictionary can be purchased online – `http://liwc.wpengine.com` – for $89. For this study, we used the 2007 version of the dictionary. In total, LIWC includes 407 "positive" emotion words, including their extensions. Here are some examples of the positive words included in the dictionary:

acceptable

benevolent

charming

devoted

elegant

In addition to these "positive" words, LIWC includes 500 "negative" emotion words, including their extensions. Here are some examples of the negative words included in the dictionary:

abusive

brutal

contempt

destructive

envious

The Harvard IV-4 dictionary is publicly available (`http://www.wjh.harvard.edu/~inquirer/homecat.htm`) and contains 1,915 and 2,291 "positive" and "negative" words, respectively. Here are some examples of the positive words included in the dictionary:

adorable

beloved

compassionate

desirable

enjoyable

Some examples of the negative words found in the Harvard IV-4 dictionary can be found here:

abrasive

brutish

condescending

derisive

exterminate

We could not find the Dictionary of Affective Language (DAL) online. We also could not obtain the dictionary from the author Cynthia Whissell. The best we could do is obtain a list of all the words included in the dictionary, regardless of category (pleasantness, activation, or imagery). Once we had the list, we looked for words that appeared in both the DAL and LIWC. Words that appeared in DAL and the LIWC positive emotion words category were said to be "positive." Words that appeared in DAL and the LIWC negative emotion words category were said to be "negative." Here are some examples of the positive words included in the our version of DAL:

admirable

beautiful

cheerful

delighted

enthusiastic

Some examples of the negative words found in our version of DAL can be found here:

alarming

bitterness

cruelty

disgusting

enrage

Undoubtedly, our version of DAL is not ideal, but it is the only version we could obtain. Given that it is a modified version, we provide all the "positive" and "negative" words we used from DAL:

`author_website/dictionaries/dal.csv`

Unfortunately, we can not provide the LIWC categories. However, to give some comparison, we provide all the "positive" and "negative" words we used from the Harvard IV-4 dictionary:

`author_website/dictionaries/harvard.csv`

Descriptive Statistics

Below we report several descriptive statistics. We begin with each Justices' baseline measures. These are reported in Table S1. Of the female Justices, Sandra Day O'Connor had the highest average vocal pitch (196.28Hz) and had the highest standard deviation (31.87Hz). For the male Justices, Thurgood Marshall had the highest average vocal pitch (147.84Hz), whereas Harry Blackmun had the highest standard deviation (46.96Hz).

On average, each question and non-question was 15 and 21 seconds long, respectively. William Burger asked the longest question which was 20 minutes long. The shortest question, asked by Antonin Scalia, was less than a second. For non-questions, Antonin Scalia gave both the longest and shortest. The former was close to four minutes long. The latter was less than a second. For reasons explained below, these incredibly short questions and non-questions were excluded from the analysis, although they do not substantively affect the results.

Ultimately, we found William Brennan spoke the least, followed by Clarence Thomas. In total, Brennan spoke for a little under 3 minutes across 14 questions/non-questions. Clarence Thomas spoke for a little over 11 minutes across 65 question/non-questions. On average, Justices spoke a little over 25 hours across 7,702 question/non-question, suggesting Brennan and Thomas barely spoke.

Antonin Scalia was the most loquacious Justice, asking 12,382 questions and making 15,523 statements. In total, Scalia spoke for 79.82 hours, which is over three times as much as the average Justice. His next closest competitor, Stephen Breyer, spoke close to 7 hours longer than Scalia, but did so using 3,288 fewer questions and 4,880 fewer non-questions. The differences found between Scalia and Brennan should be taken into consideration when evaluating our results. The same can be said for Breyer and Thomas.

Table S2 reports the mean and standard deviation of "Pitch Difference" for each Justice. In this table, positive values imply the Justice generally spoke at a higher vocal pitch towards the petitioner. Unsurprisingly, all but three Justices tended to speak at a higher vocal

Table S1: Average vocal pitch and standard deviation for U.S. Supreme Court Justices, 1982–2014.

Justice	Pitch Mean	Pitch SD	Questions	Non-Questions	Total
Sandra Day O'Connor	196.28	31.87	3,505	3,115	6,620
Elena Kagan	171.60	28.05	1,616	1,363	2,979
Sonia Sotomayor	166.14	25.07	3,345	3,165	6,510
Ruth Bader Ginsburg	163.82	22.54	7,284	8,509	15,793
Thurgood Marshall	147.84	33.75	292	173	465
Antonin Scalia	136.57	31.97	12,382	15,523	27,905
David Souter	135.62	29.43	5,404	5,816	11,220
John Roberts	133.07	28.24	5,261	9,308	14,569
William Brennan	127.48	28.63	5	9	14
Stephen Breyer	126.95	30.62	9,094	10,643	19,737
Harry Blackmun	122.24	46.96	63	31	94
Anthony Kennedy	121.52	22.31	6,208	7,191	13,399
William Rehnquist	121.07	25.25	3,098	5,659	8,757
Warren Burger	119.65	33.49	539	1,279	1,818
Lewis Powell	117.01	36.97	78	42	120
Byron White	116.58	33.42	193	240	433
John Paul Stevens	116.21	24.74	5,505	5,727	11,232
Samuel Alito	112.66	24.67	3,366	1,239	4,605
Clarence Thomas	101.81	24.30	31	34	65
Average	134.43	29.59	3,540	4,161	7,702

Note: Measurements of vocal pitch are in Hertz (Hz). To calculate each we used all of the audio from Justices' questions and non-questions. The number of questions and non-questions are reported in the corresponding columns. In the last column, we report the total number of utterances. Averages of each column are reported at the bottom of the table. The table is sorted by the mean vocal pitch.

Table S2: Average pitch difference and standard deviation for U.S. Supreme Court Justices, 1982–2014.

Justice	Pitch Difference Mean	Pitch Difference SD	Cases
John Roberts	0.05	0.59	180
Anthony Kennedy	0.02	0.49	571
Lewis Powell	0.01	0.18	2
Ruth Bader Ginsburg	−0.01	0.66	746
Stephen Breyer	−0.02	0.54	639
Thurgood Marshall	−0.02	0.58	12
John Paul Stevens	−0.03	0.42	428
David Souter	−0.03	0.50	421
Sonia Sotomayor	−0.04	0.53	225
Antonin Scalia	−0.09	0.57	774
Samuel Alito	−0.09	0.57	146
Elena Kagan	−0.10	0.53	149
Sandra Day O'Connor	−0.10	0.57	313
William Rehnquist	−0.14	0.65	274
Byron White	−0.14	1.34	9
Clarence Thomas	−0.47	–	1
Harry Blackmun	−0.68	1.22	4
Warren Burger	−0.80	1.75	44
Average	−0.15	0.69	246.90

Note: For each Justice, we converted vocal pitch to standard deviations above and below his or her average vocal pitch. The average vocal pitch (standardized) in questions directed towards the petitioner ("Petitioner Pitch") minus the average vocal pitch (standardized) in questions directed towards the respondent ("Respondent Pitch") is captured in "Pitch Difference" (Petitioner Pitch - Respondent Pitch). In the first and second columns, we report the mean and standard deviation of this measure for each Justice. We report the total number of cases in the last column. Averages of each column are reported at the bottom of the table. The table is sorted by the mean "Pitch Difference."

pitch when addressing the respondent. As indicated in the main text, petitioners win the majority of cases, so one would expect Justices to speak at a higher vocal pitch towards the respondents. It is also worth noting most of the standard deviations are all less than one, suggesting we are capturing very small changes in vocal pitch.

We also checked to see if outliers were unduly affecting our results. We estimated three different versions of the models outlined in Table 1 to assuage these concerns. First, we used Cook's distances to identify potential outliers using a $\frac{4}{n}$ cut off. When these observations (66) were identified, we added a dummy variable indicating whether the case was an "outlier" and re-estimated Models (1), (2), (3), and (4). In each of these models "Pitch Difference" was still statistically significant at the 0.0001-level. Second, we re-estimated Models (1), (2), (3), and (4) using robust logistic regression. Even though the `R` statistical software language has two packages that estimate robust logistic regressions, we used the `robust` package. Again, in each of these models "Pitch Difference" was statistically significant at the 0.0001-level. Finally, we re-estimated Models (1), (2), (3), and (4) using standard errors from 10,000 bootstrapped samples. The 95 percent confidence intervals for "Pitch Difference" in each of these models were [-0.15,-0.36], [-0.15,-0.31], [-0.15,-0.31], and [-0.15,-0.31], respectively. None of these confidence intervals overlap zero, suggesting we still find a statistically significant relationship even when the standard errors are bootstrapped.

Individual Justice Analysis

Since some Justices are more active on the bench than others, we also estimated another multilevel logistic regression in which we randomly varied "Pitch Difference" by Justice. Substantively, this model tests whether the predictive power of vocal pitch is restricted to only a handful of Justices. In Table S3, the randomly varying intercept was said to be "significant" ($p < .05$) if a simple ANOVA suggested it explained more of the variance as compared to a simple logistic regression without a randomly varying intercept. Similarly, the randomly varying slope was said to be "significant" ($p < .05$) if it was found to explain more of the variance than a multilevel logistic regression with a randomly varying intercept. Regardless of the model specification, randomly varying "Pitch Difference" did not significantly improve the performance of the model, suggesting the predictive power of vocal pitch does not vary considerably from one Justice to the next.

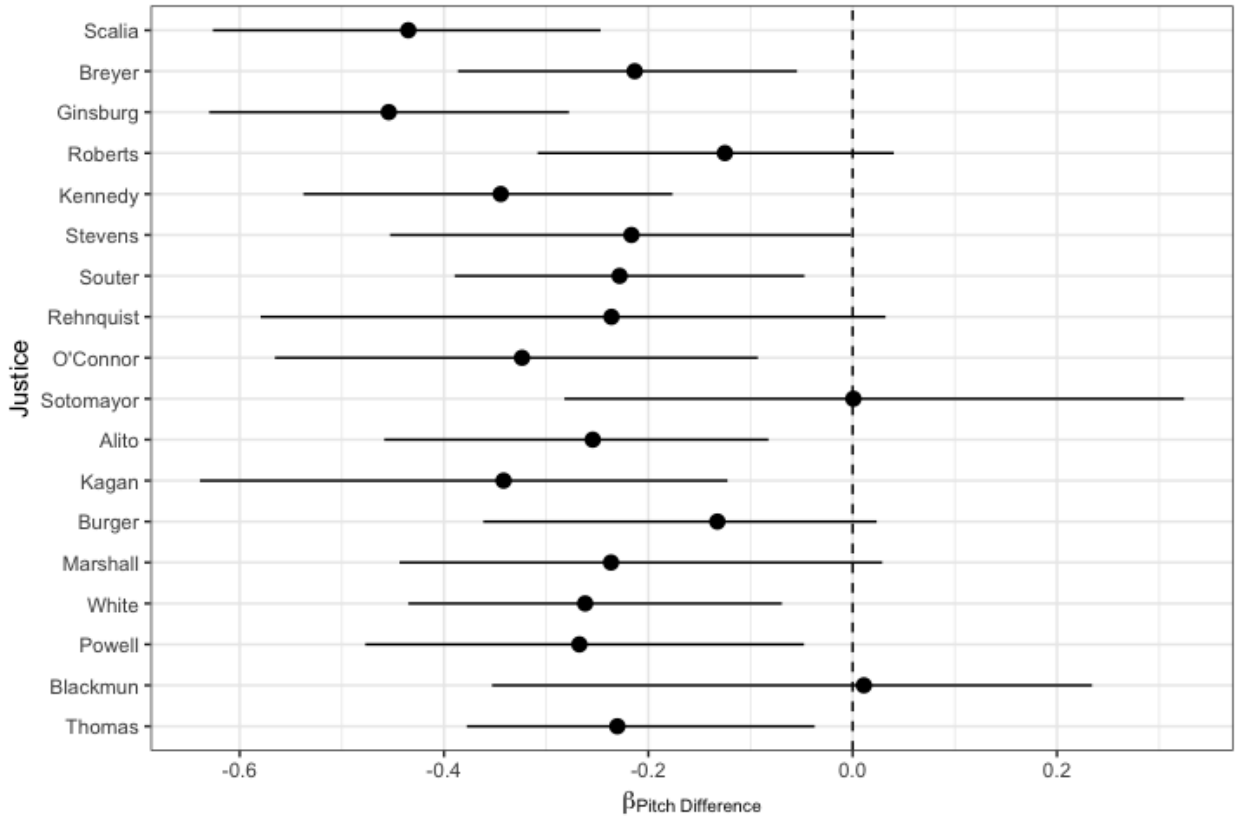
Figure S3 shows bootstrapped 95 percent confidence intervals for the coefficient estimated for each Justice in Model 1. If any of the confidence intervals overlap zero, then it would suggest vocal pitch does not help explain the corresponding Justice's votes. With the exceptions of Justices Sotomayor and Blackmun, all the random coefficients are negative, suggesting that for 16 of 18 Justices an increase in vocal pitch is predictive of a vote against the lawyer the Justice is addressing. Moreover, the confidence intervals for 12 of 16 Justices do not overlap zero, suggesting the predictive power of vocal pitch is statistically significant for the vast majority of Justices. Although not shown here, the random coefficients for Jus-

Table S3: Does Vocal Pitch Predict Votes in Favor of the Petitioner? (Random Slope)

	No Controls	DAL	Harvard IV	LIWC
	(1)	(2)	(3)	(4)
Fixed Effects				
Constant	0.185***	−0.025	−0.027	−0.026
	(0.056)	(0.161)	(0.161)	(0.161)
Pitch Difference	−0.248***	−0.206***	−0.206***	−0.206***
	(0.050)	(0.049)	(0.049)	(0.049)
Percent More Unpleasant Words Directed at Petitioner		−2.011	0.080	−2.179*
		(1.466)	(0.848)	(1.314)
Percent More Pleasant Words Directed at Petitioner		−1.631	0.285	−1.666
		(1.083)	(0.685)	(1.047)
Number More Questions Directed at Petitioner		−0.056***	−0.056***	−0.056***
		(0.008)	(0.008)	(0.008)
Political Ideology$_{t-1}$		0.158***	0.159***	0.158***
		(0.033)	(0.033)	(0.033)
Lower Court Decision Was Conservative		0.013	0.014	0.014
		(0.073)	(0.073)	(0.073)
Political Ideology$_{t-1}$ × Lower Court Decision Was Conservative		−0.263***	−0.263***	−0.263***
		(0.034)	(0.034)	(0.034)
Solicitor General as Amicus Supporting Petitioner		0.543***	0.546***	0.543***
		(0.079)	(0.079)	(0.079)
Solicitor General as Amicus Supporting Respondent		−0.671***	−0.665***	−0.671***
		(0.104)	(0.104)	(0.104)
Number of Amicus Briefs Supporting Petitioner		0.039***	0.039***	0.039***
		(0.008)	(0.008)	(0.008)
Number of Amicus Briefs Supporting Respondent		−0.058***	−0.059***	−0.058***
		(0.007)	(0.007)	(0.007)
Petitioner's Level of Resources		0.045***	0.046***	0.045***
		(0.014)	(0.014)	(0.014)
Respondent's Level of Resources		−0.003	−0.003	−0.003
		(0.014)	(0.014)	(0.014)
Random Effects				
Intercept	0.03***	0.02**	0.02**	0.02**
	(0.16)	(0.13)	(0.12)	(0.12)
Pitch Difference	0.02	0.01	0.01	0.01
	(0.12)	(0.11)	(0.11)	(0.11)
N_1	5,209	4,977	4,977	4,977
N_2	18	18	18	18
$\log L$	−3,549.531	−3,200.175	−3,202.095	−3,199.672
AIC	7,109.061	6,434.350	6,438.189	6,433.344
Percent Correctly Predicted	55.69	63.19	62.97	63.29

Note: Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Models include statements with question marks. The average vocal pitch in questions directed towards the petitioner ("Petitioner Pitch") minus the average vocal pitch in questions directed towards the respondent ("Respondent Pitch") is captured in "Pitch Difference" (Petitioner Pitch - Respondent Pitch). Model 2 uses the Dictionary of Affect in Language (DAL). Model 3 uses the Harvard-IV dictionary. Model 4 uses the Linguistic Inquiry and Word Count (LIWC) dictionary. The rest of the controls are the same as Black et al. (2011). Please refer to the Supplemental Information for more details about each dictionary, the controls, and our cross-validation approach. Levels of significance are reported as follows: *p < .1; **p < .05; ***p < .01.

Figure S3: Random Slopes for Each Justice Using Model 1 from Table S3



Note: This plot shows bootstrapped 95 percent confidence intervals using the coefficients from the first model in Table S3. The bootstrapped samples were created by first randomly sampling (with replacement) from the Justices, then randomly sampling (with replacement) from within each Justice's subset of the data. The dots represent the median of the bootstrapped sample. The lines represent the 95th percentile. If the line overlaps zero (see dashed line), then the corresponding random coefficient is not statistically significant.

tice Roberts, Rehnquist, Burger, and Marshall are statistically significant at the .10-level, suggesting for most Justices vocal pitch carries some statistical weight.

Figure S4 shows the predicted probabilities from the six Justices who appear most in our data. These are the Scalia (27,905 utterances), Stephen Breyer (19,737), Ruth Bader Ginsburg (15,793), frequent "swing vote" Anthony Kennedy (13,399), John Paul Stevens (12,232), and David Souter (11,220). These probabilities are derived using the coefficients from Model 1 in able S3. For all Justices, there is a negative relationship: when the vocal pitch in questions directed to the petitioner is greater (suggesting greater arousal), the petitioner is more likely to lose that Justice's vote. To put this into perspective, Justice Scalia spoke with an average vocal pitch of 136.57Hz (in layman's terms roughly equivalent to slight skepticism),[16] with a standard deviation of 31.97Hz. If he had spoken at 168.54Hz (stronger skepticism)[17] towards the petitioner (one standard deviation above his baseline) and 104.60Hz (neutral tone)[18] towards the respondent (one standard deviation below his baseline), then the predicted probability of the petitioner winning his vote would be 34.02 percent, or 18.60 percentage points lower than would be expected if his vocal pitch was the same towards the petitioner and respondent (52.50 percent). The magnitude of vocal pitch's predictive power is similar for Justice Anthony Kennedy, a frequent swing vote in controversial 5–4 cases. When Justice Kennedy speaks one standard deviation (22.31Hz) higher than his baseline (121.52Hz)[19] towards the petitioner (143.83Hz)[20] and speaks to the respondent one standard deviation below his baseline towards the respondent (99.21Hz),[21] the probability Justice Kennedy votes for the petitioner is 42.77 percent. When Justice Kennedy speaks at the same vocal pitch to both the petitioner and respondent, the probability he votes for petitioner is 15.33 percentage points higher (or 57.90 percent). Regardless of the Justice, vocal pitch is highly predictive of their eventual vote, suggesting even when Justices display subtle emotion, such expressions are highly informative of their underlying predispositions.

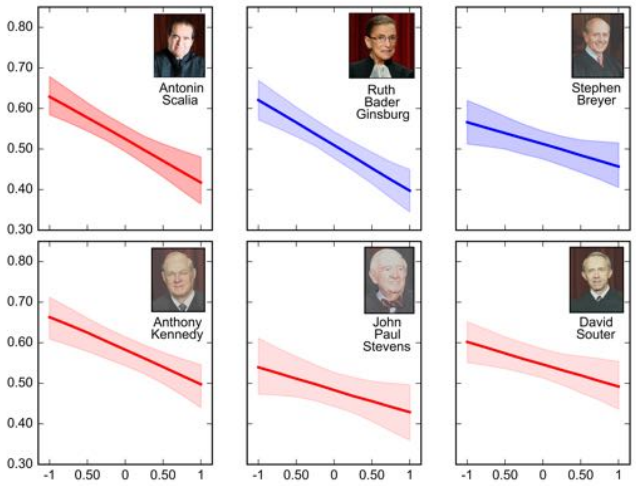
Assessing Predictive Performance

Lack of Guidance

Below we offer three different ways we assess the predictive performance of vocal pitch. In the first, we use the coefficients from Table 1, Model 2 to predict the outcome of the February 7, 2017 case involving the first Trump travel ban. Here, we find not only can vocal pitch (and only vocal) pitch be used to predict each judge's vote, but before the ruling was released we predicted there was a 68.75 percent change Trump would lose. In the second, we compare our model using vocal pitch with the {Marshall}+ algorithm which uses 95 individual- and

[16]For example, `www.project-website.com/audio/scalia_average_pitch.wav`.
[17]For example, `www.project-website.com/audio/scalia_high_pitch.wav`.
[18]For example, `www.project-website.com/audio/scalia_low_pitch.wav`.
[19]For example, `www.project-website.com/audio/kennedy_average_pitch.wav`.
[20]For example, `www.project-website.com/audio/kennedy_high_pitch.wav`.
[21]For example, `www.project-website.com/audio/kennedy_low_pitch.wav`.

Figure S4: Does Vocal Pitch Predict Justices' Votes? (Random Slope)



Note: Plotting predicted probabilities for the six most frequently speaking Justices. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Both fixed and random coefficients are from Model 1 in Table S3. The only independent variable is "Pitch Difference" (the average vocal pitch in questions directed towards the respondent from the average vocal pitch in questions directed towards the petitioner). The x-axis is "Pitch Difference" from -1 to 1. The y-axis is the probability the Justice voted for the petitioner. 95 percent confidence intervals are included. Darker ribbons imply the Justice appeared more frequently in the data. Darker pictures imply the Justice spoke at a higher vocal pitch. Ruth Bader Ginsburg spoke at an average vocal pitch of 163.82Hz. John Paul Stevens spoke at an average vocal pitch of 116.21Hz. The plots are organized from the Justice that appeared in our data the most (Scalia) to the least (Souter). Red indicates the Justice was appointed by a Republican president. Blue indicates the Justice was appointed by a Democratic president.

case-level predictors. While we were able to equal the algorithm's predictive rate for case outcomes, the algorithm outperformed our vocal pitch model by 10.37 percent in terms of Justice votes. Finally, we provided the true positive, true negative, false positive, and false negative rates – comparing each to a variety of previous studies. In some instances, our vocal pitch model does better. In other instances, our vocal pitch model does worse. For us, this begs the question, how well did we do? Said differently, does this evidence suggest vocal pitch matters?

Unfortunately, the previous literature, especially with respect to predicting Supreme Court outcomes, has provided little guidance. Whether it is Black et al. (2011), Katz, Bommarito and Blackman (2014), Knox and Lucas (2017), or Martin et al. (2004) there are a multitude of (often contradictory) ways to determine whether a modeling approach is "successful." For example, Martin et al. (2004) use expert judges as the benchmark for comparison, whereas Knox and Lucas (2017) use other algorithms and software. Conversely, Black et al. (2011) and Katz, Bommarito and Blackman (2014) use "null" models which mean different things to both sets of authors. For the former, a "null" model is choosing the modal category – also known as the "petitioner always wins" rule. In the latter, the authors begin with a lengthy paragraph (and nice figure) explaining why the "petitioner always wins" rule (or in their case "always choose reverse") is too "myopic" (9-10) then spend six paragraphs explaining why using a "null" model with "infinite" and "finite" memory (10-11) is a better approach. This is particularly interesting since Katz, Bommarito and Blackman (2014) also are responsible for the Fantasy SCOTUS website which has both amateur and expert judges predict Supreme Court cases. Not only do the people on their website consistently outperform the {Marshall}+ algorithm, but there is no reference to these amateur and expert judges in Katz, Bommarito and Blackman (2014). Indeed, they actually undercut human judges at two points in the paper when they question the general quality of human predictions (1) and suggest humans overvalue their personal experiences (9).

We do not bring up these inconsistencies to undercut previous work, but rather to suggest we should take a step back and ask what is the goal of assessing predictive performance. For example, Black et al. (2011) introduce predictive accuracy to make the broader argument that emotional expression during oral arguments matters. Similarly, at no point do Martin et al. (2004) ever suggest their algorithm should replace expert analysis. Instead, they offer their model to simply suggest that machine learning algorithms can do reasonably well when predicting Supreme Court votes. Knox and Lucas (2017) use predictive performance to achieve a similar end – they do not want to people to stop using `pyAudioAnalysis`. Rather, they want to emphasize the importance of "audio as data" (like the present paper) while simultaneously providing software in R that is noticeably lacking. A similar argument can be made for Katz, Bommarito and Blackman (2014) who use machine learning to emphasize the importance of predicting Supreme Court outcomes.

In each of these studies, predictive performance is not of value in and of itself. Instead,

predictive performance is used to advance our understanding of the Supreme Court. We think this is actually a good thing. Indeed, social scientists should not be involved in a race to produce a better mouse trap. Although not specific to predictive performance, Rusch and Zeileis (2014) argues a similar point with respect to regression trees. To use their words "the tree literature is highly fragmented." This makes it incredibly difficult to "*choose* the 'right' algorithm for a given problem" (361). Given that there are so many options, many of which yield different answers, scholars often choose to invent "a new algorithm for their specific problem rather than work through different properties of existing algorithms and benchmarking them against each other on their data" This in turn "perpetuates the fragmentation problem" (363). Not only does this drive to "re-invent the wheel" make it difficult for the literature to move forward, but it also makes it difficult to identify commonalities which prevents scholars from better understanding current machine learning algorithms and perhaps more importantly the data.

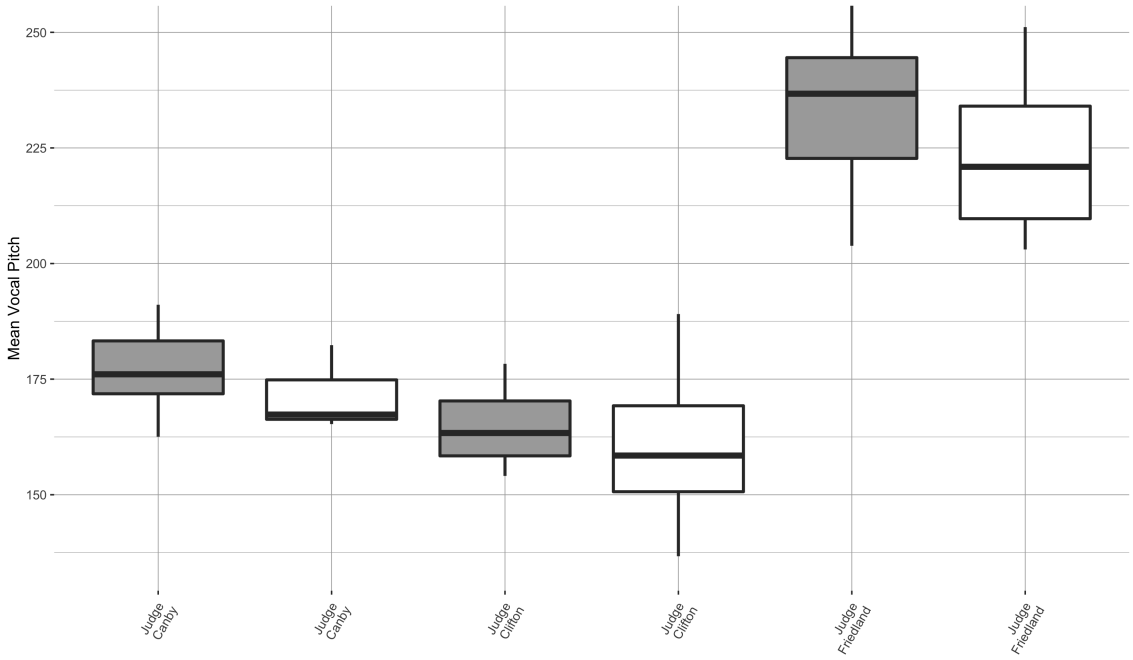
Although the present paper has no aspirations of advancing the machine learning literature, we fundamentally agree with Rusch and Zeileis (2014) assessment. For these reasons, we have went to great lengths to tie our modeling strategy as closely as possible to the one used by Black et al. (2011). Some readers may disagree with our choices, but there is currently no "gold standard" when it comes to assessing predictive performance. Not only do all previous studies use different approaches, but as we explain below they also use a variety of metrics. However, it is important to understand why these previous studies emphasize predictive performance. They do so not because they want to maximize the performance of machine learning algorithms. Rather, predictive performance is used to help us better understand the Supreme Court. In this paper, we do the same. Indeed, as we say in the concluding paragraph of the main text:

> *These comparisons are not meant to suggest that vocal pitch is the only variable scholars should use when assessing emotional expression on the Supreme Court. We argue the {Marshall}+ algorithm, text-based measures, and the "petitioner always wins" rule can (and should) be used to predict Justice votes. However, nonverbal signals, including changes in vocal pitch, also carry considerable weight. Justices choose their words carefully, but have far less control over how those words are spoken – and these subconscious vocal cues, our findings show, carry important information about eventual rulings.*

Ninth Circuit Analysis

When the Trump Administration issued an executive order that banned entry from seven majority-Muslim countries, several legal challenges were immediately filed. Within two weeks, these challenges reached the 9th Circuit, who heard oral arguments on February 7, 2017. More than 130,000 people listened to the arguments, and hundreds of experts weighed in on how the judges would vote. We saw this as a good opportunity to see whether

Figure S5: Average Vocal Pitch from Travel Ban Oral Arguments on February 7, 2017



Note: In this plot gray boxes capture the vocal pitch of questions directed towards the attorney representing the Department of Justice, whereas white boxes capture the vocal pitch of questions directed towards Washington's Attorney General. Moving from left to right, you will find box plots for Judge Canby, Clifton, and Friedland.

vocal pitch could be used to predict cases other than those found on the Supreme Court.

The oral arguments began at 7:09PM EST. Within the hour we had downloaded the one hour and eight minutes of audio from the 9th Circuit's website. Using Audacity,[22] we then split the audio into individual speech acts, focusing exclusively on Judges Canby, Clifton, and Friedland. This process took approximately three and half hours. With the segmented audio in hand, we then extracted the vocal pitch using Praat.[23] At 11:59PM EST, we posted the results shown in Figure S5. Based on the results found in the present paper, we were confident the state of Washington was going to win its case.

First, the vocal pitch was significantly higher ($t = 2.83$, df $= 75$, $p < 0.001$) in questions directed towards the DOJ attorney (203.26Hz) as compared to questions directed towards Washington's Attorney General (180.53Hz). Part of this could be due to Judge Friedland. As mentioned in the main text, women tend to speak at a higher vocal pitch than men, meaning Judge Friedland could naturally raise the vocal pitch of questions directed towards

[22] `http://www.audacityteam.org`
[23] `http://www.fon.hum.uva.nl/praat/`

the DOJ attorney by simply participating more. In the present paper, we deal with this issue by first standardizing vocal pitch to standard deviations above or below a speaker's mean vocal pitch. Unfortunately, we do not have years worth of audio data to create baselines for each judge. With that said, when the same calculation is done using only the male judges we found the same result – vocal pitch was higher in questions directed towards the DOJ attorney (168.99Hz) as compared to questions directed towards Washington's Attorney General (161.78Hz). Even though this was only statistically significant at the .10-level ($t = 1.70$, df = 43, $p < 0.10$), we think this result suggests all judges tended to have an unfavorable disposition towards the DOJ attorney.

Second, we used the models estimated throughout the present paper to try to convert our general impression into a specific estimate. Again, this is difficult because we have on average a little over six minutes of audio from each 9th Circuit judge, whereas we have a little over 25 hours from each Justice. To put this into perspective, we have more audio from Clarence Thomas (11 minutes), than any of the 9th Circuit judges. Despite the small amount of audio, we were able to obtain reasonable justice and case-level predictions. To do so, we did the following:

1. We standardized vocal pitch to standard deviations above or below each judge's baseline. Unfortunately, these baselines were derived from a single case, meaning they are not as precise as the baselines estimated for the Supreme Court.

2. In general, "Pitch Difference" is the average vocal pitch (standardized) directed towards the petitioner minus the average vocal pitch (standardized) directed towards the respondent. In this case, the plaintiff was the state of Washington and the defendant was the United States.[24]

3. With this in mind, "Pitch Difference" is simply the average vocal pitch (standardized) in questions directed towards the Washington Attorney General minus the average vocal pitch (standardized) in questions directed towards the DOJ attorney.

4. For Judges Canby, Clifton, and Friedland this resulted in a "Pitch Difference" of -0.55, -0.32, and -0.29, suggesting each judge spoke at slightly lower (less than 1 standard deviation) vocal pitch towards the Washington Attorney General.

Using the coefficients from Table 1, Model 2 we estimated there was a 58.04, 56.53, and 56.37 percent chance Judges Canby, Clifton, and Friedland would vote *for* the Washington Attorney General. This suggested there was a strong probability of a 3-0 decision ruling against President Trump's travel ban. Our prediction was even stronger at the case-level. Here, the overall "Pitch Difference" was -0.30, suggesting on average the vocal pitch of the judges was about a third of a standard deviation lower towards the Washington Attorney General as compared to the DOJ attorney. When we used this to predict the case outcome using the coefficients from our case-level models outlined below, we estimated there was a

[24]http://cdn.ca9.uscourts.gov/datastore/opinions/2017/02/09/17-35105.pdf

67.85 percent chance the Washington Attorney General would win his case. In a unanimous decision on February 10, the 9th Circuit Court of Appeals ruled for the state of Washington, upholding the stay on President Trump's travel ban. Our results show we predicted this decision three days in advance. We not only correctly estimated the direction, but we also estimated each judge's vote with some degree of certainty.

As explained in the paper, these results are not meant to suggest that vocal pitch is the only variable scholars should use when assessing emotional expression on the Supreme Court. The {Marshall}+ algorithm, text-based measures, and the "petitioner always wins" rule can (and should) be used to predict Justice votes. Our results simply suggest vocal pitch should also be included in this list, because it seems to carry considerable statistical and substantive significance.

Using the {Marshall}+ Algorithm

We chose the {Marshall}+ algorithm as a baseline because they make their results readily available at this URL:

`https://github.com/mjbommar/scotus-predict`

For us, this was extraordinarily useful because it allowed us to get some sense of how well vocal pitch predicted Supreme Court outcomes. In the associated paper (Katz, Bommarito and Blackman 2014), the authors find their algorithm predicts 69.7 percent of cases and 70.9 percent of votes. We do not find the same is true for the cases and votes we considered (see Table S4). Specifically, we took the cases and votes where we had vocal pitch measurements, then found those cases and votes in the {Marshall}+ results file. Once we did that, we counted the number of "correct" predictions as indicated by the authors. These are the results presented in our paper.

With that said, there are a couple of things to note. First, {Marshall}+ does not predict petitioner votes, rather they predict whether the Supreme Court affirmed/reversed the lower court decision. This is not the same thing as what we are predicting in our model. We have no sense of whether this would help or hinder the substantive interpretation of our results. From our experience with this study, we would suspect it is easier to predict the Court's opinion towards the lower court, but this is purely speculative.

Second, to make our predictions comparable to both Katz, Bommarito and Blackman (2014) and Martin et al. (2004) we use the present term as the testing set and use previous terms within the same natural court as the training set. In terms where the natural court changed (2005, 2009, and 2010) this became impossible since there were no previous years to use for training. In these years we used the same term as both the training and testing dataset. If we exclude these terms from Table S4 we correctly predict 65.88 percent of cases, which is 1.63 percentage points better than the {Marshall}+ algorithm (64.25 percent). If we exclude these terms from Table S4 we correctly predict 58.09 percent of votes whereas

Table S4: Comparing Vocal Pitch to the {Marshall}+ Algorithm

(a) Case Outcomes

Term	Correctly predicted using Pitch Difference	Correctly predicted using {Marshall}+	χ^2
1998 (n = 58)	74.14	67.24	0.37 (p < 0.541)
1999 (n = 63)	60.32	65.08	0.14 (p < 0.713)
2000 (n = 65)	70.77	63.08	0.56 (p < 0.456)
2001 (n = 29)	65.52	62.07	0.00 (p < 1.000)
2002 (n = 69)	62.32	65.22	0.03 (p < 0.859)
2003 (n = 55)	74.55	63.64	1.06 (p < 0.302)
2004 (n = 69)	68.12	68.12	0.00 (p < 1.000)
2005 (n = 68)	70.59	66.18	0.14 (p < 0.712)
2006 (n = 54)	64.81	68.52	0.04 (p < 0.838)
2007 (n = 57)	63.16	57.89	0.15 (p < 0.702)
2008 (n = 57)	75.44	82.46	0.48 (p < 0.491)
2009 (n = 39)	76.92	66.67	0.57 (p < 0.450)
2010 (n = 56)	62.50	67.86	0.16 (p < 0.592)
2011 (n = 46)	50.00	52.17	0.00 (p < 1.000)
2012 (n = 55)	58.18	50.91	0.33 (p < 0.566)
Total (n = 840)	66.55	64.76	0.52 (p < 0.472)

(b) Justice Votes

Term	Correctly predicted using Pitch Difference	Correctly predicted using {Marshall}+	χ^2
1998 (n = 324)	53.40	67.90	13.68 (p < 0.000)
1999 (n = 336)	52.98	70.54	21.20 (p < 0.000)
2000 (n = 355)	54.65	67.04	10.93 (p < 0.001)
2001 (n = 124)	62.99	70.16	1.14 (p < 0.285)
2002 (n = 376)	56.65	66.84	7.86 (p < 0.005)
2003 (n = 309)	60.90	69.58	4.78 (p < 0.029)
2004 (n = 365)	59.73	68.22	5.35 (p < 0.021)
2005 (n = 353)	61.89	66.95	1.75 (p < 0.186)
2006 (n = 300)	57.76	72.00	12.80 (p < 0.000)
2007 (n = 305)	54.75	69.51	13.49 (p < 0.000)
2008 (n = 293)	66.89	71.67	1.36 (p < 0.244)
2009 (n = 206)	58.25	62.14	0.50 (p < 0.481)
2010 (n = 298)	53.69	63.42	5.42 (p < 0.020)
2011 (n = 267)	59.18	65.92	2.31 (p < 0.129)
2012 (n = 310)	51.94	64.52	9.58 (p < 0.002)
Total (n = 4521)	57.42	67.79	103.33 (p < 0.000)

Note: Percentage of correctly predicted Supreme Court outcomes. In the panel (a), the outcome is whether the petitioner won the case. In the panel (b), the outcome is whether the Justice voted in favor of petitioner. The second column reports the percentage of outcomes correctly predicted using only "Pitch Difference." The third column reports the percentage of outcomes correctly predicted using the {Marshall}+ algorithm. For each, the testing dataset is the term listed in the table. The training dataset is all preceding terms within the same natural court. The χ^2 statistic tests the proportion of correctly predicted using "Pitch Difference" versus the {Marshall}+ algorithm.

the {Marshall}+ algorithm correctly predicts 68.38 percent. In all four instances, the models improve significantly over chance.

Third, the Katz, Bommarito and Blackman (2014) end their analysis in 2012, so we will exclude 2013 and 2014 from our comparisons. Similarly, Oyez does not provide a lot of audio prior to 1998, so we will begin our analysis in 1998. If we use all the available years in both datasets (1981-2012) we correctly predict 66.55 percent of cases which is 1.79 percent higher than the {Marshall}+ algorithm (64.76 percent). If we use all the available years we correctly predict 57.42 percent of votes whereas the {Marshall}+ algorithm correctly predict 67.79 percent. Although {Marshall}+ significantly outperforms our vocal pitch model when it comes to votes, we equal the algorithm's predictive rate in terms of case outcomes.

Table S5: Controlling for the 30 Best {Marshall}+ Predictors

Model 1	
(Intercept)	13.42
	(9.67)
Pitch Difference	-0.32***
	(0.05)
Case Information	
Law Type	-0.02
	(0.02)
Lower Court Disposition	0.02
	(0.02)
Issue	0.00
	(0.00)
Issue Area	-0.74
	(2.86)
Month Argument	-0.00
	(0.01)
Month Decision	-0.04**
	(0.02)
Petitioner	-0.00
	(0.00)
Petitioner Binned	-0.03*
	(0.01)
Respondent	-0.00
	(0.00)
Respondent Binned	-0.02
	(0.01)

Continued on next page

Table S5 – *Continued from previous page*

Cert Reason	0.02*
	(0.01)
Overall Historic Supreme Court Trends	
Mean Court Direction Issue	0.79*
	(0.33)
Mean Court Direction 10	-0.44
	(0.30)
Lower Court Trends	
Mean Lower Court Direction Issue	0.38
	(0.26)
Current Supreme Court Trends	
Mean Current Court Direction Issue	-0.57*
	(0.29)
Std. Dev. Current Court Direction Circuit Origin	-2.04†
	(1.11)
Std. Dev. Current Court Direction Issue	0.15
	(0.42)
Individual Supreme Court Justice Trends	
Mean Justice Direction	-8.61
	(6.89)
Mean Justice Direction 10	0.06
	(0.31)
Mean Justice Direction Z-Score	33.57
	(27.98)
Difference in Trends	
Difference Court Lower Ct. Direction	-0.35
	(0.23)
Abs. Difference Court Lower Ct. Direction	-3.43†
	(1.98)
Z-Score Abs. Difference Court Lower Ct. Direction	-0.33*
	(0.15)
Difference Justice Court Direction	-0.12
	(5.68)
Difference Justice Court Direction Issue	-0.27
	(0.19)
Z-Score Justice Court Direction Difference	-0.08

Continued on next page

Table S5 – *Continued from previous page*

	(0.56)
Agreement of Justice with Majority 10	1.17***
	(0.28)
N	3223
AIC	4231.16
BIC	4936.22
$\log L$	-1999.58

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

As final check, Table S5 predicts individual Justice votes using vocal pitch and the thirty best predictors form the {Marshall}+ algorithm. We used the thirty most pronounced predictors to make the results more interpretable. The authors (Katz, Bommarito and Blackman 2014) estimate their parameters using regression trees. This makes including all of their variables difficult. Moreover, the authors indicate most of their variables do not dramatically influence the model. In fact, the weights associated with the variables not found in Table S5 are essentially zero. As you can see, our variable ("Pitch Difference") is still highly significant ($p < 0.001$) even when these thirty variables are included as controls. A specification with all 95 variables yields the same result.

As explained in the paper, these results are not meant to suggest that vocal pitch is the only variable scholars should use when assessing emotional expression on the Supreme Court. The {Marshall}+ algorithm, text-based measures, and the "petitioner always wins" rule can (and should) be used to predict Justice votes. Our results simply suggest vocal pitch should also be included in this list, because it seems to carry considerable statistical and substantive significance.

True Positive and False Positive Rate

As explained in the introduction to this section, there is a lack of guidance in the literature about the best way to assess predictive performance. Not only do studies disagree about whether one should use a human or machine benchmark, but there is also a lack of agreement about how predictive performance should be measured.

While all previous studies seem to conclude model accuracy is important, there is little agreement about what measures should also be reported. For example, Knox and Lucas (2017) report the true positive rate, but only for a single category. Katz, Bommarito and Blackman (2014) report the true positive rate (which they call "recall") for the whole model and also report the model's "precision," which is the number of true positives divided by the number of true positives plus the number of false positives. No study predicting anything

Table S6: Assessing the Predictive Performance of Previous Models

(a) Black et al. (2011)

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 0.84	FPR = 0.48
Model = "Respondent"	FNR = 0.16	TNR = 0.52

(b) Katz, Bommarito and Blackman (2014)

	Actual = "Reverse"	Actual = "No Reverse"
Model = "Reverse"	TPR = 0.85	FPR = 0.54
Model = "No Reverse"	FNR = 0.15	TNR = 0.46

(c) Martin et al. (2004)

	Actual = "Reverse"	Actual = "No Reverse"
Model = "Reverse"	TPR = 0.76	FPR = 0.53
Model = "No Reverse"	FNR = 0.24	TNR = 0.47

(d) "Petitioner Always Wins"

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 1	FPR = 1
Model = "Respondent"	FNR = 0	TNR = 0

Note: Each table reports the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for the indicated model. Panel A includes the results from Black et al. (2011). Panel B includes the results from Katz, Bommarito and Blackman (2014). Panel C includes the results from Martin et al. (2004). Finally, Panel D includes the results from the "petitioner always wins" rule.

related to the Supreme Court has reported the whole confusion matrix. Perhaps more troubling, none of these studies reported the true negative (or false positive) rate which is where most models predicting Supreme Court outcomes run into problems.

Whether one is predicting votes or reversal decisions, the outcome variable is unbalanced. In the case of the former, the Supreme Court votes for the petitioner the vast majority of the time. In terms of the latter, the Supreme Court is much more likely to reverse a decision. Preferably, we want to be able to classify the majority and minority classes at the same rate (ideally 100 percent), but in reality we often find with imbalanced data the classifier overperforms for the majority class and under performs for the minority class (for review, see He and Garcia 2009). This problem magnifies as the variable one is predicting becomes increasingly imbalanced. Consequently, "the conventional evaluation practice of using singular assessment criteria, such as the overall accuracy or error rate, does not provide adequate information in the case of imbalanced learning" (He and Garcia 2009, 1264). To demonstrate the basic problem, in Table S6 we report the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for Black et al. (2011), Katz, Bommarito and Blackman (2014), and Martin et al. (2004).

As you can see, Black et al. (2011)'s model is better able to predict votes for the petitioner ("majority" class) as compared to votes for the respondent ("minority" class). In particular, their true positive rate is 84 percent whereas their true negative rate is 52 percent, meaning they overperform for the majority class ("petitioner" votes) and underperform

for the minority class ("respondent" votes). Similar to us, Black et al. (2011) is not a machine learning paper, so we think it is understandable that the authors may not have been too concerned about fully specifying their prediction matrix. With that said, we find it particularly troubling that Katz, Bommarito and Blackman (2014) not only do not fully specify their prediction matrix, but they also chose measures that specifically avoid including the true negative rate. Indeed, Katz, Bommarito and Blackman (2014) report "recall," "precision," and "F1." The latter certainly includes the false positive rate, but it also includes the true positive rate and the false negative rate which are precisely the measures that are inflated when one is using imbalanced data.

When people try to predict anything related to the Supreme Court, one of the first arguments levied is either petitioners win the majority of the cases or the majority of cases are reversed. These criticism are often advanced to undercut predictive models suggesting something to the effect of "67 percent correct is not very impressive because I can just choose the petitioner and be correct 65 percent of the time." Ironically here is where the true negative rate actually becomes an asset. As shown in Panel D of Table S6, the true negative rate for the "petitioner always wins" rule is zero, meaning while the model accuracy of Black et al. (2011), Katz, Bommarito and Blackman (2014), and Martin et al. (2004) may only be slightly better than the "petitioner always wins" rule they substantially outperform this same rule in terms of the true negative rate. If these previous studies would just embrace their true negative rates and explain the inherent problem with imbalanced classifiers, we think the literature would be in a much better place.

More specifically, there are two types of intrinsic imbalance (or imbalance due to the nature of the dataspace). "Absolute rarity" is essentially the problem described by Knox and Lucas (2017). This occurs where the category of interest (in their case "skepticism) is severely underrepresented in the data. "Relative imbalance" occurs when the minority class is not necessarily rare, but is disproportionate to the minority class. This is essentially what exists in both Supreme Court vote and case outcome data. As He and Garcia (2009) suggests, several studies have shown that even when there is relative imbalance, the minority class can be accurately classified. These results demonstrate that the degree of imbalance is not the only thing that hinders the ability to learn the minority class. Instead, "data set complexity is the primary determining factor of classification deterioration, which, in turn, is amplified by the addition of a relative imbalance" (1265).

Thus, increasing model complexity without accounting for the imbalanced nature of the data can at best have little affect on the ability to learn the minority class and at worst can actually decrease the true negative rate. One need not look any further than the performance matrices of Black et al. (2011) and Katz, Bommarito and Blackman (2014). Here, the former has a true positive rate of 84.44 percent whereas the latter has a true positive rate of 84.91 percent. Similarly, Black et al. (2011) has a true negative rate of 52.03 percent whereas Katz, Bommarito and Blackman (2014) has a true negative rate of 54.26 percent.

Table S7: Assessing the Predictive Performance of Vocal Pitch

(a) No Controls

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 0.88	FPR = 0.83
Model = "Respondent"	FNR = 0.12	TNR = 0.17

(b) With Controls (DAL)

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 0.76	FPR = 0.53
Model = "Respondent"	FNR = 0.24	TNR = 0.47

(c) With Controls (Harvard-IV)

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 0.76	FPR = 0.53
Model = "Respondent"	FNR = 0.24	TNR = 0.47

(d) With Controls (LIWC)

	Actual = "Petitioner"	Actual = "Respondent"
Model = "Petitioner"	TPR = 0.76	FPR = 0.52
Model = "Respondent"	FNR = 0.24	TNR = 0.48

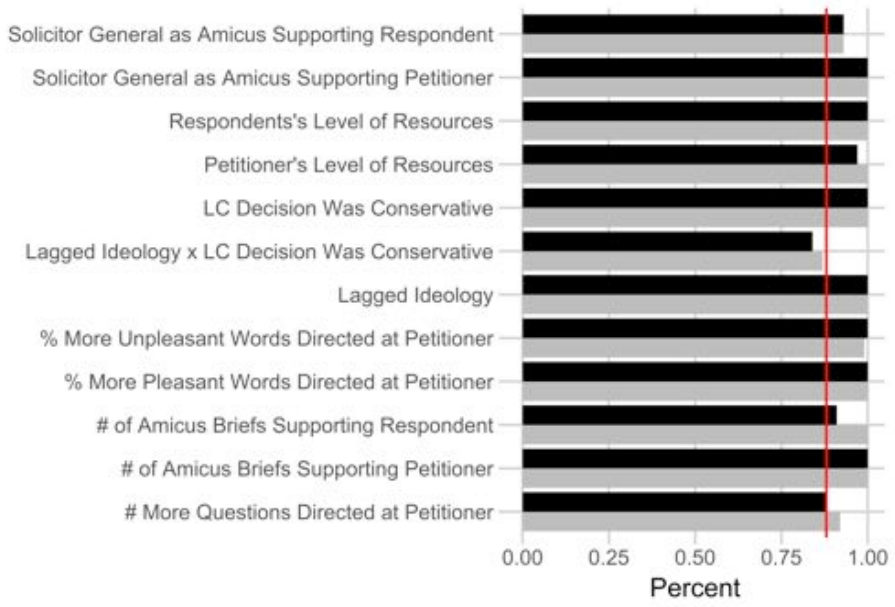
Note: Each table reports the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for the indicated model. Panel A includes the results from a model where vocal pitch is the only predictor (see Table 1, Model 2). Panel B adds the Black et al. (2011) controls to this base model (see Table 1, Model 3). Panels C and D are the same, but the Harvard-IV (see Table 1, Model 4) and LIWC (see Table 1, Model 5) dictionaries are used to create the text-based measures, respectively.

The reason why these rates are so similar is because simply adding additional parameters to the model does little to help to predict the minority class unless the model takes steps to adjust the imbalance. Thus, for the present paper we chose to estimate more reduced models, similar to those reported by Black et al. (2011).

In Table S7, we report the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for each of the models we estimated in Table 1. As you can see, our results are nearly identical to the model used by Martin et al. (2004). Their best model returns a true positive rate of 0.76 and a true negative rate of 0.47, whereas our best model (Model 5) returns a true positive rate of 0.76 and a true negative rate of 0.48. To assess the degree to which vocal pitch helps improve model performance, we ask the reader to look to Table S7, Panel A. Here, we report the performance matrix for a model in which vocal pitch is included as the only predictor (see Table 1, Model 2). It is readily apparent we have difficulty predicting respondent votes. Indeed, while a true negative rate of 0.17 is a noticeable improvement over the "petitioner always wins" model, it does not seem not very impressive. However, this is primarily because the other variables we used in our more robust models are not held to a similar standard. More specifically, if each of these variables were used as the only predictor, would the true negative rate improve?

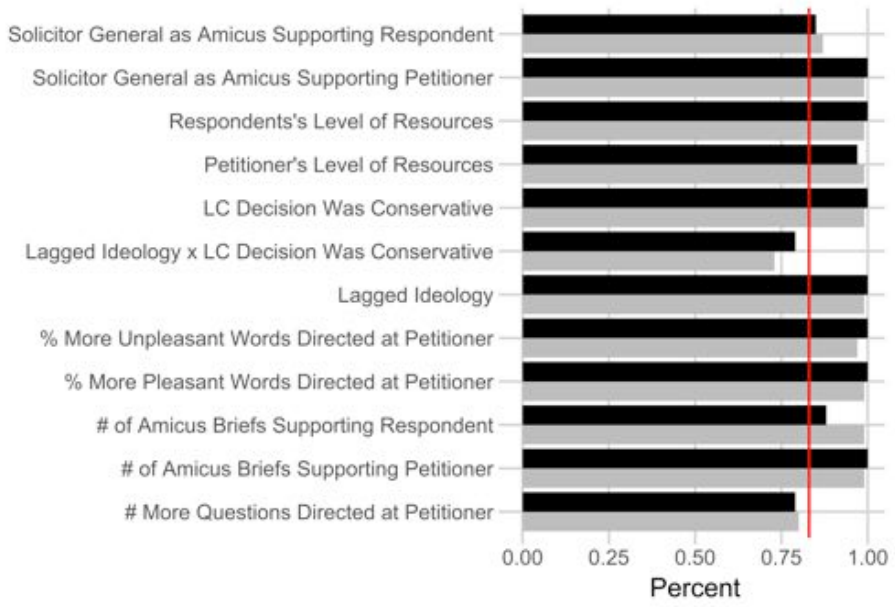
In Figures S6 – S9, we report the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for models in which each of the variables reported

Figure S6: Black et al. (2011) True Positive Rate



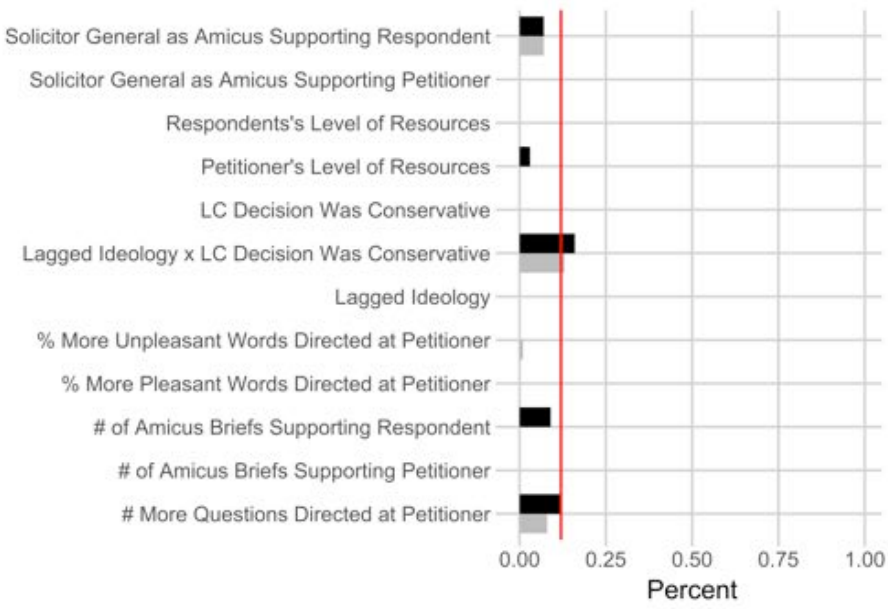
Note: In this plot we plot the true positive rate for each variable used by Black et al. (2011). Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Outside of the random intercept, the only predictor included in these models is the variable labeled in the y-axis. Gray boxes indicate the models were estimated using Black et al. (2011)'s data. Black boxes indicate the models were estimated using the data we collected. The red line is the true positive rate of a model only using vocal pitch. This number can be found in the top left of Table S7, Panel A.

Figure S7: Black et al. (2011) False Positive Rate



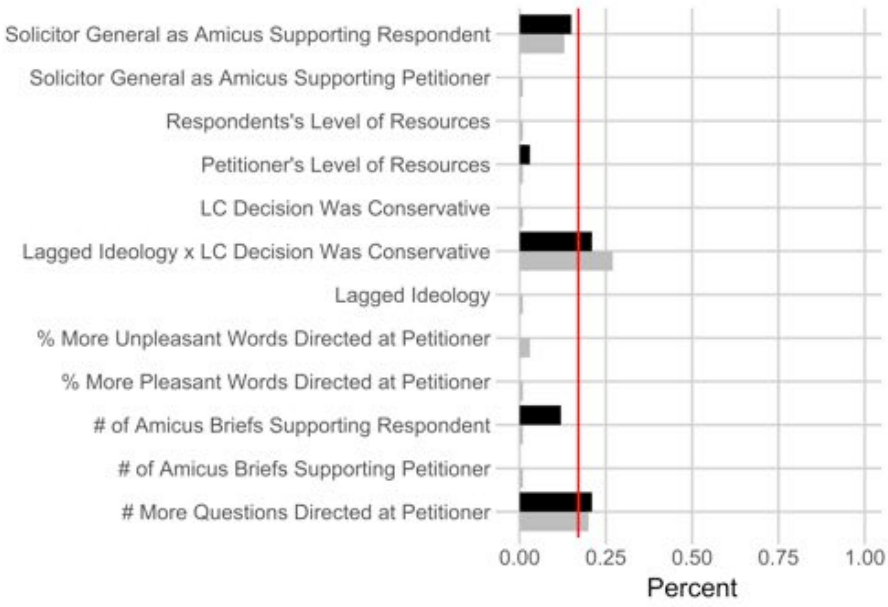
Note: In this plot we plot the false positive rate for each variable used by Black et al. (2011). Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Outside of the random intercept, the only predictor included in these models is the variable labeled in the y-axis. Gray boxes indicate the models were estimated using Black et al. (2011)'s data. Black boxes indicate the models were estimated using the data we collected. The red line is the false positive rate of a model only using vocal pitch. This number can be found in the top right of Table S7, Panel A.

Figure S8: Black et al. (2011) False Negative Rate



Note: In this plot we plot the false negative rate for each variable used by Black et al. (2011). Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Outside of the random intercept, the only predictor included in these models is the variable labeled in the y-axis. Gray boxes indicate the models were estimated using Black et al. (2011)'s data. Black boxes indicate the models were estimated using the data we collected. The red line is the false negative rate of a model only using vocal pitch. This number can be found in the bottom left of Table S7, Panel A.

Figure S9: Black et al. (2011) True Negative Rate



Note: In this plot we plot the true negative rate for each variable used by Black et al. (2011). Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Outside of the random intercept, the only predictor included in these models is the variable labeled in the y-axis. Gray boxes indicate the models were estimated using Black et al. (2011)'s data. Black boxes indicate the models were estimated using the data we collected. The red line is the true negative rate of a model only using vocal pitch. This number can be found in the bottom right of Table S7, Panel A.

by Black et al. (2011) is used as the only predictor similar to Table 1, Model 2. Here, we estimated these models using our own data (black bars) and the data used in the original study (gray bars). The red line represents the performance of a model in which vocal pitch is the only predictor. For the specific number used to produce the line, please refer to the appropriate cell in Panel A of Table S7.

Although we encourage the reader to refer to each figure, we are going to focus on Figure S9. This figure reports the true negative rate which is where our model only using vocal pitch did the worst. Here, it is apparent our true negative rate is actually pretty good relative to the Black et al. (2011) variables. Indeed, we out perform every variable except for the interaction term and the number of questions directed at the petitioner. Perhaps more importantly, we do substantially better than any of the text-based measures. For example, when our data is used and the only predictor is the percent of unpleasant words, the true negative rate is the same as the "petitioner always wins" rule, meaning the model produces the same confusion matrix. The same can be said when our data is used and the only predictor is the percent of pleasant words. This again returns the same confusion matrix as the "petitioner always wins" rule, suggesting these text-based measures have little predictive value.

We found the same results using Black et al. (2011)'s data. In terms of the percent of unpleasant words, the true negative rate is 0.03 when the authors' data is used. Similarly, the true negative rate improves to 0.01 when the authors' data is used and the only predictor is the percent of pleasant words. However, even these "high water" marks fall well short of the true negative rate returned when vocal pitch is used as the only predictor. The authors may contend these are unreasonable tests because the text-based measures should both be used in the same model. Ironically, when this is done the true negative rate actually gets worse when using the authors' data. Indeed, when both variables are used the true negative rate drops to zero in both models suggesting that text-based measures have little to no predictive value when included by themselves or when they are both included in the same model.

Collectively, these results are important because the accuracy of Table 1, Model 2 may seem unimpressive. However, the fact that we significantly improve model accuracy at all over the "petitioner always wins" rule using a single variable is actually quite impressive. There are only two variables used by Black et al. (2011) that can say the same: the interaction term and the number of questions. We think this is partially due to the substantive value of vocal pitch, but also because of the nature of the dependent variable. Attempting to learn from a variable where the majority and minority class is relatively imbalanced is incredibly difficult which is why previous studies and the present study have such difficulty outperforming the "petitioner always wins" rule while simultaneously predicting the minority class.

The question becomes, what are we supposed to do? Said differently, how are judicial

scholars supposed to predict Supreme Court votes with unbalanced data? We refer readers to He and Garcia (2009) for a detailed discussion of sampling and other algorithmic solutions that can be employed, but for us two initial steps should be taken. First, as He and Garcia (2009) suggests, political scientists need to have standardized evaluation metrics. Whether this is using human judges versus other algorithms or using model accuracy versus the true positive rate, previous studies considering the Supreme Court have all used different evaluation procedures. Although there is no "gold standard," we suggest future predictive papers should report the (1) true positive rate, (2) false positive rate, (3) true negative rate, and (4) false negative rate for all models, including models in which the variable of interest is included as the only predictor.

Second, we need to be clear about what we are trying to achieve by reporting predictive performance. If the goal is to incrementally improve machine learning algorithms, then more care needs to be taken when using imbalanced data. However, in the social sciences, this is generally not the goal. Instead, the goal is to learn whether something like vocal pitch has substantive as well as statistical value. In these instances, predictive performance should be used in combination with a variety of metrics, all designed to demonstrate the variable of interest matters. We should not race to build a better mousetrap. Instead, we should use machine learning to better our understanding of the Supreme Court. Even though the present paper is not a machine learning paper, we essentially say this much in the concluding paragraph of the main text:

> *These comparisons are not meant to suggest that vocal pitch is the only variable scholars should use when assessing emotional expression on the Supreme Court. We argue the {Marshall}+ algorithm, text-based measures, and the "petitioner always wins" rule can (and should) be used to predict Justice votes. However, non-verbal signals, including changes in vocal pitch, also carry considerable weight. Justices choose their words carefully, but have far less control over how those words are spoken – and these subconscious vocal cues, our findings show, carry important information about eventual rulings.*

Additional Robustness Checks

Below we report several additional robustness checks. Table S8 re-estimates the models outlined in Table 1 while including the number of words directed towards the petitioner and respondent. Since our vocal pitch measure could theoretically be influenced by irregular speech patters, we included these controls because such irregularities are more likely to occur in longer statements. It is readily apparent the relationship between vocal pitch and the Justice votes is unaffected when these variables are included. Please also notice that the variable controlling for the number of questions is no longer statistically significant. We included this variable in our original models precisely to address concerns associated with utterance length. The fact that this variable is no longer statistically significant when the

Table S8: Does Vocal Pitch Predict Votes in Favor of the Petitioner?

	No Controls	DAL	Harvard IV	LIWC
	(1)	(2)	(3)	(4)
Fixed Effects				
Constant	0.178***	−0.025	−0.027	−0.026
	(0.055)	(0.160)	(0.160)	(0.160)
Pitch Difference	−0.266***	−0.215***	−0.216***	−0.216***
	(0.036)	(0.038)	(0.038)	(0.038)
Percent More Unpleasant Words Directed at Petitioner		−1.924	0.189	−2.020
		(1.476)	(0.851)	(1.315)
Percent More Pleasant Words Directed at Petitioner		−1.542	0.324	−1.581
		(1.093)	(0.690)	(1.054)
Number More Questions Directed at Petitioner		0.008	0.009	0.008
		(0.012)	(0.012)	(0.012)
Political Ideology$_{t-1}$		0.151***	0.151***	0.151***
		(0.031)	(0.031)	(0.031)
Lower Court Decision Was Conservative		0.004	0.005	0.005
		(0.073)	(0.073)	(0.073)
Political Ideology$_{t-1}$ × Lower Court Decision Was Conservative		−0.247***	−0.247***	−0.247***
		(0.034)	(0.034)	(0.034)
Solicitor General as Amicus Supporting Petitioner		0.538***	0.540***	0.538***
		(0.080)	(0.080)	(0.080)
Solicitor General as Amicus Supporting Respondent		−0.676***	−0.670***	−0.676***
		(0.104)	(0.104)	(0.104)
Number of Amicus Briefs Supporting Petitioner		0.039***	0.039***	0.039***
		(0.008)	(0.008)	(0.008)
Number of Amicus Briefs Supporting Respondent		−0.057***	−0.057***	−0.057***
		(0.007)	(0.007)	(0.007)
Petitioner's Level of Resources		0.043***	0.043***	0.043***
		(0.014)	(0.014)	(0.014)
Respondent's Level of Resources		−0.003	−0.003	−0.003
		(0.014)	(0.014)	(0.014)
Words Directed Towards Petitioner		0.150***	0.150***	0.150***
		(0.031)	(0.031)	(0.031)
Words Directed Towards Respondent		−0.240***	−0.241***	−0.240***
		(0.031)	(0.031)	(0.031)
Random Effects				
Intercept	0.03***	0.01**	0.01**	0.01**
	(0.16)	(0.11)	(0.11)	(0.11)
N_1	5,209	4,977	4,977	4,977
N_2	18	18	18	18
$\log L$	−3,551.721	−3,171.066	−3,172.714	−3,170.663
AIC	7,109.441	6,376.132	6,379.429	6,375.326

Note: Each model is a multilevel logistic regression with a random intercept for each Justice. Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Models include statements with question marks. The average vocal pitch in questions directed towards the petitioner ("Petitioner Pitch") minus the average vocal pitch in questions directed towards the respondent ("Respondent Pitch") is captured in "Pitch Difference" (Petitioner Pitch - Respondent Pitch). Model 2 uses the Dictionary of Affect in Language (DAL). Model 3 uses the Harvard-IV dictionary. Model 4 uses the Linguistic Inquiry and Word Count (LIWC) dictionary. The rest of the controls are the same as Black et al. (2011) except we included the total number of words directed at the petitioner and respondent. Given that these new variables are highly correlated with the number of questions asked, we had difficulty getting these models to converge. Ultimately, this required we divide the total number of words by one hundred. Levels of significance are reported as follows: *p < .1; **p < .05; ***p < .01.

total number of words is included in the model gives us some degree of confidence that we are adequately controlling for this issue in our original model specification. Indeed, we actually had a lot of difficulties getting these models to converge precisely because utterance length is highly correlated with the number of questions and the variables associated with the number of pleasant and unpleasant words.

We also disaggregated our data into questions and non-questions. The results from these models can be found in Table S9. Beginning with Panel A, we show when one only uses questions the results are identical, regardless of whether one uses questions directed only at the petitioner (see Model 1) or questions directed only at the respondent (see Model 2). In fact, Model 3 shows when both are included in the same model you find both variables are highly significant ($p < 0.001$). In this model, the coefficients suggest when Justices speak at a higher vocal pitch towards the petitioner the petitioner is more likely to lose, whereas the same is true for questions directed towards the respondent.

Panel B replicates the petitioner results, but does not replicate the results for the respondent. In these models, we only include statements. These results emphasize the importance of questions when it comes to emotional activation. When Justices are asking questions they are actively trying to acquire new information. Mood congruent processing becomes much more likely during this process, suggesting Justices are more actively sorting information using their priors. The lack of a statistically significant relationship for non-questions further underlines this argument, even though consistent predictions are still found for questions directed towards the petitioner.

From this point, both Panels C and D demonstrate when questions and non-questions are considered simultaneously our results are extremely robust. In fact, every model yields highly significant results regardless of the specification. Beginning with Panel C, one finds essentially the same results as Panel A. When Justices raise their vocal pitch towards petitioners, the petitioners are more likely to lose (see Model 1). The same can be said for respondents (see Model 2), and each of these relationships hold when both are included in the same model (see Model 3). Panel D shows these results also hold even when Justice ideology is included as a control.

References

Bachorowski, Jo-Anne and Michael J. Owren. 1995. "Vocal Expression of Emotion: Acoustic Properties of Speech are Associated with Emotional Intensity and Context." *Psychological Science* 6(4):219–224. 11

Bagshaw, Paul C., S.M. Hiller and M.A. Jack. 1993. "Enhanced Pitch Tracking And The Processing Of F0 Contours For Computer Aided Intonation Teaching." Presented at the 3rd Annual European Conference on Speech Communication and Technology. 8

Table S9: Considering Petitioner Pitch and Respondent Pitch Separately

(a) Questions

	Petitioner Only	Respondent Only	Petitioner and Respondent
	(1)	(2)	(3)
Constant	0.13***	0.24***	0.16***
	(0.02)	(0.03)	(0.03)
Pet. Pitch	−0.16***		−0.28***
	(0.03)		(0.04)
Res. Pitch		0.11***	0.25***
		(0.03)	(0.04)
N	7,036	6,554	5,098
Log Likelihood	−4,845.04	−4,492.75	−3,484.95
AIC	9,694.08	8,989.50	6,975.90

(b) Non-Questions

	Petitioner Only	Respondent Only	Petitioner and Respondent
	(1)	(2)	(3)
Constant	0.11***	0.26***	0.20***
	(0.02)	(0.02)	(0.03)
Pet. Pitch	−0.12***		−0.15***
	(0.03)		(0.03)
Res. Pitch		0.02	0.04
		(0.02)	(0.03)
N	7,324	6,544	5,003
Log Likelihood	−5,055.31	−4,481.05	−3,430.79
AIC	10,114.62	8,966.09	6,867.57

(c) Both

	Petitioner Only	Respondent Only	Petitioner and Respondent
	(1)	(2)	(3)
Constant	0.12***	0.25***	0.18***
	(0.02)	(0.02)	(0.02)
Pet. Pitch	−0.14***		−0.19***
	(0.02)		(0.03)
Res. Pitch		0.05***	0.11***
		(0.02)	(0.02)
N	14,360	13,098	10,101
Log Likelihood	−9,900.95	−8,977.17	−6,926.01
AIC	19,805.91	17,958.33	13,858.01

(d) Both with Controls

	Petitioner Only	Respondent Only	Petitioner and Respondent
	(1)	(2)	(3)
Constant	0.23***	0.30***	0.24***
	(0.02)	(0.02)	(0.03)
Pet. Pitch	−0.13***		−0.18***
	(0.02)		(0.03)
Res. Pitch		0.05***	0.11***
		(0.02)	(0.02)
Lib. Petitioner	−0.25***	−0.14***	−0.16***
	(0.03)	(0.04)	(0.04)
Ideology	0.14***	0.18***	0.16***
	(0.01)	(0.01)	(0.02)
Lib. Petitioner × Ideology	−0.17***	−0.29***	−0.23***
	(0.02)	(0.02)	(0.02)
N	14,360	13,098	10,101
Log Likelihood	−9,804.93	−8,852.79	−6,854.41
AIC	19,619.86	17,715.58	13,720.82

Note: Outcome is whether the Justice voted in favor of petitioner. Unit of analysis is each Justice's vote. Models in panel A includes only statements with question marks ("Question" = 1). Models in panel B include only statements without question marks ("Question" = 0). Models in panels C and D include both, with models in the latter controlling for ideology. The average vocal pitch directed towards the petitioner ("Pet. Pitch") and the average vocal pitch directed towards the respondent ("Res. Pitch") are included as predictors. The column names indicate which combinations are used. Levels of significance are reported as follows: $^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$. Standard errors are reported in parentheses. All models include Justice fixed effects.

Bänziger, Tanja and Klaus R Scherer. 2005. "The Role of Intonation in Emotional Expressions." *Speech Communication* 46(3):252–267. 11

Black, Ryan C., Sarah A. Treul, Timothy Johnson and Jerry Goldman. 2011. "Emotions, Oral Arguments, and Supreme Court Decision Making." *The Journal of Politics* 73(2):572–581. 6, 23, 24, 32, 33, 34, 35, 36, 37, 38, 39

Boersma, Paul. 1993. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound." *Proceedings of the Institute of Phonetic Sciences* 17:97–110. 6, 7

Collins, Paul. 2004. "Friends of the Court: Examining the Influence of Amicus Curiae Participation in U.S. Supreme Court Litigation." *Law and Society Review* 38(4):807–832. 6

Collins, Paul. 2008. *Friends of the Supreme Court: Interest Groups and Judicial Decision Making*. Oxford University Press. 6

DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton and Harris Cooper. 2003. "Cues to Deception." *Psychological Bulletin* 129(1):74–118. 13

Elkins, Aaron, Stefanos Zafeiriou, Maja Pantic and Judee Burgoon. 2014. Unobtrusive Deception Detection. In *The Oxford Handbook of Affective Computing*, ed. Rafael A. Calvo, Sidney D'Mello, Johnathan Gratch and Arvid Kappas. New York, NY: Oxford University Press chapter 38, pp. 503–515. 12, 13

Epstein, Lee, Andrew D Martin, Kevin M Quinn and Jeffrey A Segal. 2007. "Ideological Drift Among Supreme Court Justices: Who, When, and How Important." *Northwestern University Law Review* 101(4):1483. 6

Farnsworth, Ward. 2007. "The Use and Limits of Martin-Quinn Scores to Assess Supreme Court Justices, with Special Attention to the Problem of Ideological Drift." *Northwestern University Law Review* 101(4):1891. 6

He, Haibo and Edwardo A Garcia. 2009. "Learning From Imbalanced Data." *IEEE Transactions on Knowledge and Data engineering* 21(9):1263–1284. 32, 33, 40

Heilman, Kenneth M. 2000. Emotional Experience: A neurological Model. In *Cognitive Neuroscience of Emotion*, ed. Richard D. Lane and Lynn Nadel. New York, NY: Oxford University Press. 10

Heilman, Kenneth M., Robert T. Watson and Edward Valenstein. 2011. Neglect and Related Disorders. In *Clinical Neuropsychology*, ed. Kenneth M. Heilman and Edward Valenstein. New York, NY: Oxford University Press pp. 296–348. 10

Hess, Wolfgang J. 2007. Pitch and Voicing Determination of Speech with an Extension Toward Music Signal. In *Springer Handbook of Speech Processing*, ed. Jacob Benesty, M. Mohan Sondhi and Yiteng Huang. 1st ed. Springer chapter 10, pp. 181–211. 3, 9, 10

Jones, Barbara E. 2003. "Arousal Systems." *Frontiers in Bioscience* 8:438–451. 10

Katz, Daniel Martin, Michael James Bommarito and Josh Blackman. 2014. "Predicting the Behavior of the Supreme Court of the United States: A General Approach." *Available at SSRN 2463244* . 23, 27, 29, 31, 32, 33

Knox, Dean and Christopher Lucas. 2017. "A General Approach to Classifying Mode of Speech: The Speaker-Affect Model for Audio Data." Unpublished manuscript. 23, 31, 33

Koch, Michael and Ulrich Ebert. 1993. "Enhancement of the Acoustic Startle Response by Stimulation of an Excitatory Pathway from the Central Amygdala/Basal Nucleus of Meynert to the Pontine Reticular Formation." *Experimental Brain Research* 93(2):231–241. 10

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley and Alfons O. Hamm. 1993. "Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions." *Psychophysiology* 30(3):261–273. 10

Laukka, Petri, Patrik Juslin and Roberto Bresin. 2005. "A Dimensional Approach to Vocal Expression of Emotion." *Cognition & Emotion* 19(5):633–653. 11

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis* 10(2):134–153. 5

Martin, Andrew D., Kevin M. Quinn, Theodore W. Ruger and Pauline T. Kim. 2004. "Competing Approaches to Predicting Supreme Court Decision Making." *Perspectives on Politics* 2(4):761–767. 23, 27, 32, 33, 34

Mauss, Iris B. and Michael D. Robinson. 2009. "Measures of Emotion: A Review." *Cognition and Emotion* 23(2):209–237. 11

Owren, Michael J. and Jo-Anne Bachorowski. 2007. Measuring Emotion-Related Vocal Acoustics. In *Handbook of Emotion Elicitation and Assessment*, ed. James A. Coan and John J. B. Allen. New York, NY: Oxford University Press pp. 239–265. 10

Pisanski, Katarzyna, Judyta Nowak and Piotr Sorokowski. 2016. "Individual Differences in Cortisol Stress Response Predict Increases in Voice Pitch During Exam Stress." *Physiology & Behavior* 163:234–238. 11, 12

Posner, Jonathan, James A. Russell and Bradley S. Peterson. 2005. "The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive development, and Psychopathology." *Development and Psychopathology* 1y:715–734. 10

Puts, David Andrew, Steven J.C. Gaulin and Katherine Verdolini. 2006. "Dominance and the Evolution of Sexual Dimorphism in Human Voice Pitch." *Evolution and Human Behavior* 27:283–296. 10

Rohde, David W. and Harold J. Spaeth. 1976. *Supreme Court Decision Making*. San Francisco, CA: W. H. Freeman. 5

Rosen, Jeffery B., Janice M. Hitchcock, Catherine B. Sananes, Mindy J. D. Miserendino and Michael Davis. 1991. "A Direct Projection from the Central Nucleus of the Amygdala to the Acoustic Startle Pathway: Anterograde and Retrograde Tracing Studies." *Behavioral Neuroscience* 105(6):817–825. 10

Rothenberg, Martin. 1992. "A Multichannel Electroglottograph." *Journal of Voice* 6(1):36–43. 8

Rusch, Thomas and Achim Zeileis. 2014. "Discussion on Fifty Years of Classification and Regression Trees." *International Statistical Review* 82(3):361–367. 24

Russell, James A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110:145–172. 11

Schubert, Glendon. 1965. *The Judicial Mind: The Attitudes and Ideologies of Supreme Court Justices, 1946-1963*. Evanston, IL: Northwestern University Press. 5

Schubert, Glendon. 1974. *The Judicial Mind Revisited: Psychometric Analysis of Supreme Court Ideology*. London, UK: Oxford University Press. 5

Segal, Jeffery A. and Albert D. Cover. 1989. "Ideological Values and the Votes of U.S. Supreme Court Justices." *American Political Science Review* 83:557–565. 5

Segal, Jeffrey A. 1988. "Amicus Curiae Briefs by the Solicitor General During the Warren and Burger Courts: A Research Note." *The Western Political Quarterly* 41(1):135–144. 6

Songer, Donald R. and Reginald S. Sheehan. 1993. "Interest Group Success in the Courts: Amicus Participation in the Supreme Court." *Political Research Quarterly* 46(2):339–354. 6

Spaeth, Harold J., Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger and Sara C. Benesh. 2015. "Supreme Court Database, Version 2015, Release 01." `http://www.http://supremecourtdatabase.org`. 5

Titze, Ingo R. 2000. *Principles of Voice Production*. Iowa City, IA: National Center for Voice and Speech. 10

Vogel, Adam P., Paul Maruff, Peter J. Snyder and James C. Mundt. 2009. "Standardization of Pitch Range Settings in Voice Acoustic Analysis." *Behavior Research Methods* 41(2):318–324. 9, 10

Weenink, David. 2012. "Speech Signal Processing with Praat." Available online at `http://www.fon.hum.uva.nl/david/sspbook/sspbook.pdf`. 6, 7

Zuckerman, Miron and Robert E. Driver. 1985. Telling Lies: Verbal and Nonverbal Correlates of Deception. In *Multichannel Integrations of Nonverbal Behavior*, ed. Aaron W. Siegman and Stanley Feldstein. New York, NY: Lawrence Erlbaum chapter 3, pp. 129–147. 12