Online Supplement
Emotions, Oral Arguments, and Supreme Court Decision Making*

Ryan C. Black
rcblack@msu.edu
Michigan State University

Sarah A. Treul
streul@unc.edu
University of North Carolina at Chapel Hill

Timothy R. Johnson
trj@umn.edu
University of Minnesota

Jerry Goldman
jgoldman@oyez.org
Northwestern University

Contents

*Published in the *Journal of Politics*, volume XX, number YY (MONTH YEAR), pg. PP-TT. Replication data are available (ADD URL HERE).

1 Whissell's Dictionary of Affect in Language (DAL)

1.1 DAL User's Guide

The follow pages come from the technical manual and user's guide to Whissell's Dictionary of Affect in Language, which we downloaded from the Human Development Consulting website (`http://www.hdcus.com/manuals/wdalman.pdf`) on March 2, 2010. Note that we did not use the HDC program to perform our content analysis (pages 6-7) but rather wrote our own basic computer code using the DAL's word lists.

Whissell's Dictionary of Affect in Language
Technical Manual and User's Guide

Cynthia Whissell
Laurentian University

The new Dictionary of Affect in Language (called DAL or Dictionary for short) is an instrument designed to measure the emotional meaning of words and texts. It does this by comparing individual words to a word list of 8742 words which have been rated by people for their activation, evaluation, and imagery.

Dictionary Construction

WORD SELECTION

Words included in the DAL were selected in an ecologically valid manner. There were three steps involved in the selection.

STEP 1: The Kucera and Francis 1969 corpus of 1,000,000 words was sampled from print media in the early 1960's. Words from this corpus with frequencies greater than 10 which also appeared in more than one subsample were included in the DAL list. This insured that the starting words in the set would not be rare ones, or ones specific to one type of print source. Proper names were removed from the sample.

STEP 2: The word set was then compared to four text samples generated by individuals rather than media. It was also compared to a large sample from juvenile literature. Unique words found in these sources were added to the list.

All of the samples employed at this step had been gathered by researchers at Laurentian University:

1. Students' retelling of a story, 16309 words (source: Terri-Lynn Dittburner, Dr. M. Persinger)
2. Interviews on the topic of abuse, 6085 words (source: Carolyn Djaferis)
3. Adolescents' descriptions of their emotions, 15929 words (source: Louise Wood)
4. University students' essays, 14807 words (source: Katie Lemega)
5. Juvenile fiction of the 50's, 60's, 70's, 80's, and 90's, 82865 words (source: Michael Dewson and Laurie Steven)

STEP 3: The DAL list which contained approximately 8700 words at the end of step 2 was tested on 16 new, blindly selected, samples. It was also tested on a corpus of 350,000 words of English text collected by Whissell from many sources. The DAL demonstrated a hit rate or matching rate of approximately 90%.
The hit rate of 90% meant that one would expect NINE OUT OF TEN WORDS IN MOST ENGLISH TEXTS to be matched by the DAL.
The 90% hit rate compares very favourably with that of the original DAL (15%-25%).

WORD RATING

Data Collection Period

Data were collected in the latter half of the 1990's.

Volunteers

Volunteers for the rating task were mostly university students. Both men and women participated. A very small number of volunteers (less than 5%) were paid. Many volunteers received experimental participation credits. A total of over 200 volunteers was involved.

Rating Dimensions

The words of the DAL list were rated along the dimensions of PLEASANTNESS, ACTIVATION, and IMAGERY. In each case the scale used was a three-point scale.

(1) Unpleasant	(2) In between	(3) Pleasant
(1) Passive	(2) In between	(3) Active
(1) Hard to imaging	(2) In between	(3) Easy to imagine

Method

Roughly 50% of the ratings for Pleasantness and Activation were gathered using a computer-administered task. The remaining 50% and all ratings for Imagery were gathered in a paper and pencil task.

Different volunteers rated different numbers of words, and some rated words along more than one dimension. Occasionally volunteers returned to be retested on a second set of words.

Most volunteers were able to make about 200 rating judgments before showing signs of boredom, inattention, or fatigue (the task was self-paced and could be terminated).

The DATA used to create the DAL involved MORE THAN 186,000 DIFFERENT RATING JUDGMENTS ABOUT WORDS. Each word was rated for Activation and Pleasantness an average of 8 times and for Imagery 5 times.

THE 8742 WORDS IN THE DAL LIST

Descriptive Statistics

	Mean	SD	Skewness	Kurtosis
Pleasantness	1.84	.44	.27	-.37
Activation	1.85	.39	.39	-.29
Imagery	1.94	.63	.18	-1.18

Correlations

Correlations among rating dimensions were extremely weak, and shared variance among scales was minimal (<1%).

	Pleasantness	Activation	Imagery
Pleasantness	1.0	.10	.06
Activation	.10	1.0	.05
Imagery	.06	.05	1.0

POPULATION VALUES FOR TEXT SAMPLES

Not all DAL words are used with equal frequency. As a result, population parameters for text samples are somewhat different than DAL parameters for individual words. The broad corpus of 350,000 words of English text collected by Whissell was used as the normative text corpus. Descriptive statistics were calculated for this corpus.

	Mean	**SD**
Pleasantness	1.85	.36
Activation	1.67	.36
Imagery	1.54	.63

In addition, the text corpus was evaluated for the appearance of certain classes of extreme words.

IN TEXTS,

Very Pleasant Words (in the top 10% of all rated words) occurred .060 of the time(sd=.24).
Fun or Cheerful Words (in the top 25% of all rated words for Pleasantness and in the top 25% for Activation) occurred .049 of the time (sd=.215).
Very Active Words (top 10% of all rated words) occurred .042 of the time (sd=.201).
Nasty Words (top 25% for Activation, bottom 25% for Pleasantness) occurred .032 of the time (sd=.175).
Very Unpleasant words (bottom 10%) occurred .038 of the time (sd=.191).
Very Sad words (bottom 10% for both Activation and Pleasantness) occurred .052 of the time (sd=.221).
Very Passive words (bottom 10% for Activation) occurred .195 of the time (sd=.396).
Nice or Soft words (top 25% for Pleasantness, bottom 25% for Activation) occurred .046 of the time (sd=.21).
Highly Imaged words (top 10%) occurred .045 of the time (sd=.209) and
Poorly Imaged words (bottom 10%) .399 of the time (sd=.490).

Differences in word usage patterns explain the fact that DAL list means for Activation and Imagery are higher than text means. The more frequent use of more passive and more poorly imaged words brings the text mean down lower than the list mean (1.84 to 1.67 for Activation, 1.93 to 1.53 for Imagery). There were weak negative correlations between the frequency of use for individual words and Activation (-.06) and Imagery (-.08).

RELIABILITY AND VALIDITY

Reliability Information

Full DAL ratings for Pleasantness and Activation were correlated with subsamples of ratings.

Subsample	N	Pleasantness r	Activation r
1	2746	.68	.55
2	2118	.72	.61
3	8740	.87	.86
4	8466	.81	.69

Validity Information

The new and original DAL share 2165 words. The correlation for Activation of these words was .45, that for Pleasantness .71.
The new DAL shares 1556 words with the Children's DAL. The correlation for Activation was .51, that for Pleasantness .62.
The new DAL shares 1703 words with a list of nouns which had been rated for imagery. The correlation was .47.

POSSIBLE USES OF THE DICTIONARY

The DAL could be used to:

1. Select words for learning, memory, and cognitive experiments.
2. Score texts of many different kinds including descriptions of subjective feelings, essays, newspaper reports, poetry, literature.
3. Score responses in a free-association task, or other projective task.
4. Score memory for stories for emotional memory.
5. Score song lyrics for emotional content.

LANGUAGE is not the most frequent human behaviour (heartbeat and breathing exceed it in number of occurrences per lifetime!) but it is VERY COMMON. **ANY sample of language gathered in ANY manner can be scored for its emotionality and imagery using the Dictionary of Affect in Language**.

Oscar Wilde's poem ballad of Reading Goal was written in 1896, and was dedicated (in memoriam) to a member of the Royal Horse Guards who died there (Reading Goal was a prison).

The poem ends with the lines:
And all men kill the thing they love,
By all let this be heard,

Some do it with a bitter look,
Some with a flattering word,
The coward does it with a kiss,
The brave man with a sword!

HDC Program Output for The Ballad of Reading Goal

Mean Pleasantness 1.80
Mean Activation 1.65
Mean Imagery 1.60
-Known Adult Words 3605
Child Pleasantness 4.22
Child Activation 3.99
-Known Child Words 1019
Total words 4188
Mean Frequency 2479.2
Sentence Length 32.98
Sentences (periods) 26
Exclamation Marks 26
Question Marks 7
% Nice 2.75
% Pleasant 5.05
% Fun 3.74
% Active 3.08
% Nasty 3.52
% Unpleasant 6.05
% Sad 4.94
% Passive 19.39
% High Imagery 6.24
% Low Imagery 45.8

From the data here it is obvious that the Ballad is somewhat less pleasant (1.80 vs 1.85), better imaged (1.60 vs 1.53), and marginally less active (1.65 vs 1.67) than the normative text corpus. 3605 of the 4188 words in the poem were matched, giving the DAL a hit rate of 86%. Values for the emotional dimensions and categories could all be compared to the text corpus using z tests and the means and standard deviations reported earlier for the corpus.

For example, the standardized score for pleasantness would be calculated as

z = (Sample mean - Corpus mean) / Corpus sd
or 1.80-1.85/.36
or .14

The z test testing the null of equality of sample and corpus means would involve additionally dividing the standard error by the square root of the number of scored words, namely

(1.80 - 1.85) / .36 / sqr(3605)
or 8.4 (p<.01)

USING THE HDC SCORING PROGRAM

1. Download the Whissell Dictionary of Affect from the HDC pages.
2. Run setup.
3. Run the .exe file or double click the icon.
4. Clear the screen
5. Type in any desired text OR
6. Drag in any highlighted material from a WORD or WORD PERFECT program OR
7. Cut and Paste material from such programs OR
8. Name a file which is an ASCII or DOS file for analysis.
9. To analyze more than 1 samples at a time, separate each sample by entering aaa between samples

You will receive two output files with a name chosen by yourself, one is a summary file and will look like the Reading Goal example, the other is a data file and will list each and every scored word with all information for that particular word.

1.2 Example Words Figure

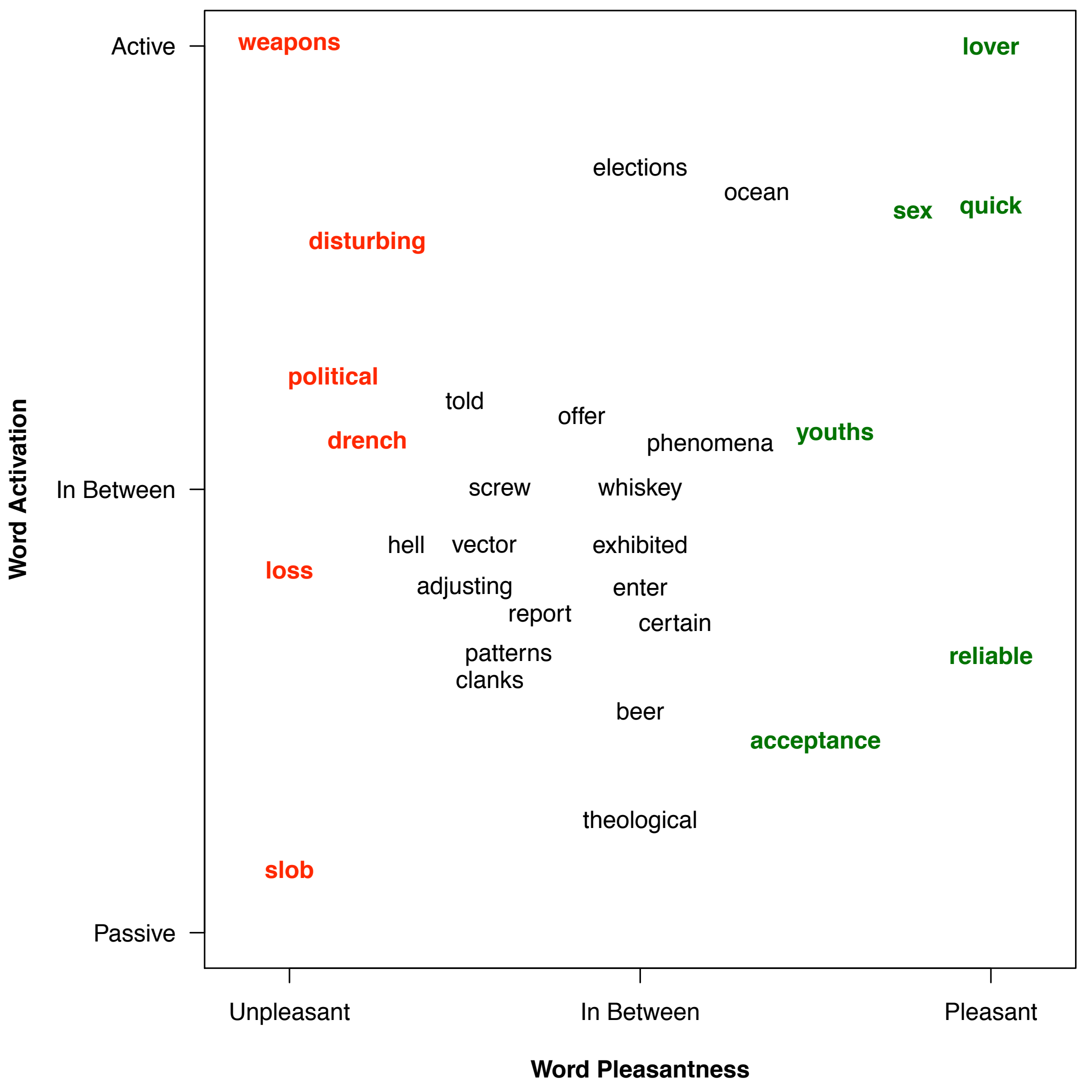

Figure 1: Word pleasantness (x-axis) and activation (y-axis) for a representative sample of words taken from Whissell's DAL. Words displayed in red font are "unpleasant words," which fall in the 10th percentile (or lower) of pleasantness. Words displayed in green font are "pleasant words," which fall in the 90th percentile (or higher) of pleasantness.

2 Ideology and Emotional Language

As we note in the main text (see footnote #14), one potential concern is that our emotion-based variables are simply acting as proxies for the Court's (or a justice's) ideological stance towards a side. To evaluate the extent to which this argument is accurate we conducted a series of auxiliary regressions to explore the relationship between our emotional variables and the variables we use to tap into ideological agreement with a given side. Each column corresponds to parameter estimates for regressing the ideology variables on each of our four independent variables of interest. Parameter estimates are ordinary least squares with robust standard errors reported in parentheses (* denotes $p < 0.05$).

	Court Outcome		Justice Votes	
	Unpleasant	Pleasant	Unpleasant	Pleasant
Political Ideology	0.040	0.076	-0.085*	-0.058*
	(0.102)	(0.090)	(0.029)	(0.024)
Lower Court Conservative	-0.034	0.117	-0.133	0.094
	(0.093)	(0.085)	(0.114)	(0.109)
Ideology x Lower Conservative	0.193	-0.013	0.277*	0.168*
	(0.155)	(0.143)	(0.043)	(0.038)
Constant	-0.122*	-0.118*	-0.055	-0.025
	(0.061)	(0.052)	(0.075)	(0.070)
Observations	2996	2996	3042	3042
Adjusted R2	0.00	0.00	0.01	0.00
Root MSE	1.11	0.99	3.01	2.90

To interpret these results we then calculate predicted values using the SPost series of commands implemented in Stata by Long and Freese (2006). The table below presents a substantive snapshot of the results for the Court Outcome regression where the unpleasant words variable is the dependent variable. Recall that positive values of this correspond to relatively more harsh language being directed at the petitioner and negative values indicate more abuse being heaped on the respondent. Cell entries represent the predicted value of unpleasant words, conditional on the median's ideology and the lower court direction.

Table 1: Court Outcome Unpleasant Words Predicted Values

	Lower Conservative	Lower Liberal	Difference (and 95% C.I.)
Median Conservative	0.08	-0.08	0.16 [-0.01, 0.33]
Median Liberal	-0.17	-0.12	-0.04 [-0.24, 0.15]

When the median is conservative and the lower court decision is also conservative, we find that the Court appears to direct more unpleasant language towards the petitioner than when the lower court decision is liberal (i.e., $0.08 > -0.08$). This is ideologically intuitive since the petitioner in the former instance is asking the Court to reverse a decision that the median would presumably like. Similarly, it appears as though the Court is more negative to the petitioner when the median justice is liberal and the petitioner is asking the Court to reverse a liberal (as opposed to a conservative) lower court decision (i.e., $-0.12 > -0.17$).

Despite these apparent differences in point estimates, we are ultimately unconcerned for several reasons. First, we note that in both instances the within row differences are not statistically significant at the conventional 95 percent level. What is more, the magnitude of the average estimated effect is exceedingly small. For the unpleasant words variable a single standard deviation is equal to 1.1 percent. In other words, the (nearly statistically significant) effect observed when the median is conservative amounts to less than 15 percent of a single standard deviation in the variable of interest. The magnitude of the potential effect when the median is liberal is even smaller (less than 4 percent of a standard deviation). Finally, taking the model as a whole, a very small amount of the overall variation in our emotional variables can be explained by these ideology measures. Indeed, the adjusted R2 for each of the four models never exceeds 0.01 (i.e., each model explains less than one percent of the variation in the emotional variables).

Repeating this process for our pleasant words variable and then both emotional variables in the justice votes model yields roughly similar results, which we present in the three tables below. Note that the differences in our justice votes tables—Table 3 and Table 4—are larger than those based on the Court outcome modes reported in Table 1 and Table 2. Despite this increase in the difference, the relative magnitude is still quite small. The 1.08 difference found for the most conservative justice in Table 3 is only slightly larger than one third of a single standard deviation in the unpleasant words variable (which has a larger range at the justice vote level than it does at the Court outcome level).

Taken together, we ultimately find very little evidence to believe that ideology exerts a large substantive influence on the emotional language used by the Court and its members during oral arguments.

Table 2: Court Outcome Pleasant Words Predicted Values

	Lower Conservative	Lower Liberal	Difference (and 95% C.I.)
Median Conservative	0.06	-0.04	0.10 [-0.05, 0.25]
Median Liberal	-0.12	-0.00	0.12 [-0.06, 0.30]

Table 3: Justice Votes Unpleasant Words Predicted Values

	Lower Conservative	Lower Liberal	Difference (and 95% C.I.)
Justice Conservative	0.65	-0.43	1.08* [0.71, 1.46]
Justice Liberal	-0.67	0.16	-0.82* [-1.17, -0.48]

Table 4: Justice Votes Pleasant Words Predicted Values

	Lower Conservative	Lower Liberal	Difference (and 95% C.I.)
Justice Conservative	0.55	-0.28	0.83* [0.49, 1.17]
Justice Liberal	-0.21	0.12	-0.33* [-0.64, -0.01]

3 Result Robustness and Specific Justices

We note in footnote #16 that our results are not being driven by a single justice. At the Court Outcome level, we re-estimated our model 11 times, each time excluding a particular natural court, which refers to every unique combination of justices that sit together.[1] The unpleasant words variable is always negative and statistically significant at the $p < 0.05$ level (two-tailed test). We describe the natural court results for our pleasant words variable in the next section below.

At the Justice Vote level we conducted a similar analysis, re-estimating the model a total of 11 times, each time excluding an individual justice. The coefficient is always negative and, with three exceptions, statistically significant at the $p < 0.05$ level (two-tailed test). While the coefficient is still negative for three justices, the two-tailed p-value is above the 0.05 threshold. These excluded justices and their p-values (in parentheses) are: Alito (0.07), Breyer (0.051), and Souter (0.051). Following the advice of others (see, e.g., Gelman and Stern 2006, Gill 1998),[2] we do not read much into the "non-significance" of these results but rather believe they demonstrate that our results are not unduly sensitive to a single justice.

4 Time-Boundedness of Pleasant Words

In the Court Outcome model we find that as the relative amount of pleasant words directed at the petitioner increase, so too does the Court's probability of voting to reverse the lower court decision (i.e., ruling in favor of the petitioner). In the Justice Votes model, by contrast, we fail to find any evidence of this relationship. One explanation we offer for this incongruence in results is that the effect of pleasant words may be time-bound to the earlier terms for which we lack justice-level data. To gain some empirical leverage on this intuition, we simply re-estimated our Court Outcome model a series of times, each time omitting an individual natural court. We use natural court – as opposed to each term – as the intuition is driven by personnel-based changes to the overall tenor of the language used by the justices (as opposed to some strictly time-based feature).

Interestingly, when we exclude the 715 case outcomes decided by the 7th (and final) configuration

[1]For example, upon his confirmation in September 2005, the Court became "Roberts 1," which still included Justice Sandra Day O'Connor. Roberts 1 lasted only through the end of January 2006, when Samuel Alito was confirmed by the Senate and replaced Justice O'Connor. This led to the dissolution of Roberts 1 and the creation of Roberts 2, which continued until the retirement of Justice David Souter at the end of the 2008 term. The confirmation of Sonia Sotomayor to replace Souter created Roberts 3.

[2]Full citations: Gelman, Andrew and Hal Stern. 2006."The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant." *American Statistician.* 60(4): 328-331; Gill, Jeff. 1998. "The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52(3): 647-675 (September).

of the Burger Court,[3] the size of the coefficient on our pleasant words variable plummets by 28 percent and its p-value increases from 0.054 (in the fully pooled model) to 0.26. For the exclusion of any other natural court, however, the coefficient remains relatively stable and the p-value generally stays at conventionally accepted significance levels (i.e., $p < 0.10$, two-tailed test).

We submit that this evidence is at least suggestive that the predictive value of pleasantness is being driven, in large part, by the last five terms of the Burger Court, 1981-1985. Unfortunately, as we lack justice-level data for this time period – or any term prior to 2004 – we cannot definitively speak to whether this intuition is accurate.

5 Court Outcome Model

5.1 Missing Transcripts

We note in footnote #8 that we are missing 51 transcripts from LexisNexis. While this constitutes only 1.7 percent of the total number of potential cases, 36 of the 51 missing transcripts come from the 1979 term, which constitutes a missing rate of roughly 28 percent for that single term (the other 15 cases are evenly distributed across the 1980 and 1981 terms and have missing rates equal to 5.6 percent for both).

After verifying with LexisNexis that the transcripts are indeed unavailable, we took several descriptive steps to assess whether these cases were missing at random. Ultimately we believe they are essentially random omissions. There appears to be no systematic differences between our sample and missing cases with regards to the issue area, the rate at which the petitioner wins, the ideological direction of the lower court decision, and the average size of the majority coalition.

Finally, we re-estimated our Court outcome model and excluded the terms affected by the missing transcripts. The substantive results we report in the main text remain virtually identical for each of these re-estimations.

---

[3] Burger 7 consisted of Chief Justice Burger and Justices Brennan, White, Marshall, Blackmun, Powell, Rehnquist, Stevens, and O'Connor.

Interestingly, even though Burger 7 lasted for just five years, it decided only 131 fewer cases than Rehnquist 7 (i.e., Chief Justice Rehqnuist and Justices Stevens, O'Connor, Scalia, Kennedy, Souter, Thomas, Ginsburg, and Breyer), which lasted for more than 11 years.

5.2 Additional Results Figures

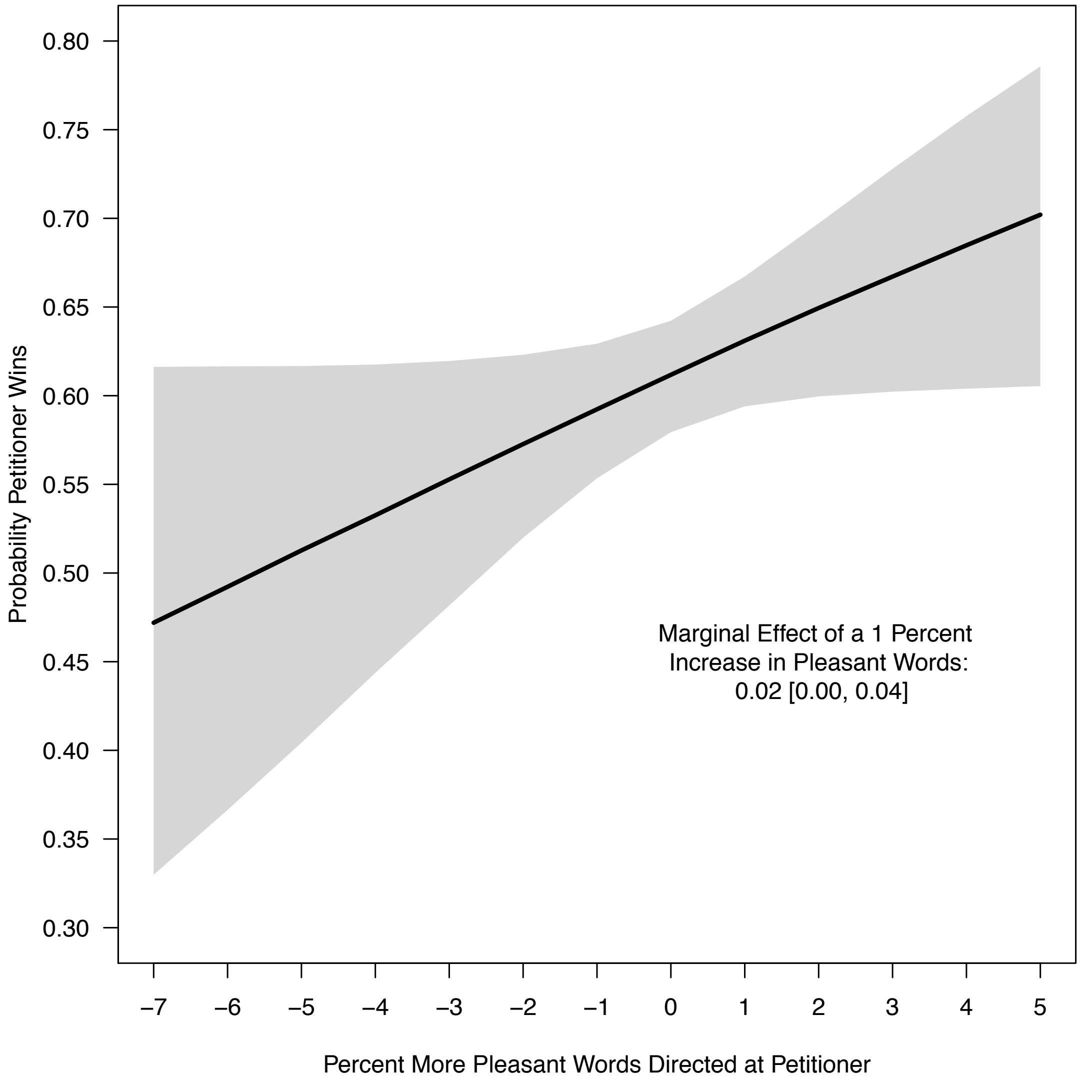

Figure 2: Effect of pleasant language on the Court's likelihood of ruling for the petitioner. 1 percent is slightly larger than one standard deviation (1 S.D. = 0.99). Marginal effect calculated with baseline value of pleasant words at -7. A one-percent increase is always statistically significant for all baseline values at the 95 percent level (two-tailed). All other variables held at their sample mean or mode as appropriate. Shaded region represents 95 percent confidence interval obtained through stochastic simulations.

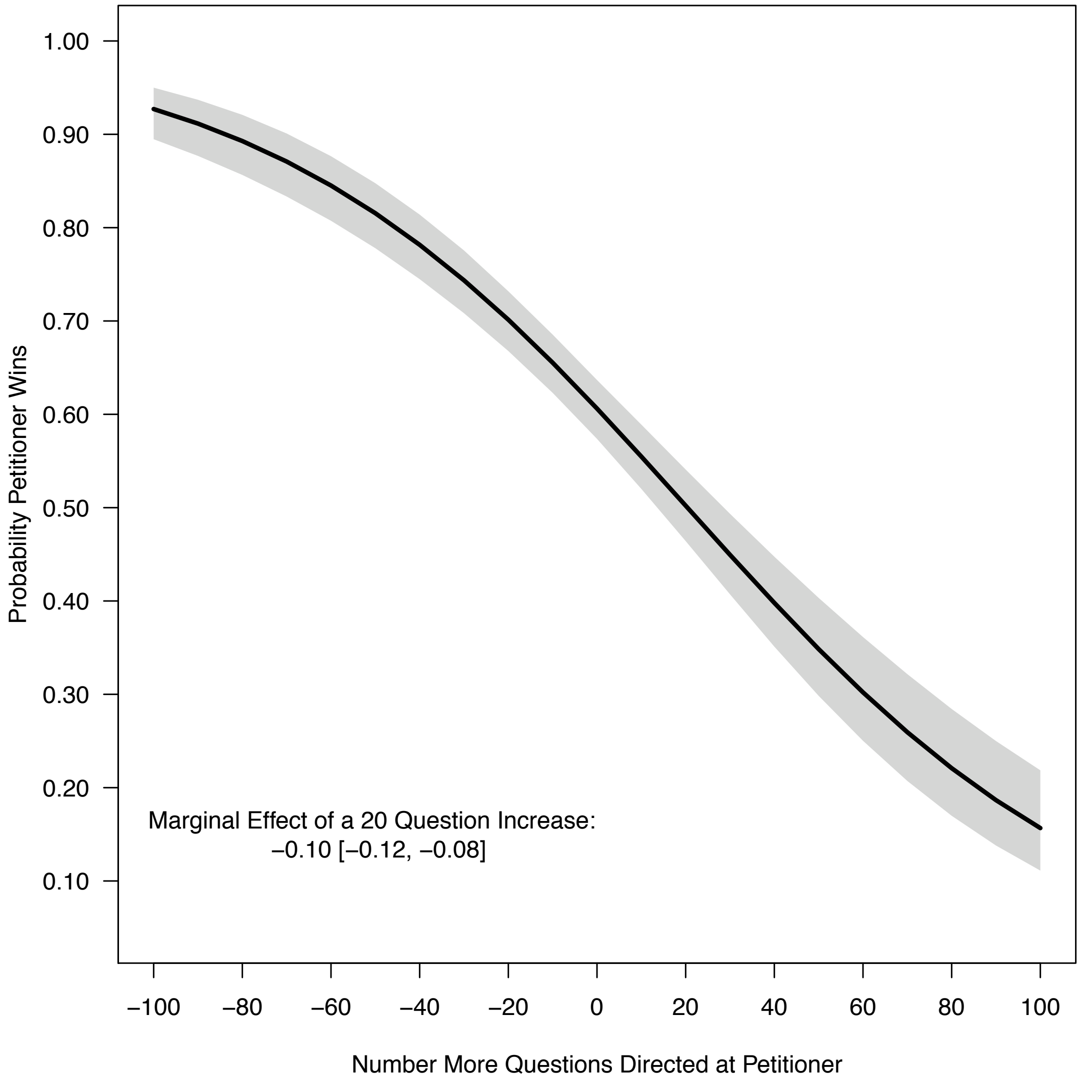

Figure 3: Effect of number of questions on the Court's likelihood of ruling for the petitioner. 20 questions is slightly smaller than one standard deviation (1 S.D. = 24.7). Marginal effect calculated with baseline value of question difference at 0. A twenty question increase is always statistically significant for all baseline values at the 95 percent level (two-tailed). All other variables held at their sample mean or mode as appropriate. Shaded region represents 95 percent confidence interval obtained through stochastic simulations.

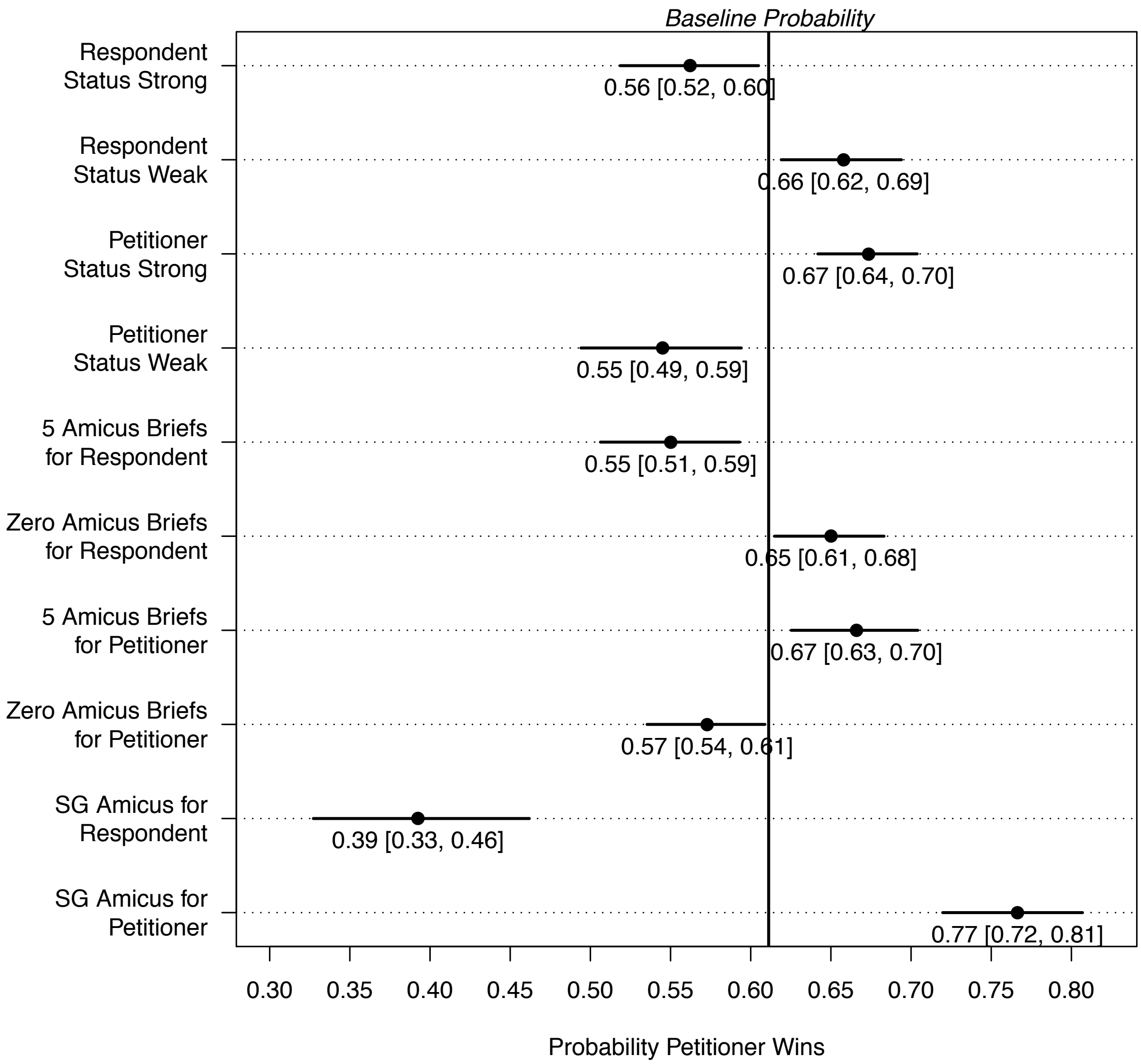

Figure 4: Effect of control variables on the Court's likelihood of ruling for the petitioner. Strong status denotes a value of 9 (i.e., state government), whereas weak denotes a value of 3 (i.e., individual). The sample mean value used in the baseline calculation is 6, which is a business. The baseline value for amici briefs is 2 briefs. 0 briefs represents the minimum and 5 briefs represents roughly one standard deviation above the mean. The SG participated as amicus supporting the respondent or petitioner in 18 percent and 9 percent of the cases in our data, respectively.

6 Justice Votes Model

6.1 Additional Results Figures

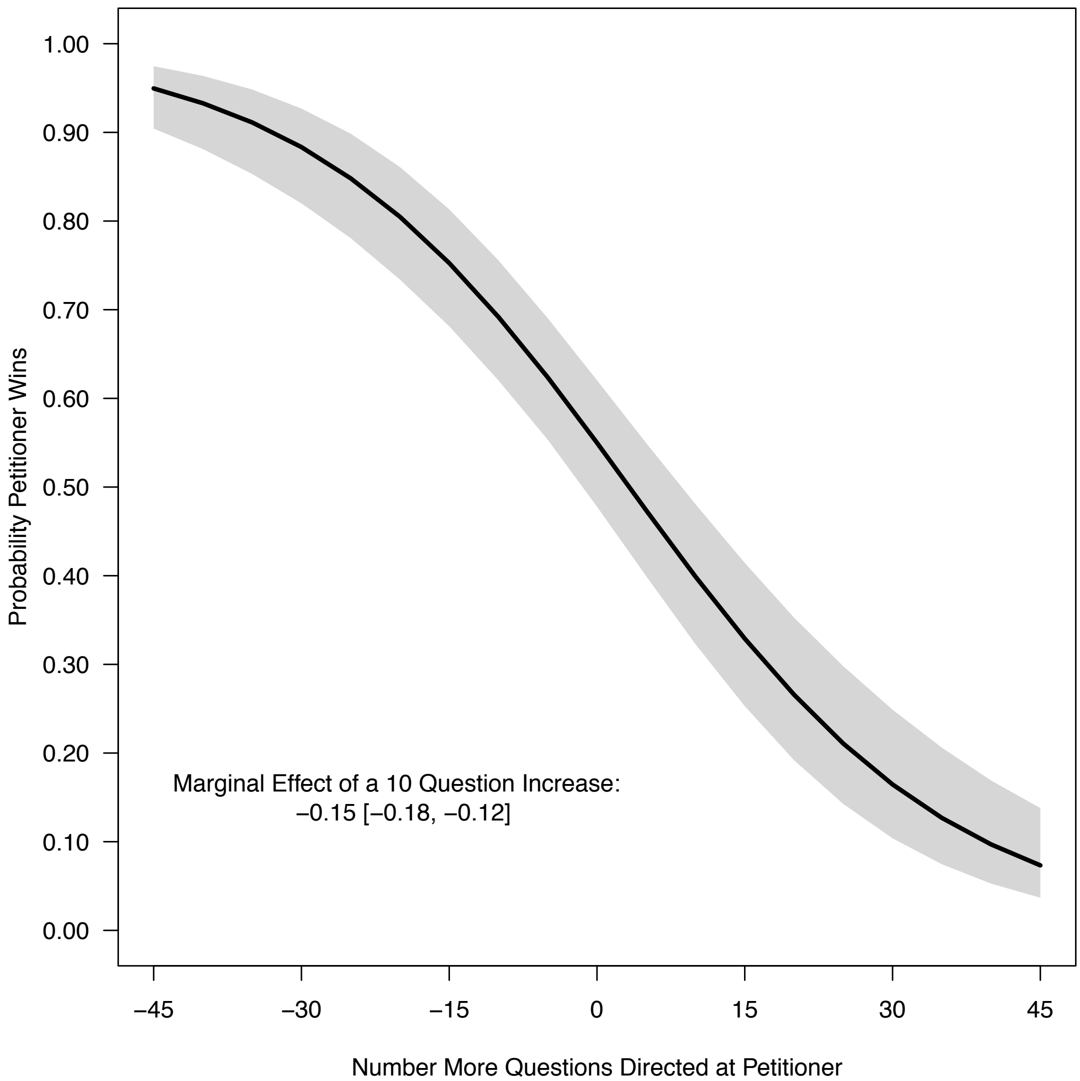

Figure 5: Effect of number of questions asked on a justice's probability of voting for the petitioner. 10 questions is slightly larger than one standard deviation (1 S.D. = 8.5). Marginal effect calculated with baseline value of question difference at 0. A ten question increase is always statistically significant for all baseline values at the 95 percent level (two-tailed). All other variables held at their sample mean or mode as appropriate. Shaded region represents 95 percent confidence interval obtained through stochastic simulations.

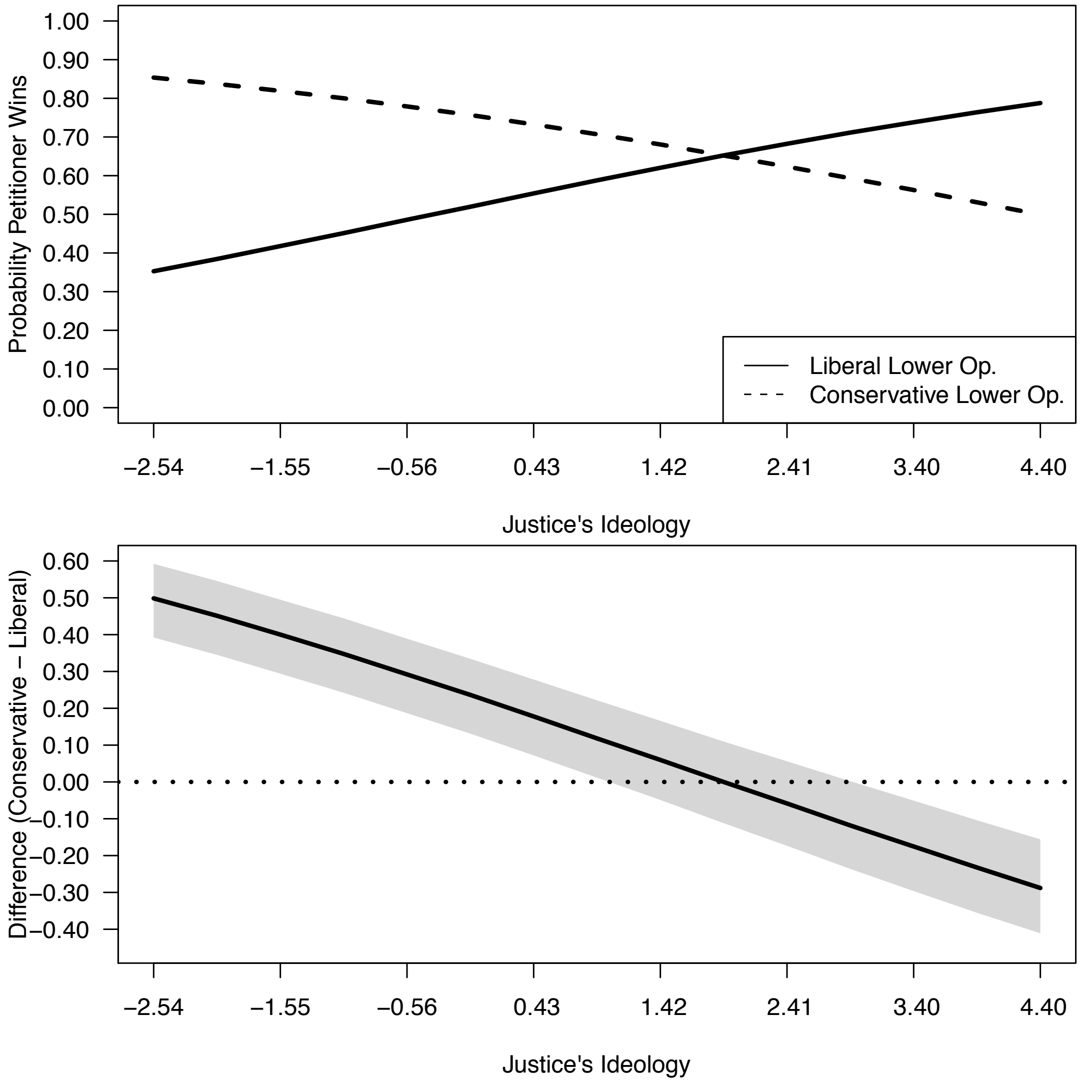

Figure 6: Conditional effect of a lower court opinion's ideological direction on a justice's probability of voting for the petitioner. The top panel is probability of petitioner winning, conditional on a justice's ideology and the ideological direction of the lower court opinion. The bottom panel is the difference in probability between a conservative and liberal lower court opinion, conditional on a justice's ideology (i.e., the size of the gap between the solid and dashed lines in the top panel). All other variables held at their sample mean or mode as appropriate. Shaded region in the bottom panel represents 95 percent confidence interval obtained through stochastic simulations.

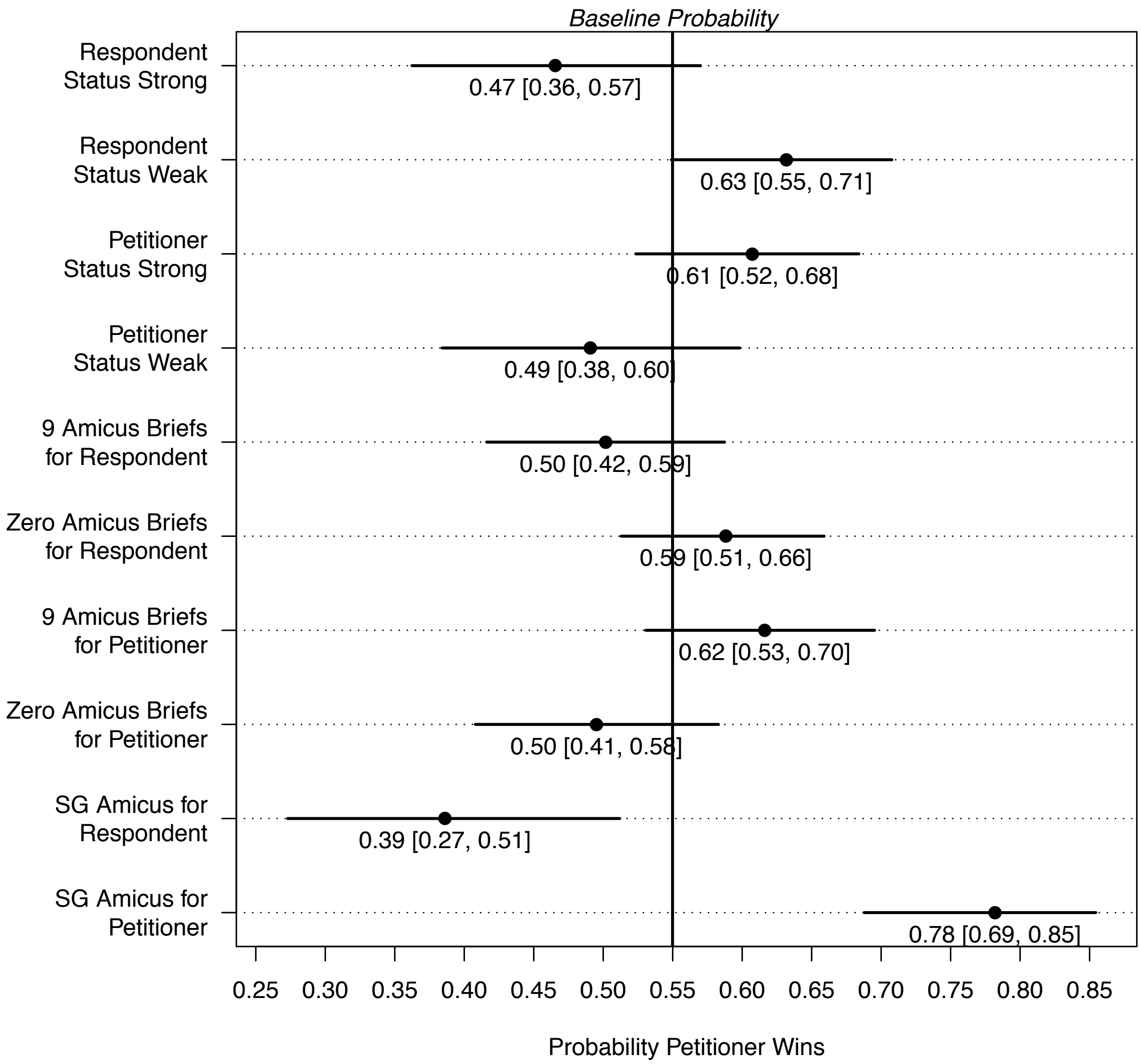

Figure 7: Effect of control variables on a justice's probability of voting for the petitioner. Strong status denotes a value of 9 (i.e., state government), whereas weak denotes a value of 3 (i.e., individual). The sample mean value used in the baseline calculation is 6, which is a business. The baseline value for amici briefs is 4 briefs. 0 briefs represents the minimum and 9 briefs represents roughly one standard deviation above the mean. The SG participated as amicus supporting the respondent or petitioner in 18 percent and 9 percent of the observations in our data, respectively.

6.2 Descriptive Aspects of Emotion

While space constraints prevented us from providing a comprehensive descriptive summary of our data, we suspect some readers would be interested to see how each justice differentially uses emotional language. Here we present a simple table that provides the total number of words uttered by each justice and the percentage of those words that were either pleasant or unpleasant, as measured by Whissell's DAL.

Justice Name	Observations	Words Uttered	Percent Pleasant	Percent Unpleasant
Alito	235	42053	2.78	4.59
Breyer	342	282541	3.17	4.40
Ginsburg	347	181275	2.70	4.14
Kennedy	346	117471	3.36	4.09
O'Connor	93	19511	4.06	4.03
Rehnquist	30	5357	3.86	3.86
Roberts	266	160126	3.28	4.35
Scalia	346	228189	2.97	5.05
Souter	347	210396	3.12	4.49
Stevens	346	108434	2.99	4.14
Thomas	344	278	2.16	3.24