Implicit Associations in Legal Language

Elliott Ash, Daniel L. Chen, Arianna Ornaghi[*]

January 29, 2019

Abstract

This paper provides a quantitative analysis of implicit language associations among judges and legislators using recent machine learning tools designed to assess semantic biases in text corpora. Our measure proxies for implicit associations by looking at relative co-occurrence of sentiment words (e.g. positive versus negative, career versus family) for gender identifiers (man versus woman). Using the universe of published opinions in U.S. Circuit Courts, we document that judicial language displays a stronger associations between men and positive versus negative attributes, and career versus family, with respect to women. Judges displaying higher language bias against women tend to be older, male, and Protestant. Having daughters and increased exposure to female judges in a court reduces bias. Finally, language bias predicts conservative votes on women rights' issues. A preliminary analysis for political language, based on U.S. Congressmen's speeches, shows similar results.

1 Introduction

An active literature in economics and other social sciences has begun to reveal the pervasive social impacts of subtle discrimination towards disadvantages groups, such as women and racial minorities. Implicit attitudes have been shown to influence choices ranging from how physicians' make clinical decisions (Green et al.,

[*] Arianna Ornaghi, University of Warwick, a.ornaghi@warwick.ac.uk (corresponding author); Elliott Ash, ETH Zurich, ashe@ethz.ch; Daniel Chen, Toulouse School of Economics, daniel.chen@iast.fr. We thank David Cai and Christoph Goessmann for helpful research assistance.

2007) to which candidates receive call-backs for job applications (Rooth, 2010). Therefore a research question of major policy relevance is the degree to which policymakers have implicit biases, and whether it affects their decisions. However, standard measures used to proxy for implicit attitudes in the existing literature, such as the Implicit Association Test (IAT), are not generally available in the context of public officials.

This paper addresses this challenge in the context of judges and legislators. Our key contribution is to develop a text-based measure of implicit associations that explains a unique feature of the setting – the large corpus of written text that is available for appellate judges, and a large corpus of spoken text that is available for congressmen. We combine these texts with recent machine learning applications aimed at measuring semantic biases in text corpora (Caliskan et al., 2017).

Our measure proxies for implicit attitudes by looking at the difference between the relative co-occurrence of attribute words (e.g. positive versus negative, innocent versus guilty, arts versus science) and words for protected social group distinctions (man versus woman, white versus black, white versus hispanic). The language context is the universe of U.S. Circuit Court opinions for the years 1880 through 2013, and the universe of floor speeches in the U.S. Congress for the years 1870 through 2016.

We focus on gender associations. Judges that are more biased in their language against women tend to be older, male, and Protestant. Having daughters reduces language bias against women. Language bias predicts conservative votes on abortion and sex discrimination decisions and congressional votes on reproductive rights. When female representation increases in a court, language bias among male judges decreases. These results are robust to adjusting for a wide variety of biographical covariates. Preliminary analysis on political language suggests similar results.

The rest of the paper is organized as follows. Section 2 provides a literature background. Section 3 describes the data and methodology. Section 4 provides descriptives on the language associated with relevant social and cultural dimensions. Section 5 reports our main results and Section 6 the preliminary analysis based on political language. Section 7 concludes.

2 Literature Review

2.1 Implicit Attitudes

The first major literature to which this paper contributes is the large literature in social psychology on implicit attitudes. The literature on the Implicit Association Test – the preferred instrument in psychology to measure implicit bias – is vast (Greenwald et al., 1998). It has been shown that the test results are highly correlated with judgments and choices (Bertrand et al., 2005). Green et al. (2007) shows that implicit bias affects physician treatment choices. Friese et al. (2007) show that implicit attitudes are predictive of subsequent voting. Rooth (2010) show that measured implicit bias is related to which candidates receive call-backs for job applications.

Glover et al. (2017) analyze to which extent a biased prior may affect the productivity level of minority workers. In the study, cashiers are randomly allocated to managers and the bias of managers is measured through an IAT. The authors show that biased managers induce workers from minority groups to under-perform. The evidence is consistent with less manager-worker interaction being the key mechanism (McConnell and Leibold, 2001; Dovidio et al., 2002; Hebl et al., 2002; Dovidio and Gaertner, 2008).

2.2 Using Word Embeddings to Measure Associations in Language

The second literature that this paper contributes to is the work on understanding human motivations and preferences using written or spoken natural language. There is an emerging literature that represents natural language objects (words, phrases, and documents) in a vector space and analyzes their spatial relations, for instance in order to detect biases.

In the past, the common approach was qualitatively oriented, with either a deep reading of the text or a subjective coding of important themes (see Glaser and Strauss (2017) for an example of the latter approach). However, these approaches lack a rigorous method to replicate them (Ricoeur, 1981; DiMaggio, 1997). As a consequence, more formal methods to analyse texts were developed (Andrade, 1995; Mohr, 1998), with semantic networks and topic modeling being the favourites in this regard. The first method perceives words as nodes in a network and textual co-occurences as links (Kaufer and Carley, 1993; Carley, 1994; Corman et al., 2002; Pachucki and

Breiger, 2010; Lee and Martin, 2015). Topic modeling discovers underlying topics and themes through an inductive method (Blei et al., 2003; Blei, 2012; DiMaggio et al., 2013; Mohr and Bogdanov, 2013).

Recent approaches have gone beyond the traditional network or topic methods by mapping word relations into a high-dimensional vector space (Mikolov et al., 2013; Pennington et al., 2014). This method is generally called word embedding. Since word embedding generally positions connected words close to each other, it can be used to detect biases. Kulkarni et al. (2015) use the method to trace the environment of the word *gay* through the 20th century. Another approach to word embeddings is to analyse analogies that computers induce from texts (Bolukbasi et al., 2016; Caliskan et al., 2017). A further refinement of the method is to translate text into a hyperbolic space whereby not only analogies but also more general word connections are possible to detect (Handler, 2014; Rei and Briscoe, 2014).

The results of word embedding associations have been shown to be close to results of implicit association tests, thus suggesting that they can help to detect unconscious attitudes (Caliskan et al., 2017). Garg et al. (2018) trace for example gender and ethnic stereotypes over time. Kozlowski et al. (2018) present word embedding as a method to detect biases in texts. As case study, the paper uses word embedding to detect class and gender biases. The current paper seeks to apply these methods to judicial and legislative texts.

3 Data and Methodology

3.1 U.S. Circuit Court Data

The analysis utilizes a corpus of all U.S. Circuit Court cases, for the years 1870 through 2013. We have the full text of all opinions related to these cases. We also have detailed metadata for each case, from which we use in particular the court and authoring judge. The cases are linked to biographical information on the judges obtained from the Federal Judicial Center. This includes birth year, gender, race, religion, and political affiliation of appointing president. When possible, we additionally merge these data with information on whether judges have children, and their gender, from Glynn and Sen (2015).

3.2 Word Embeddings

We model language using word embeddings. Word embeddings represent words as dense, relatively low-dimensional vectors in a Euclidean space. In our case, a very sparse term frequency matrix with 50,000 columns (representing counts for a vocabulary of 50,000 words) is reduced to a dense embedding matrix with just 300 dimensions. The defining characteristic of word embeddings is that words with similar meaning have similar vector representations (i.e. are represented by vectors with a high cosine similarity): word embeddings preserve semantic relations.

The model we use is GloVe, implemented in Python. The model's objective function is to predict the context of a given word (a window of neighboring words). Given the prediction exercise behind the model, it follows that words that are used in similar contexts will have similar representations, which explains why semantic relationships are preserved. Key parameters are: 20 epochs, 300 dimensional vectors, 0.05 learning rate, window of 10 words.

We construct the input of the model as follows. First, we clean the raw text (e.g. removing HTML markup and citations) so that each opinion is represented as a list of words, segmented by sentence. Then, we remove from these lists uncommon words that are not part of the chosen vocabulary. These lists of words provide the inputs for the embeddings model.

We train both global embedding models based on the full corpus, and judge-specific word embedding models in which opinions written by each individual are treated as a separate corpus. A challenge when training individual-specific embeddings is that, as shown in Antoniak and Mimno (2018), embeddings trained on relatively small corpora might be sensitive to specific documents included in the corpus. To address this issue, we first restrict the sample to individuals that have a corpus of at least 250'000 separate tokens after cleaning.[1]

In addition, we train embeddings while bootstrapping the corpus, as suggested in Antoniak and Mimno (2018). To this end, sentences are treated as documents and sampled with replacement. As shown in Antoniak and Mimno, documents segmented at the sentence level produce embeddings with less variability across bootstrapped samples, compared to larger segments. The number of sentences we include in each bootstrapped sample is the same as the number of total sentences

[1] We explored robustness to selecting different thresholds and most of our results hold with more than 100,000 tokens.

written by the individual. We use 25 bootstrapped samples in the full corpus, and then 10 bootstrapped samples in the individual-level corpora.

3.3 Using Word Embeddings to Identify Cultural Dimensions in Language

A key feature of word embeddings is that, as pointed out by Kozlowski et al. (2018), they are constructed in a space that respects Euclidean geometry. This means that it is possible to identify semantic or cultural dimensions as vectors that define a "step" in a particular direction. For example, we can identify a gender dimension as a vector that takes a step from female toward male. As a result, calculating cosine similarities of other words to these dimensions allows us to extrapolate meaning and understand how words are connotated along these dimensions. Considering the gender dimension from before, a more positive correlation will be associated with more masculine words, whereas a more negative correlation will be associated to more feminine words.

We construct cultural dimensions following Kozlowski et al. (2018). They identify a cultural dimension starting from a set of word pairs "such that the difference between each word in a pair is a step along the dimension of interest." We define each dimension by taking the average of the vector difference between all pairs:

$$\frac{\sum_{p}^{|P|} (\overrightarrow{p}_1 - \overrightarrow{p}_2)}{|P|}$$

where $\vec{p}_1$ and $\vec{p}_2$ represent the vector endpoints for the two comparison words (e.g. "man" and "woman"). Using this method, we identify not only the gender dimension (male/female), but also two attribute dimensions that we use to study the connotation that the gender dimension takes on in each corpus: a positive/negative dimension and career/family dimension.

To validate the method and check that the gender cultural dimension does convey meaning, we projected names onto the dimension and tested whether they can be correctly classified using a cutoff rule. As shown in Table 1, we find that sorting names by their cosine similarity, and predicting whether a first name is a male or female name based on whether the cosine similarity with the gender dimension is positive or negative, does indeed recover the correct classification in the vast majority of the cases.

Table 1: Classification Accuracy of Group-Distinctive Names using Cutoff Rule

	% correctly identified	F1 score
Gender	96.50	0.965

3.4 Testing for Cultural Associations using Word Embeddings

To formally test whether there exists an association between the gender dimension and the two attribute dimensions, we use the Word Embedding Association Test developed by Caliskan et al. (2017). The idea behind WEAT is to test whether stereotypical associations are more likely to occur than non-stereotypical ones. More precisely, given two sets of target words (e.g. male and female words) and two set of attribute words (e.g. positive and negative words), WEAT asks whether there is a difference between the relative similarity of two set of target words with respect to the two sets of attribute words.

More formally, let X, Y be the two sets of target words and A, B be the two set of attribute words. Then, the WEAT test statistic is defined as:

$$WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in X} s(y, A, B)$$

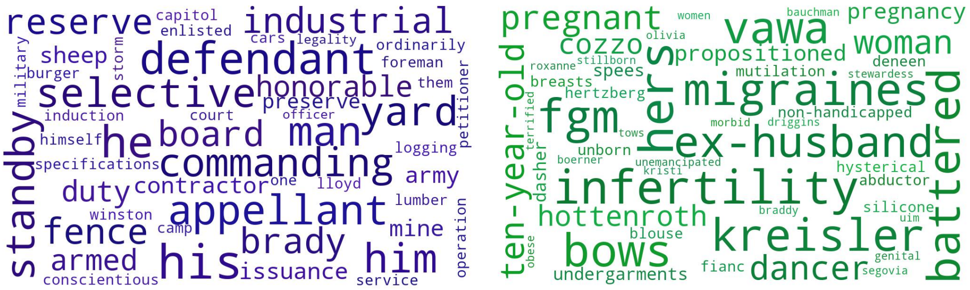
$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

where $\cos(w, a)$ is the cosine similarity between the two word vectors w and a. To enhance comparability across embeddings, we also define a WEAT effect size, which normalizes the WEAT score by the standard deviation of the cosine similarity across all target vectors, as follows:

$$WEAT = \frac{\sum_{x \in X} s(x, A, B) - \sum_{y \in X} s(y, A, B)}{SD_{w \in W} s(w, A, B)}$$

The sets of target and attribute words we use are the same as the ones used to define the dimension poles, ordered such that a higher value of the score corresponds to a stronger stereotypical association (male/positive versus female/negative; male/career versus female/family). We compute WEAT score effect sizes for all embeddings, and assign to each judge the median WEAT effect size across the different bootstrapped samples.

Figure 1: Words with Strongest Gender Association in Judicial Corpus



(a) Male

(b) Female

4 Cultural Dimensions in Judicial Language

We begin by studying implicit associations in judicial and political language using global embeddings trained on the full corpus of Circuit Court opinions. Equivalent figures for Congress are available upon request.

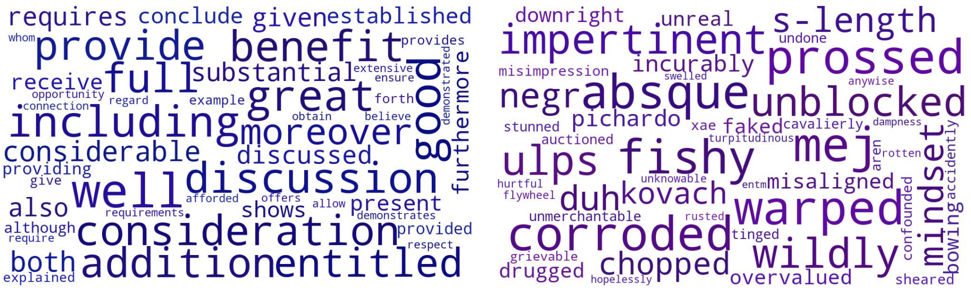
4.1 Most Similar Words

To understand how cultural dimensions are used in our corpora, we begin by asking what words have the highest similarity to either direction of the dimension (e.g. the direction pointing to female versus the direction pointing to male) we consider. This provides a first measure of how the dimensions are connotated in our corpus. We present the results using world clouds in which the larger the word, the stronger the cosine similarity of the adjectives with the respective vector. The examples here are from a randomly selected bootstrap sample of the judicial corpus.

We begin by looking at the gender dimension. The two panels of Figure 1 show words that are more strongly correlated with male gender (left panel) and female gender (right panel). Words associated with the male gender not surprisingly include some of the target words used to identify the dimension (e.g. man, he, his, him), words related to the judicial system (e.g. defendant, appellant, court), but also some stereotypical associations (e.g. honorable, industrial, conscientious). Words associated with the female gender include again some of the target words used to identify the dimension (e.g. woman, her, women) but also interestingly many words associated with family (e.g. pregnancy, infertility, ex-husband).

Figure 2 shows word clouds constructed in the same way for the positive (panel

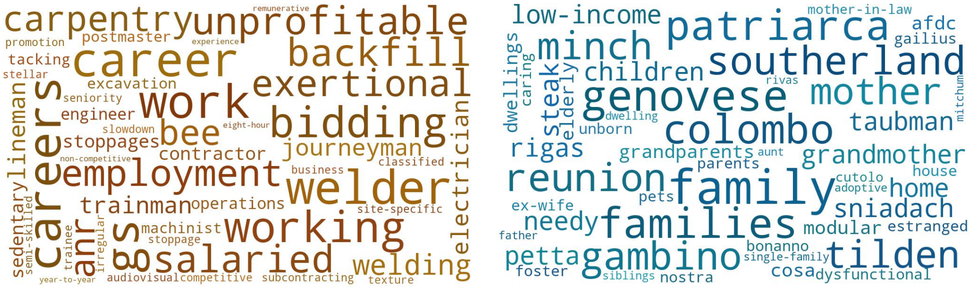
Figure 2: Words with Strongest Positive Association in Judicial Corpus



(a) Positive

(b) Negative

Figure 3: Words with Strongest Career Association in Judicial Corpus



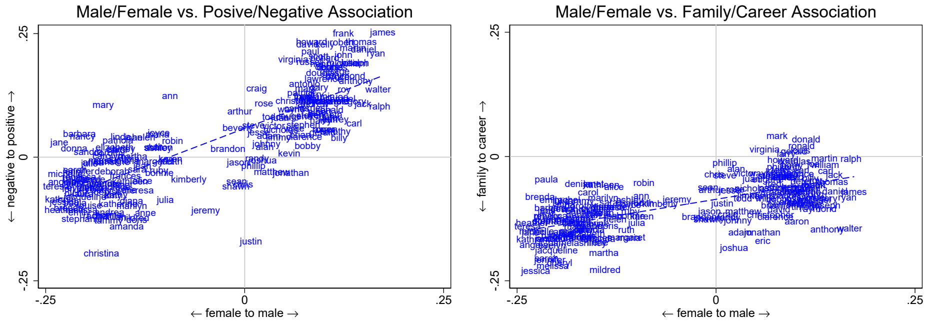
(a) Career

(b) Family

to the left) and negative (panel to the right) poles of the respective attribute dimension. Reassuringly, most similar words associated to the dimension make intuitive sense. Words with the most positive connotation include substantial, opportunity, considerable, respect, while words with the most negative connotation include warped, corroded, fishy, impertinent.

Similarly, Figure 3 shows word clouds for the career (panel to the left) and family (panel to the right) poles of the respective attribute dimension. Again, words most closely associated with the career pole of the dimension are all about employment (e.g. employment, salaried, welder), while those most closely associated with the family pole related to the home (e.g. children, mother, grandmother), although not all of them place it in a positive light (e.g. Gambino, Cutolo - family names related to organized crime). Overall, the word clouds show that the types of implicit social attributes that we are trying to measure come through quite well in the language associations.

Figure 4: Gender Associations in Judicial Language



4.2 Gender Associations

In this section, we show some suggestive association between the gender dimension and the two attribute dimensions by showing the attribute connotation of words with a strong gender connotation: the most common male and female first names from the 1990 census (the most recent census for which this information is available). For each name, we consider the median similarity across the 25 bootstrap samples of the global embeddings. Figure 4 visually presents the results. The x-axis reports the cosine similarity between the vector representing a given first name and the vector representing the gender dimension. Higher values (further to the right) correspond to names with a stronger male connotation and lower values (further to the left) correspond to a stronger female connotation. The y-axis reports the cosine similarity between the name and a corresponding attribute dimension. The y-axis label shows how to interpret the direction of the correlation (e.g. ← negative to positive → means that higher (lower) values correspond to names with stronger (weaker) positive sentiment connotation.

The left panel of Figure 4 shows that male names tend to have an overall more positive connotation with respect to female names. Interestingly, all first names have a family connotation to some extent, but consistent with stereotypical views that women tend to be more closely associated with family with respect to career, female names are more strongly associated with the family dimension with respect to male names.

Table 2: WEAT Scores and Effect Sizes in Judicial Language

	Score	Effect Size	Share Signif. at 95% level	Share Signif. at 90% level
Male/Female vs. Positive/Negative	0.317	1.212	0.56	0.88
Male/Female vs. Career/Family	0.294	0.827	1.00	1.00

4.3 WEAT scores

Figure 4 shows suggestive evidence that men are presented in judicial language as having a more positive connotation than women, and are also more closely associated with career versus family with respect to women. We now show that these associations are confirmed when we formally test for them using the test presented in Section 4.3. In particular, Table 2 shows the median WEAT score and effect size across bootstrapped samples for the gender/good and gender/career association. In addition, we display the share of bootstrap samples for which the score is significantly different than 0 based on the permutation test developed in Caliskan et al. (2017). The table shows a positive WEAT score for both associations, meaning that men have a stronger association with positive (career) versus negative (family) attributes with respect to women, and that this association is significantly different than 0 in the majority of the bootstrap samples.

5 Language Associations in Judicial Decisions

The results shown thus focus on the entire corpus, but we might be interested in understanding whether the overall level of language bias is correlated with judge characteristics. To do that, we exploit judge specific embeddings that treat all of the opinions authored by a certain judge as a separate corpus. We assign to each judge their median effect size across the ten bootstrap samples. We begin by showing descriptively correlates of WEAT effect sizes, and then move on to discussing how female representation impacts the WEAT scores of male judges in the circuit and the relationship to votes.

5.1 Implicit Language Associations and Biographical Characteristics

This section analyzes how language bias is related to judge characteristics. Each graph in Figure 5 shows the mean WEAT effect size for judges with different biographical characteristics, together with 95% confidence intervals for the mean. All graphs to the left refer to the male/female versus positive/negative association, while all graphs to the right refer to the male/female versus career/family association. There is a large difference across genders in the gender WEAT, with male judges displaying higher lexical gender bias. However, there is no difference between the political party affiliation of a judge's nominating president. There is a significant difference by religion, with Protestant judges having a high WEAT score and non-religious judges having a low WEAT score. The biggest differences are by the cohort of the judges, with early cohorts having much higher WEAT scores than the more recent cohort.

The graphs in Figure 5 show descriptively the variation in the raw data separately for the different characteristics. Table 3 show the same descriptive correlations in regression form. In particular, it shows the coefficients from a regression of the WEAT effect size for the two associations of interest on biographical characteristics and circuit fixed effects, which allows to potentially gain precision and control for all characteristics jointly. Table 3 columns (1) and (3) shows an overall similar pattern as the graphs, in particular with judges from older cohorts having a much higher WEAT effect size with respect to younger judges. Interestingly, conditional on other characteristics, judges appointed by a Democratic president now display a lower WEAT scores in the gender/career association.

We also explore whether having a daughter has an effect on the language bias displayed by judges. Table 3 columns (2) and (4) display the coefficients from the same regression as before but also including a dummy for having at least one daughter and number of children fixed effects, which are fundamental to control for family composition and thus move towards causality. The sample size is significantly lower as the information on children is only available for a subset of our judges. Interestingly, it appears that having a daughter induces judges to display a lower association between gender and career/family.

Figure 5: Gender WEAT Tests by Judge Characteristics

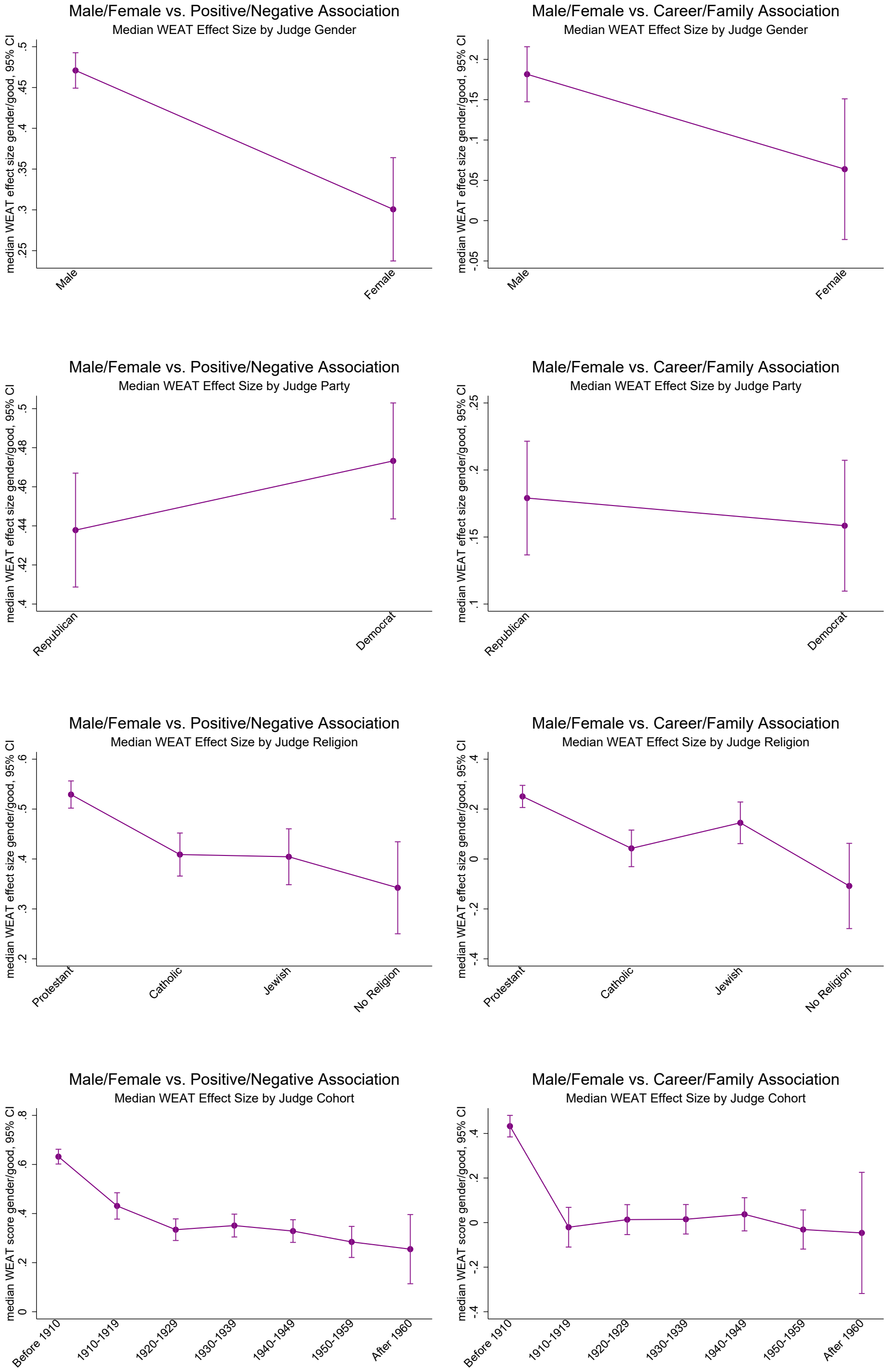


Table 3: WEAT Effect Sizes and Judges' Biographical Characteristics

	Male/Female vs. Positive/Negative Association		Male/Female vs. Career/Family Association	
	(1)	(2)	(3)	(4)
Democrat	0.002	0.043	-0.063**	-0.111**
	(0.019)	(0.033)	(0.031)	(0.055)
Female	-0.036	-0.085**	0.037	0.014
	(0.036)	(0.038)	(0.058)	(0.070)
Minority	0.006	-0.031	-0.024	0.017
	(0.033)	(0.048)	(0.060)	(0.095)
Protestant	0.014	0.040	0.044	-0.075
	(0.039)	(0.063)	(0.060)	(0.089)
Catholic	-0.011	0.013	-0.021	-0.186*
	(0.043)	(0.065)	(0.064)	(0.097)
Jewish	-0.016	-0.011	0.054	-0.043
	(0.044)	(0.068)	(0.068)	(0.094)
Born in 1910s	-0.184***	-0.163***	-0.439***	0.285
	(0.031)	(0.060)	(0.054)	(0.294)
Born in 1920s	-0.286***	-0.306***	-0.424***	0.329
	(0.027)	(0.052)	(0.045)	(0.285)
Born in 1930s	-0.264***	-0.298***	-0.436***	0.262
	(0.029)	(0.053)	(0.047)	(0.286)
Born in 1940s	-0.274***	-0.276***	-0.395***	0.270
	(0.032)	(0.055)	(0.053)	(0.287)
Born in 1950s	-0.341***		-0.473***	
	(0.045)		(0.067)	
Born after 1960	-0.459***		-0.590***	
	(0.107)		(0.101)	
Has Daughters		0.011		-0.154**
		(0.040)		(0.063)
Observations	616	223	617	223
Circuit FEs	yes	yes	yes	yes
Children FEs		yes		yes

Table 4: Effect on Gender Bias of Greater Female Representation

	Male/Female vs. Positive/Negative Association				Male/Female vs. Career/Family Association			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share Judge Female	-0.116	-0.361	-0.439*	-0.449*	0.109	-0.0819	-0.0242	-0.0913
	(0.129)	(0.200)	(0.167)	(0.150)	(0.275)	(0.354)	(0.184)	(0.184)
Observations	11565	11565	11565	11293	11733	11733	11733	11733
Clusters	10	10	10	10	10	10	10	10
Year FEs	yes	yes	yes	yes	yes	yes	yes	yes
Circuit FEs		yes	yes	yes		yes	yes	yes
Circuit Trends			yes	yes			yes	yes
Lagged DV				yes				yes

5.2 Female Representation and Language Bias

What are the determinants of changes in language bias over time? One reason could be that as women join the courts, exposure to women would reduce these biases among male judges. To get at this issue, we computed the WEAT scores at the court-year level for all circuit courts. We paired this with a treatment variable, the proportion of judges that are female on a court. Because new judges are assigned to circuits by the president as slots become available, the timing across circuits of getting more women is likely exogenous. We regressed language bias on proportion female with court and year fixed effects.

The coefficients from these regressions are reported in Table 4. Cross-sectionally (within year), there aren't differences between courts in language bias (Columns 1 and 5). And the effect for the family/career association is consistently zero. However, the relative association of maleness with positiveness is negative, as one can see upon the inclusion of circuit fixed effects (Column 2), circuit trends (Column 3), and the lagged dependent variable (Column 4).

Holding court-level factors constant, and allowing for arbitrary national trends, more female judge representation decreases male judge language bias on the positive/negative margin. As these judges are exposed to more females, they change their language to be less biased. However, there is no effect for gender/career language. This could be that career language tends to be reflected more in case facts, while positive/negative language is more at the discretion of the judge.

5.3 Effects of Language Bias on Decisions and Votes

While the fact that judges display lexical gender bias in their writings is interesting per se, it does not necessarily imply policy relevance. If judges are aware of their biases and correct for them when making decisions, we might expect how judges write about women to have no impact. For these preferences to matter, then, it must be the case that they impact judicial decisions, and this is the dimension we explore in this section. In particular, we ask whether conditional on a number of biographical controls and circuit-year fixed effects, our lexical gender bias measure predicts how judges vote on gender rights cases. Standard errors are clustered at the circuit-year level. The inclusion of circuit-year fixed effects is especially important for identification, as cases are randomly assigned to judge panels within circuit-years. As a result, the effect of the lexical bias measure is not going to be driven by judges with different lexical gender bias selecting to serve on panels for different types of cases, thus addressing an important source of potential endogeneity. We analyse two separate datasets in which judges' decisions in a sample of cases are coded to be pro- and against- women rights.

Table 5 focuses on votes reported by the Chicago Judges Project. We pool cases that relate to reproductive rights, sexual discrimination and sexual harassment and include issue fixed effects to ensure we are only using within issue variation. Column (1) shows a regression of the WEAT scores without demographic controls. Judges with a higher WEAT effect size, i.e. judges that display a stronger stereotypical association between men and career versus family with respect to women, are less likely to vote in favor of plaintiffs in women-rights cases. The inclusion of demographic controls does not affect this result, although the coefficient is indeed smaller. The same is true when we additionally include a dummy for the judge having a daughter and children fixed effects. The WEAT effect size for the gender/good association has a negative coefficient throughout, but is never significant.

Table 6 focuses on votes reported by Glynn and Sen (2015). We pool cases that relate to reproductive rights, employment discrimination and Title IX, again including issue fixed effects. Column (1) shows that, similarly as before, when we have no demographic controls judges with a higher WEAT effect size in the gender/career association are less likely to vote in favor of plaintiffs in women-rights cases. Includ-

Table 5: WEAT Scores and Judge Decisions in Gender Rights Cases (CJP dataset)

	Pro-plaintiff Vote		
	(1)	(2)	(3)
Male/Female vs. Positive/Negative Association	-0.024	-0.069	-0.042
	(0.040)	(0.043)	(0.050)
Male/Female vs. Career/Family Association	-0.061**	-0.054**	-0.052**
	(0.025)	(0.024)	(0.026)
Democrat		0.127***	0.124***
		(0.020)	(0.023)
Female		0.019	0.022
		(0.021)	(0.023)
Has Daughters			-0.003
			(0.026)
Observations	3804	3784	3327
Clusters	163	163	163
Circuit-Year FEs	yes	yes	yes
Additional Demographic Controls	no	yes	yes
Issue FEs	yes	yes	yes
# of Children FEs			yes

ing demographic controls weakens the result: the coefficient is smaller and no longer statistically significant, although controlling for whether a judge has a daughter and children fixed effects restores the original result. Again, the WEAT effect size for the gender/good association has a negative coefficient throughout, but is never significant. Overall, Table 6 and Table 7 show that the lexical bias of judges matter for how they vote, a result that we plan to further explore using District Court data and sentencing decisions.

6 Preliminary Results on Legislator Speech

Are these associations unique to judicial language, or do we see them reflected across different domains? To begin answering this question, we are currently performing an analogous analysis for another set of lawmakers – legislators in the U.S. Senate and House. In particular, we use the digitized Congressional Record, which consists of transcripts of the speeches given by U.S. Congressmen from 1870 through 2015. Each speech is tagged to a Congressman, for which we have a range of metadata on personal characteristics, including gender, party affiliation, race, religion and year of birth. When possible, we merge these data with information on congressmen's children from Washington (2008).

Table 6: WEAT Scores and Judge Decisions in Gender Rights Cases (Glynn and Sen dataset)

	Pro-plaintiff vote		
	(1)	(2)	(3)
Male/Female vs. Positive/Negative Association	0.054	0.023	0.064
	(0.043)	(0.048)	(0.053)
Male/Female vs. Career/Family Association	-0.051*	-0.035	-0.055*
	(0.028)	(0.028)	(0.032)
Democrat		0.114***	0.093***
		(0.022)	(0.024)
Female		0.020	0.043*
		(0.021)	(0.022)
Has Daughters			0.023
			(0.025)
Observations	3602	3601	3319
Clusters	83	83	83
Circuit-Year FEs	yes	yes	yes
Additional Demographic Controls	no	yes	yes
Issue FEs	yes	yes	yes
# of Children FEs	no	no	yes

We clean the data for the U.S. Congress corpus following the same procedure as for the judicial opinions, which results in a sample of 880 congressmen who have more than 250'000 tokens. The main difference with respect to the methodology is that in the bootstrapping procedure, we treat speeches, and not sentences, as documents and sampled with replacement. Given that speeches in congress are generally short, this should minimally impact the results.

We begin by showing how the biographical characteristics of congressmen relate to their language bias. Figure 6 and Table 7 show that for the gender/good association, the qualitative results of judges are replicated. Instead, the correlations for gender/career are inconsistent, even to the point of going in the opposite direction. Interestingly, having a daughter appears to have no effect on the language bias of congressmen.

Next, Table 8 provides preliminary results on language bias and congressional votes on reproductive rights bills using the voting data in Washington (2008). With respect to the judges analysis, there is not randomized assignment, so the analysis relies on controlling for observables for identification. Overall, the results support the view that congressmen with higher positive/negative bias are more likely to vote conservatively on bills that expand reproductive rights. There is no effect of

Figure 6: Gender WEAT Tests by Congressman Characteristics

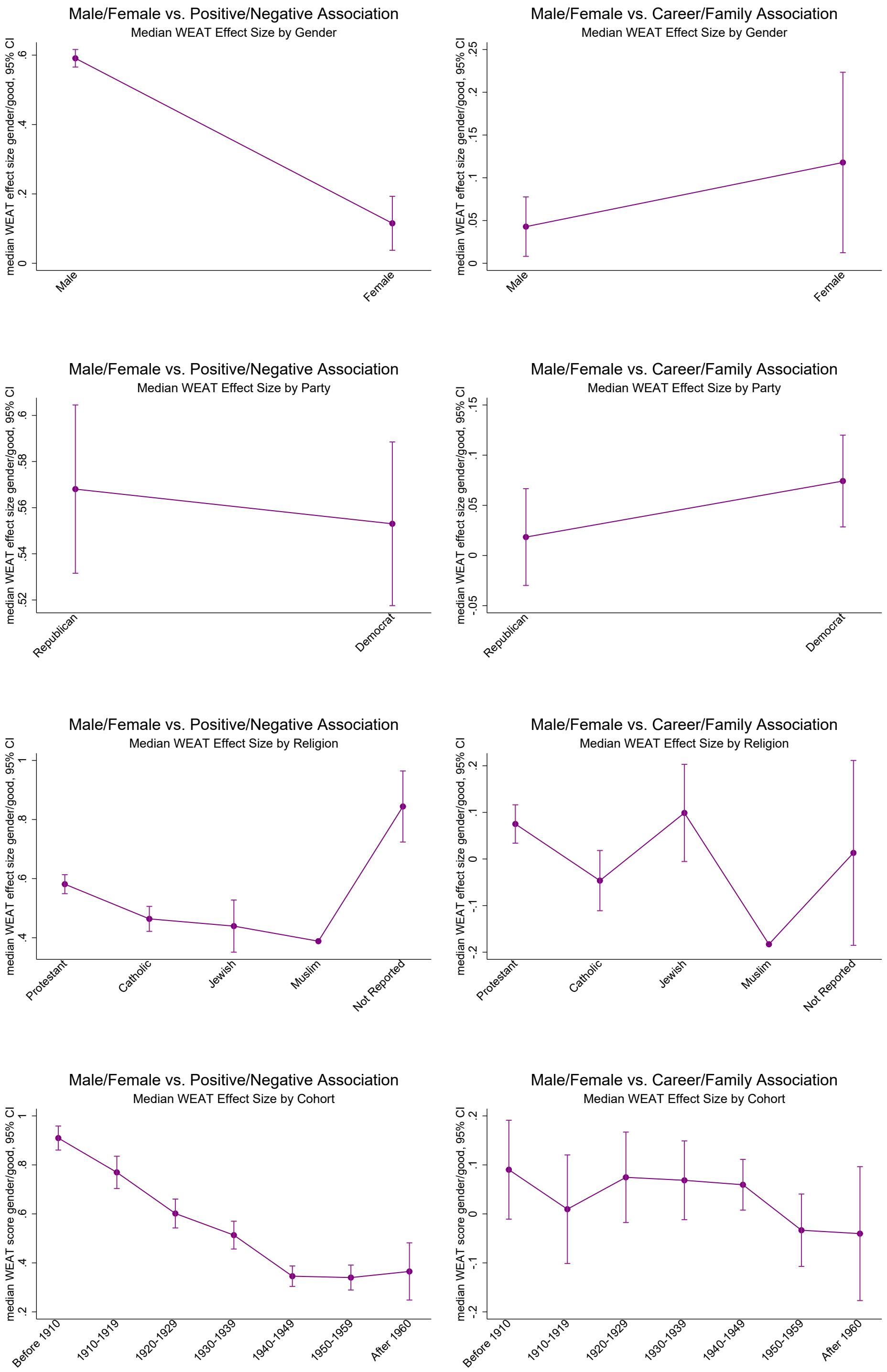


Table 7: WEAT Effect Sizes and Congressmen's Biographical Characteristics

	Male/Female vs. Positive/Negative Association		Male/Female vs. Career/Family Association	
	(1)	(2)	(3)	(4)
Democrat	0.003	0.010	0.084**	-0.067
	(0.022)	(0.062)	(0.037)	(0.071)
Female	-0.316***	-0.324***	0.092*	0.125
	(0.046)	(0.050)	(0.056)	(0.077)
Minority	-0.228***	-0.168**	-0.111**	0.041
	(0.038)	(0.068)	(0.051)	(0.082)
Christian	-0.067	0.194**	0.092	-0.259*
	(0.045)	(0.098)	(0.107)	(0.145)
Catholic	-0.121**	0.207**	-0.044	-0.320**
	(0.049)	(0.095)	(0.108)	(0.145)
Jewish	-0.160***	0.131	0.075	-0.113
	(0.058)	(0.086)	(0.116)	(0.153)
Muslim	0.112		-0.040	
	(0.077)		(0.122)	
Born in 1910s	-0.118***		-0.093	
	(0.042)		(0.077)	
Born in 1920s	-0.276***	-0.175	-0.017	0.217
	(0.038)	(0.184)	(0.070)	(0.199)
Born in 1930s	-0.340***	-0.150	-0.014	0.264*
	(0.038)	(0.175)	(0.066)	(0.139)
Born in 1940s	-0.489***	-0.214	-0.031	0.262**
	(0.032)	(0.172)	(0.059)	(0.116)
Born in 1950s	-0.512***	-0.187	-0.106	0.148
	(0.037)	(0.170)	(0.065)	(0.124)
Born after 1960	-0.481***		-0.104	
	(0.066)		(0.089)	
Daughters		-0.048		0.056
		(0.032)		(0.043)
Observations	880	162	880	162
Children FEs		yes		yes

Table 8: WEAT Scores and Congressman Votes on Reproductive Rights Bills

Issue	Abortion ban	Teen access to abortion	Contraceptives for federal employees	RU2486	Teen access to contraceptives	International family planning
	(1)	(2)	(3)	(4)	(5)	(6)
Male/Female vs.	-0.167	-0.221**	-0.035	-0.211*	-0.097	-0.266**
Positive/Negative Association	(0.108)	(0.107)	(0.116)	(0.109)	(0.124)	(0.107)
Male/Female vs. Career/Family	0.027	0.003	0.088	-0.040	-0.014	-0.054
Association	(0.081)	(0.082)	(0.098)	(0.098)	(0.099)	(0.089)
Democrat	0.489***	0.454***	0.507***	0.471***	0.512***	0.462***
	(0.084)	(0.087)	(0.083)	(0.094)	(0.095)	(0.090)
Female	0.288***	0.211**	0.285**	0.151	0.180*	0.131
	(0.100)	(0.096)	(0.110)	(0.111)	(0.093)	(0.121)
# Daughters	0.107**	0.101**	0.066	0.060	0.084*	0.141***
	(0.042)	(0.042)	(0.047)	(0.048)	(0.045)	(0.041)
Observations	136	136	136	136	136	136
Additional Demographic Controls	yes	yes	yes	yes	yes	yes
# of Children FEs	yes	yes	yes	yes	yes	yes

family/career language bias.

7 Conclusion

This work shows that judicial language exhibits implicit associations between social groups and socially and legally relevant attributes. Future work can look at differences in these measures across individual judges, and even within judge over time. We are interested in looking at other courts, such as the U.S. Supreme Court, U.S. District Courts, and state courts.

References

Andrade, R. (1995). *The development of cognitive anthropology*. Cambridge University Press. 2.2

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119. 3.2

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, 95(2):94–98. 2.1

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84. 2.2

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022. 2.2

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357. 2.2

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. 1, 2.2, 3.4, 4.3

Carley, K. (1994). Extracting culture through textual analysis. *Poetics*, 22(4):291–312. 2.2

Corman, S. R., Kuhn, T., McPhee, R. D., and Dooley, K. J. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human communication research*, 28(2):157–206. 2.2

DiMaggio, P. (1997). Culture and cognition. *Annual review of sociology*, 23(1):263–287. 2.2

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606. 2.2

Dovidio, J. F. and Gaertner, S. L. (2008). New directions in aversive racism research: Persistence and pervasiveness. In *Motivational aspects of prejudice and racism*, pages 43–67. Springer. 2.1

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology*, 82(1):62. 2.1

Friese, M., Bluemke, M., and Wänke, M. (2007). Predicting voting behavior with implicit attitude measures: The 2002 german parliamentary election. *Experimental Psychology*, 54(4):247–255. 2.1

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. 2.2

Glaser, B. G. and Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge. 2.2

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260. 2.1

Glynn, A. N. and Sen, M. (2015). Identifying judicial empathy: Does having daughters cause judges to rule for women's issues? *American Journal of Political Science*, 59(1):37–54. 3.1, 5.3

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., and Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine*, 22(9):1231–1238. 1, 2.1

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464. 2.1

Handler, A. (2014). An empirical study of semantic similarity in wordnet and word2vec. 2.2

Hebl, M. R., Foster, J. B., Mannix, L. M., and Dovidio, J. F. (2002). Formal and interpersonal discrimination: A field study of bias toward homosexual applicants. *Personality and social psychology bulletin*, 28(6):815–825. 2.1

Kaufer, D. S. and Carley, K. M. (1993). Condensation symbols: Their variety and rhetorical function in political discourse. *Philosophy & rhetoric*, pages 201–226. 2.2

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2018). The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*. 2.2, 3.3

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee. 2.2

Lee, M. and Martin, J. L. (2015). Coding, counting and cultural cartography. *American Journal of Cultural Sociology*, 3(1):1–33. 2.2

McConnell, A. R. and Leibold, J. M. (2001). Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of experimental Social psychology*, 37(5):435–442. 2.1

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 2.2

Mohr, J. W. (1998). Measuring meaning structures. *Annual review of sociology*, 24(1):345–370. 2.2

Mohr, J. W. and Bogdanov, P. (2013). Topic models: What they are and why they matter. 2.2

Pachucki, M. A. and Breiger, R. L. (2010). Cultural holes: Beyond relationality in social networks and culture. *Annual review of sociology*, 36:205–224. 2.2

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. 2.2

Rei, M. and Briscoe, T. (2014). Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 68–77. 2.2

Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation*. Cambridge university press. 2.2

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3):523–534. 1, 2.1

Washington, E. L. (2008). Female socialization: how daughters affect their legislator fathers. *American Economic Review*, 98(1):311–32. 6, 6