AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General


ISSN: 0096-3445

2023, Vol. 152, No. 9, 2591–2602
https://doi.org/10.1037/xge0001412

# Quantifying Cultural Change: Gender Bias in Music

Reihane Boghrati[1] and Jonah Berger[2]
[1] W. P. Carey School of Business, Arizona State University
[2] Wharton School, University of Pennsylvania

Cultural items (e.g., songs, books, and movies) have an important impact in creating and reinforcing stereotypes. But the actual nature of such items is often less transparent. Take songs, for example. Are lyrics biased against women, and how have any such biases changed over time? Natural language processing of a quarter of a million songs quantifies gender bias in music over the last 50 years. Women are less likely to be associated with desirable traits (i.e., competence), and while this bias has decreased, it persists. Ancillary analyses further suggest that song lyrics may contribute to shifts in collective attitudes and stereotypes toward women, and that lyrical shifts are driven by male artists (as female artists were less biased to begin with). Overall, these results shed light on cultural evolution, subtle measures of bias and discrimination, and how natural language processing and machine learning can provide deeper insight into stereotypes, cultural change, and a range of psychological questions more generally.

***Public Significance Statement***
Gender bias is pervasive. One reason biases may be so persistent is that they are continually reinforced through cultural items like songs. We use natural language processing and analyze over 50 years of lyrics to quantify gender bias in music over time. Lyrics tend to associate desirable traits (i.e., competence) with men, and while this bias has decreased over time, it persists. The temporal change is mainly driven by male artists, as female artists were less biased to begin with. Overall, these results shed light on cultural evolution, subtle measures of bias and discrimination, and how natural language processing and machine learning can provide deeper insight into stereotypes, cultural change, and a range of psychological questions more generally.

*Keywords:* gender stereotypes, misogyny, music, natural language processing

*Supplemental materials:* https://doi.org/10.1037/xge0001412.supp



Gender bias is pervasive (Carlana, 2019; Charlesworth et al., 2021; Garg et al., 2018; Gruber et al., 2021; Kesebir, 2017; Moss-Racusin et al., 2012). Across a range of disciplines (e.g., science, medicine, and business) and outcomes (e.g., hiring, evaluation, and recognition), women are often perceived less favorably and treated less fairly. The same job applicant, for example, is seen as more competent and offered a higher starting salary if they have a male rather than a female name (Moss-Racusin et al., 2012).

One reason such biases may be so persistent is that they are continually reinforced through culture. Songs, books, and other cultural items not only reflect the setting in which they were produced, but also shape the attitudes and behaviors of the audiences that consume them (Anderson et al., 2003;Berger et al., 2021; Lennings & Warburton, 2011; Packard & Berger, 2020). Song lyrics that are aggressive toward women, for example, or portray them negatively, increase antifemale attitudes, and misogynistic behavior (Fischer & Greitemeyer, 2006). Lyrics that espouse equality, however, can boost attitudes toward women and encourage pro-female behavior (Greitemeyer et al., 2015). Consequently, language plays a critical role in reproducing and disseminating bias.

But while cultural items like songs have impact, their actual nature is less clear. Are song lyrics biased against women, for example, and if so, have they become any less biased over time? Further, what are the potential drivers and consequences of these changes?

This article was published Online First April 13, 2023.
Reihane Boghrati https://orcid.org/0000-0002-5281-1310
Preliminary results were presented at the Marketing Science, Society for Consumer Psychology, and Association for Consumer Research conferences. An earlier version is available on SSRN.

The authors thank Jordan Etkin, Grant Packard, and Sudeep Bhatia for helpful feedback on research and early drafts of the paper and Rishabh Tagore for assistance with data collection. The authors declare no conflicts of interest.

Jonah Berger designed the idea and theory. Reihane Boghrati performed the computations. Jonah Berger and Reihane Boghrati verified the analytical methods. Jonah Berger and Reihane Boghrati wrote the manuscript. Reihane Boghrati served as lead for data curation, formal analysis, methodology, validation, and served in a supporting role for conceptualization and writing–original draft. Jonah Berger served as lead for conceptualization, writing–original draft, and served in a supporting role for data curation, formal analysis, methodology, and validation. Reihane Boghrati and Jonah Berger contributed to writing–review and editing equally.

Correspondence concerning this article should be addressed to Jonah Berger, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, United States. Email: jberger@wharton.upenn.edu

Quantifying culture, and cultural change, can be challenging. While researchers can collect surveys of individual's attitudes or stereotypes, such methods often involve self-report, and can be challenging to collect over long time horizons. Alternatively, researchers can try to manually code archival data, but this is often subjective and difficult to scale.

We suggest that an emerging machine-learning approach may provide a powerful tool to address some of these challenges. More and more textual data is available across all areas of life, and natural language processing provides a useful key to unlock a range of insights (Berger & Packard, 2022; Jackson et al., 2022). We examine how word2vec (Mikolov et al., 2013), a well-adopted word vector representation technique (DeFranza et al., 2020; Mishra et al., 2019), can shed light on questions regarding cultural change (collective representations, Durkheim, 2009; Moscovici, 2001) that might otherwise be difficult or impossible to address.

In particular, we integrate research on culture, language, and social cognition to study gender bias. We start by building a novel data set of more than a quarter of a million songs from six music genres over more than 50 years. This is 250 times larger than any prior investigation and allows for deeper and more comprehensive analyses. We focus on music in particular because of its impact on culture. People of all ages listen to music and the lyrics of popular songs are heard by millions. Then, using state-of-the-art machine learning techniques, we investigate potential gender biases and whether they have changed over time. Further, we examine the potential consequences of such lyrical shifts for social stereotypes. The results address the longstanding debate about gender bias in music, shed light on cultural evolution, and showcase how emerging natural language processing techniques can provide insight into cultural change and a range of psychologically relevant questions more broadly.

Note that while there are many types of bias one could examine, given competence is a universal dimensions of social cognition (Fiske et al., 2007), we focus our attention there. Further, given the importance of competence perception for a variety of judgments and actions (e.g., hiring and promotion), it is particularly important to understand cultural content that may contribute to gender biases in this area. We also explore related biases (i.e., warmth, intelligence, and masculine and feminine stereotypes).

## Sources of Bias

Two major sources of information contribute to stereotypes and biases (Lewis & Lupyan, 2020). The first is direct experience (E. R. Smith & DeCoster, 1998). If most nurses someone observes are women, for example, and most chemical engineers they observe are men, they might conclude that women are better suited for nursing while men are better suited for chemical engineering. Similarly, observing objects in a computer science classroom can provide insight into the type of people who tend to be computer scientists, and thus impact female students' interest in the discipline (Cheryan et al., 2009).

Outside direct experience, however, language plays an important role. Even without any direct experience with nurses or chemical engineers, hearing someone talk about these two groups may provide information. If people use the pronoun "she" when talking about nurses, for example, listeners may assume most nurses are female. Similarly, hearing someone recount a story about their computer science class may provide insight into the gender makeup and stereotypes associated with the discipline.

But while language provides information, and thus may impact stereotypes and biases, the language embedded in cultural products (e.g., songs, books, and movies) should be particularly impactful. Culture is often conceived of cultural background, and the particular country or social group in which one is enculturated has an important impact on attitudes, cognitions, and behaviors (Markus & Kitayama, 1991). Beyond cross-cultural differences, though, the norms, practices, and items that make up collective culture (i.e., cultural products; Morling & Lamoreaux, 2008) also play an important role (Kashima, 2008).

Cultural products like music, books, and movies should be especially influential given their reach. While one person's personal experience may be recounted a few times, or transmitted to other friends, information transmitted through a book or song, particularly a popular one, should reach a much larger audience. This, in turn, should increase the impact of the stereotypes, norms, and norms contained. Research on violent media, for example, has shown that it has a causal impact on aggression (e.g., Anderson & Bushman, 2002; Berkowitz, 1993; Bushman & Anderson, 2002). Consequently, cultural products can be an important carrier of stereotypes and biases.

But while it is clear that cultural items can shape attitudes and action, their actual nature is less clear. Take song lyrics, for example. Are they actually biased against women, and if so, have any such biases changed over time?

## Gender Bias in Music

Researchers and cultural critics alike have long debated whether music is misogynistic or biased against women. The music industry is dominated by men (e.g., only around 2% of producers are women, S. L. Smith et al., 2019), and many have suggested that this has contributed to a sexist environment that shapes the content produced (e.g., Harding & Nett, 1984).

Further, though some have argued that rap music paints a negative picture of women (e.g., Adams & Fuller, 2006; Herd, 2009), the same arguments have previously been made for other genres. In the mid-1980s, for example, Harding and Nett (1984) argued that "Rock music is the most misogynistic and aggressive form of music" (p. 60) and similar complaints have been leveled at country music (see Evans, 2014). More generally, looking across genres, authors have argued that songs objectify women (i.e., focusing on their bodies, body parts, or provocative dress, Flynn et al., 2016) or contain some sort of female stereotype (e.g., women as evil or as sex objects, Cooper, 1985).

But while researchers across psychology, sociology, and a range of other disciplines have long theorized about gender bias in music, actually quantifying such potential biases have been hampered by issues of scale and measurement. Most perspectives are based on small samples or one genre over a short period of time (Adams & Fuller, 2006). Harding and Nett (1984), for example, examined 40 rock songs and Armstrong (2001) examined 13 rap artists. While a couple papers have examined slightly larger samples in one genre (e.g., 340 rap songs, Herd, 2009) or cross-genre samples over a short time period (e.g., 600 songs over 4 years, Flynn et al., 2016) without more comprehensive data, it is difficult to draw strong conclusions. One genre may not be representative of music as a whole, and without examining longer time horizons, it is hard to

know whether bias has shifted over time. Truly examining gender bias in music requires analyzing tens of thousands of songs over multiple genres across multiple decades.

Further, even if one were able to compile such a dataset, measuring gender bias would be challenging. Most work has identified gender bias through manual coding. This involves the authors (or research assistants) personally examining each song's lyrics and noting whether it appeared biased. But while this is easy for a few dozen songs, and possible for even a few hundred, manually rating tens of thousands of song lyrics would be unduly time consuming.

Relying on human judgment also makes such analyses susceptible to bias. One can instruct coders on the key dimensions of interest, train them of a test set, and give them feedback to ensure they are reliable. But biases can still leak in. The same lyrics, for example, may seem more or less misogynistic depending on the gender of the coder reading them.

## The Current Research

To address these challenges, we use natural language processing. Rather than having individuals read song lyrics, and manually rate them based on how biased they seem, we use automated textual analysis.

Prior work sometimes measures gender bias using aggression toward women (Fischer & Greitemeyer, 2006), so one could measure how frequently women are the object of aggression. But while we test this possibility, note that even misogynists might not explicitly sing about wanting to hurt women, for fear that it would restrict their audience. More generally, if gender bias does exist it may be more implicit.

Consequently, we use an emerging computational linguistics approach to measure a subtler form of bias. Not whether lyrics are explicitly aggressive toward women, but whether woman are less likely to be linked with desirable traits (e.g., competence).

As discussed in the methods section, word embedding is a natural language processing technique that allows words to be represented based on their meaning (Devlin et al., 2018; Mikolov et al., 2013; Pennington et al., 2014). This machine learning framework represents each word by a vector and uses a high-dimensional space to map each word or entity based on the other words with which it frequently appears. The relationship between these vectors captures the semantic relationship between words and can be used to measure whether certain words are associated with one gender more than the other. This method has been used to quantify everything from whether gender prejudice occurs more in gendered languages (DeFranza et al., 2020) to whether gender stereotypes are reflected in the large-scale distributional structure of natural language semantics (Lewis & Lupyan, 2020).

We leverage this approach to examine gender bias over time. We collect more than a quarter of a million songs from six main music genres over more than 50 years. Then, by separately analyzing lyrics from each 5-year time period, we examine whether and how the association between women and desirable traits changes over time.

Analyzing things over time is particularly important because it provides potential insight into where things might move in the future. While recent lyrics might show evidence of bias or not, without knowing the prior trajectory, it is hard to estimate how things might evolve moving forward. Are any biases decreasing, for example, or are they increasing? Consequently, we examine both whether recent lyrics are biased, and whether such biases have changed.

# Method

## Data

First, we compiled data on songs and their lyrics. Given copyright concerns, there are limited open-access lyrics datasets, so we compiled information from different sources. We started by extracting all songs on each major Billboard chart (i.e., pop, rock, country, dance, rap, and r&b) every 3 months from 1965 (or whenever a given chart started) to 2018. This was then paired with lyrics from SongLyrics.com. To collect additional songs, we gathered all songs and their lyrics from datasets on kaggle.com[1] and scraped all available songs and their lyrics from a major lyrics website. Then, we used the million songs dataset (Mahieux et al., 2011) and freeDB (FreeDB, 2018) to append year and genre for any song that did not already include this information. We removed any song with missing information and used artist and title combination to remove duplicates in the data.

The final dataset includes 258,937 songs from 1965 to 2018. It includes pop, rock, country, r&b, dance, electronic, rap, and hip-hop genres. To ensure enough data for each genre in as many time points as possible, we combined dance and electronic music, and rap and hip-hop.

Second, we cleaned the lyrics. Given the focus on English lyrics, we used the langdetect package (Version 1.0.7) in Python to remove any non-English lyrics. We also removed non-informative texts in brackets, such as Verse 1 or Pre-Chorus 1. Finally, we shifted all lyrics to lower case so that variations of the same word are treated similarly (e.g., Woman and woman).

## Transparency and Openness

Our data and codes are publicly available at https://osf.io/kgjea/?view_only=a714f62bbdb2433dae873551db2a18e1. Lyrics from the kaggle datasets are publicly available already, but given we do not own the copyrights, we cannot make the raw lyrics from the major lyrics website publicly available. That said, we provide the word embedding scores for competence and other dimensions, both within genre and across genre data, Python codes for training the models, and the R codes for statistical analyses.

## Measuring Aggression

Prior work often measures bias using aggression toward women (Fischer & Greitemeyer, 2006), so we begin by testing how frequently women are the object of aggression using a violence word list (see the online supplemental materials). Such analyses suggest (a) that women and men are recipients of explicit aggression and violence a similar amount and (b) that this difference has not changed over time (see the online supplemental materials). That said, note that even misogynists might not explicitly sing about wanting to

[1] https://www.kaggle.com/artimous/every-song-you-have-heard-almost; https://www.kaggle.com/mousehead/songlyrics; https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

hurt women, for the fear that it would restrict their audience. More generally, if gender bias does exist it may be more implicit. Indeed, scholars have long argued that stereotypes are maintained through subtle or indirect language (Charlesworth et al., 2021; Durkheim, 2009; Moscovici, 2001).

## Measuring Subtler Bias

To address these limitations, as noted previously, we use a state-of-the-art machine learning approach to examine a subtler form of bias. Not whether lyrics are explicitly aggressive toward women, but whether woman are less likely to be linked with desirable traits (i.e., competence).

Competence and warmth are two universal dimensions of social cognition (Fiske et al., 2007), but while women are often described as warm (e.g., kind and supportive), they are less often described as competent (e.g., smart and ambitious).

We use word2vec (Mikolov et al., 2013), a well-adopted word vector representation technique, to quantify how likely women (relative to men) are to be linked to competence, as well as other traits (i.e., intelligence, warmth, masculine stereotypes, and feminine stereotypes), over time (see Berger & Packard, 2022 for a review).[2]

Word2vec is a two-layer neural network which receives a corpus of text and transforms it to numerical vectors. This approach assigns each word a high-dimensional vector such that the relationship between vectors captures the semantic relationship between the words. Words that relate to one another such as fruit (e.g., apple and orange) or vehicles (e.g., car and truck) appear close together, but words that are not as related (e.g., apple and truck) appear further apart.

To position words, the approach considers words co-occurrence, distance, and occurrence in similar contexts. If "men are smart" shows up much more frequently than "women are smart," for example, that would increase the relative similarity (i.e., shrink the distance) between men and smart. The distance between occurrences of words also matters. Even though both "women are smart" and "women do great research and are also smart" contain the words women and smart, the first example has them closer together, which increases the similarity in word2vec space. Occurring in similar contexts also shapes similarity. If "women are scientists" and "smart people are scientists" both appear frequently, it increases the similarity between "women" and "smart" even if the two words never appear together.

The similarity between word2vec vectors has been shown to be a powerful tool for studying language, perceptions, and stereotypes. Consequently, recent work has begun to use similar methods to measure stereotypes. Embedding associations captured human ratings of whether target words were more associated with women or men ($r > .76, p < .001$; Kozlowski et al., 2019), for example, and tracked the percentage of different genders in different occupations over time ($r = .70, p < .001$; Garg et al., 2018). This can be seen as analogous to an implicit association test (IAT), where people demonstrate faster response time when asked to pair similar (than dissimilar) concepts (Caliskan et al., 2017).

To measure gender bias in music, we quantify the relationship between words related to each gender, and key dimension of interest (i.e., competence), over time. We used word lists from prior work (Garg et al., 2018) to identify words related to women (e.g., "woman," "women," and "she") and men (e.g., "man," "men," and "he," see Table A1 in the online supplemental materials for full list). Then, we use word2vec to train a separate word embedding model per time-bucket and generate 300-dimensional real-valued word vectors. That means for each word (e.g., woman) and each period in the dataset (e.g., 1965–1970), we use the set of song lyrics over that period to position the vector for that word. Next, we take the word vector representation of words in our gender word lists and average the vectors to obtain one single 300-dimensional vector per each gender.

We use a similar approach to obtain vectors for key dimensions of interest (e.g., competence). In the case of competence, for example, we use a list of word from prior work (Nicolas et al., 2021) that are known to relate to competence (e.g., smart, persistent, and knowledgeable, see Table A1 in the online supplemental materials for more examples).

Then, to measure the association, we capture the similarity between gender vectors and key dimension vectors (e.g., competence) at any given period in time. We use cosine similarity metric because it is frequently used in the literature and in the original word2vec article (Mikolov et al., 2013). If a gender word and a competence word occur in similar contexts, for example, their vector representations are close in the word2vec space, which results in a larger cosine similarity score and indicates that the two words are highly associated. Conversely, if the words do not occur in similar contexts, the cosine similarity is smaller, and the words are considered less associated.

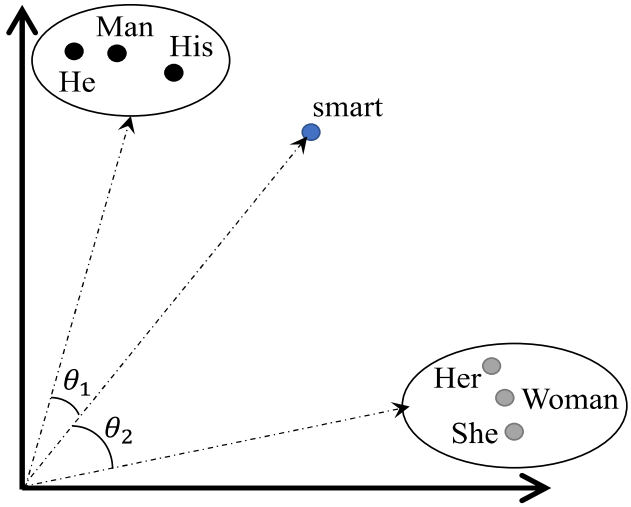
To give a sense of how this approach works, Figure 1 provides a simplified version (words are represented with 300-dimensional vectors in our analyses). As shown in the figure, semantically similar words (e.g., the group related to women or the group related to men) appear close to one another. As discussed, we create a single vector for each gender and compare how close it is to words representing key dimension of interest (e.g., the vector for the word "smart" which relates to competence). The figure shows that compared to the female vector, the male vector is closer to "smart," indicating that, in this example, smart is more associated with men.

Formally, to calculate similarity of trait $t$ (e.g., competence) to gender $g$, we calculate the cosine similarity of each word vector in $W_t$ to gender vector $V_g$, given that $W_t$ is a matrix of $n$ by 300 where $n$ is the number of words in trait $t$ (e.g., competence) and 300 is the vector size. To calculate how biased trait $t$ (e.g., competence) is, for each trait word (e.g., smart), we subtract cosine similarity between the trait word vector and the vector representing female from the cosine similarity between the trait word vector and the vector representing male. Positive values indicate that trait word is more associated with males, negative values indicate the opposite. This difference score also removes any general time trends that affect both genders equally (e.g., maybe more recent songs talk about

[2] We tested several other well-known methods on a random small subset of our data and found that word2vec was most appropriate. We compared truncated Singular Value Decomposition (a.k.a. latent semantic allocation), Positive Pointwise Mutual Information, GloVe (Pennington et al., 2014), and word2vec. The preliminary results and consistency of word2vec results along with previous research (Sahlgren & Lenci, 2016) showed that word2vec using Continuous Bag Of Words is the appropriate method for our question.

**Figure 1**

*Simplified Illustration of Word Embeddings in Two Dimensions*




*Note.* Semantically similar words (e.g., the group related to women) appear close to one another. Because $\theta_1$ is less than $\theta_2$, cosine similarity of the male vector with the smart vector is larger than female vector with smart. This indicates that, compared to the female vector, the male vector is closer to smart, suggesting that smart is more associated with men. See the online article for the color version of this figure.

competence more in general):

$$t \in \{\text{competence}\}$$

$$\overline{\text{bias}_t} = \frac{1}{n}\sum_{i=1}^{n}\Big(\text{cosine\_similarity}\Big(\overrightarrow{W_{ti}}, \overrightarrow{V_{\text{male}}}\Big) - \text{cosine\_similarity}\Big(\overrightarrow{W_{ti}}, \overrightarrow{V_{\text{female}}}\Big)\Big)$$

In addition to prior work, further validation tests (see the online supplemental materials) demonstrate that such difference scores accurately capture gender biases.

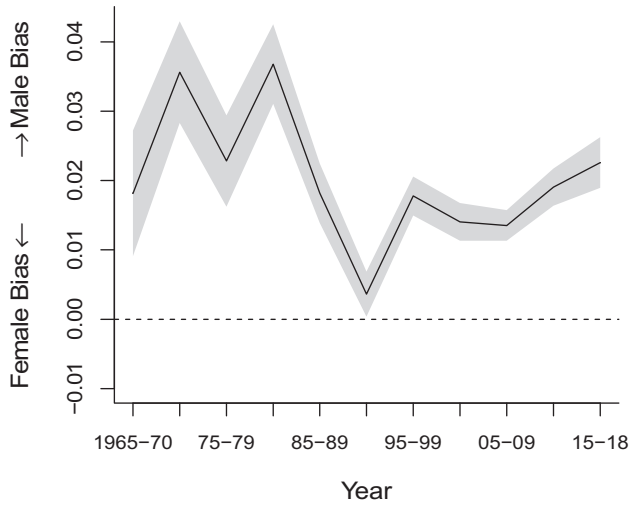
We quantify the association for each time period in the dataset (each contains over a half million words), and then analyze whether the associations change over time. We use linear mixed-effect models with time as the independent variable and bias as the dependent variable. Given association with gender may vary across words, we add a random effect for words. For ancillary analyses, we also run separate linear mixed-effect models for each genre (e.g., rock).

## Results

Results suggest that lyrics have become less biased against women, but remain biased (Figure 2). Words related to competence (e.g., smart) have become more associated with women, from strongly biased toward men to less so, $\beta = -0.03$, $SE = 0.01$, $t(5{,}827.89) = -2.95$, $p < .01$. That said, the gains level off, and potentially even reversed in the late 1990s, $\beta = -0.21$, $SE = 0.04$, $t(5{,}673.88) = -4.93$, $p < .001$, $\beta^2 = 0.17$, $SE = 0.04$, $t(5{,}625.28) = 4.30$, $p < .001$. Further, the confidence intervals do not overlap with zero in recent years suggesting that lyrics still associate competence with men.[3] Words related to intelligence (e.g., precocious and inquisitive, Garg et al., 2018) show similar results, $\beta = -1.56$, $SE = 0.34$, $t(120.28) = -4.63$, $p < .001$, $\beta^2 = 1.34$, $SE = 0.31$, $t(118.21) = 4.32$, $p < .001$ (Figure A2 in the online supplemental materials).

**Figure 2**

*Gender Bias in Lyrics Over Time*




*Note.* Dashed line represents equal association with either gender and values greater (less) than 0 indicate the trait is more associated with men (women). Grey regions represent 95% confidence intervals around the estimates for each period.

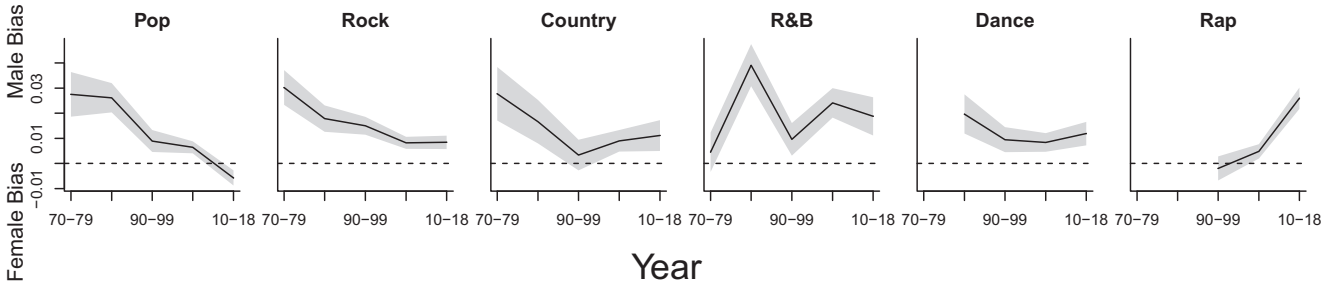
## Within Genre Analysis

Note that the leveling off is not driven solely by the introduction of new genres. Examining different genres separately (Figure 3) show that while the introduction of rap and dance in the 1980s contributed to the leveling off, similar patterns hold for genres that existed previously (i.e., pop, rock, and country, see the online supplemental materials). More generally, ancillary analyses suggest that while competence has become more associated with women in pop, $\beta = -0.14$, $SE = 0.02$, $t(2{,}033.37) = -6.67$, $p < .001$; rock, $\beta = -0.08$, $SE = 0.02$, $t(2{,}642.50) = -5.07$, $p < .001$); and country, $\beta = -0.05$, $SE = 0.02$, $t(1{,}166.33) = -2.14$, $p = .03$,

[3] While these results demonstrate that women are less associated with competence in song lyrics, one might wonder what this effect actually means. For example, is it mere collocation (e.g., "she likes dumb guys") or description (e.g., "she is dumb")? Prior work has demonstrated that embeddings are a useful means to measure bias (Garg et al., 2018; Kozlowski et al., 2019), but to examine how women are being talked about differently than men, we applied a dependency parser algorithm on sentences which included either a female or male word as the subject and a competent or incompetent word. Dependency parsers label the role of each word in a sentence, such as subject or adverb modifier. Then, we examined the most prevalent combination of dependency tags between female/male and competent/incompetent words. The top combinations and their prevalence were similar across genders, except for the adjectival complement role (e.g., object of the verb or "She is _____") for incompetent words which appears twice as frequently when female words are the subject. Examples include sentences such as "she was **afraid** that somebody would see" or "she's just as **fragile** as a child." Further, note that this relation is the most common one for female and incompetent words overall. This suggests that one reason women may be less associated with competence is that they are more likely to be described using incompetent words.

**Figure 3**

*Within Genre Gender Association with Competence*




*Note.* Dashed line represents equal association with each gender and values greater (less) than 0 indicate the trait is more associated with men (women). Grey regions represent 95% confidence intervals around the estimates for each period.

such associations have not changed in R&B, $\beta = 0.009$, $SE = 0.02$, $t(1{,}258.60) = 0.38$, $p = .70$ and dance, $\beta = -0.02$, $SE = 0.03$, $t(1{,}283.27) = -0.92$, $p = .36$. Competence has become less associated with women in rap lyrics, $\beta = 0.12$, $SE = 0.02$, $t(1{,}351.89) = 6.23$, $p < .001$, but women were relatively more associated with competence in rap lyrics than other genres to begin with, so the decrease may be partially due to the fact that women were more associated with competence there initially.

## Other Dimensions

For completeness, we also analyze warmth related words (e.g., kind, friendly, and caring; Nicolas et al., 2021, see Table A2 in the online supplemental materials for more examples). Results indicate that warmth has become less associated with women, $\beta = 0.17$, $SE = 0.03$, $t(10{,}382.90) = 5.26$, $p < .001$, $\beta^2 = -0.12$, $SE = 0.03$, $t(10{,}305.62) = -4.05$, $p < .001$.

Further analysis of words associated with masculine stereotypes (e.g., leader and ambitious) and feminine stereotypes (e.g., cooperate and trustworthy; Gaucher et al., 2011) suggests that lyrics have become somewhat less gendered overall. Words related to masculine stereotypes have become less associated with men, $\beta = -0.23$, $SE = 0.08$, $t(163.35) = -2.86$, $p < .01$ and words related to feminine stereotypes have become less associated with women, $\beta = 0.14$, $SE = 0.05$, $t(161.70) = 2.58$, $p < .05$. See the online supplemental materials for more details.

## Magnitude of the Effects

To provide a sense of the magnitude of these effects, following prior work (Charlesworth et al., 2021; DeFranza et al., 2020; Kurdi et al., 2019), we perform the Single Category Word-Embedding Association Test (SC-WEAT) which computes a standardized effect-size measure for bias in word embedding space. We computed the bias score for each word following the process in the Measuring Subtler Bias section, calculated the average and standard deviation of the bias scores, and divided the average by the standard deviation to get the effect size, or SC-WEAT D score. Similar to a single-category IAT D score, this is a difference normalized by the standard deviation.

Results suggest that the effects observed here are similar in magnitude to those observed in prior work on language and bias. Looking at the average association between competence and gender over time, for example, shows an SC-WEAT D score of 0.20, which is similar to the average meta-analytic estimate (0.29). Charlesworth et al. (2021) report across various contexts (e.g., books and produced speech) and audiences (e.g., children and adults, see their table S3). Lyrics show a similar (albeit slightly larger) effect size for the association between women and warmth ($D = -0.24$), a smaller effect size for the association between women and female stereotypes ($D = -0.10$) and a larger effect size for the association between men and intelligence ($D = 0.56$) and masculine stereotypes ($D = 0.97$).

## Drivers of Lyrical Shifts

One might wonder what is driving the observed changes. Are the linguistic shifts similar for male and female artists, for example? Or might they simply reflect an increase in the number of female artists, who may use less misogynistic and gendered language?

To begin to address these questions, we analyze artist gender. Building on prior research (Santamaría & Mihaljević, 2018; Squire, 2019), we use the gender_guesser package (Version 0.4.0) in Python along with national name list (Kaggle, 2017) to infer an artist's gender. For individual artists (e.g., Katy Perry or Billy Joel) we simply use their name, but for artists that go by pseudonyms (e.g., Snoop Dogg) or bands that have multiple members (e.g., Metallica or the Supremes) we first use their Wikipedia page to extract individual name(s). In cases where no such page is available, research assistants searched for this information manually. Then, we use a similar approach as individual artists to detect gender and assign the band a score based on the percentage of members that are female. Gender was able to be assigned for 99.5% of artists.[4]

Results cast doubt on the notion that the decrease in gender bias is driven by the introduction of more female artists. There is no significant change in the percentage of female artists over time, $\beta = -0.08$, $SE = 0.33$, $t(9) = -0.23$, $p = .82$. Further, while one could argue that this is driven by new genres emerging, even within genres there is no significant change, $\beta_{pop} = 0.85$, $p = .06$, all other genres

[4] Note that there are three to four times more male artists in the dataset, so the effects for male language may be more precisely estimated. Further, while there are enough solo female artists or all female groups, and solo male artists or all male groups to estimate effects for each, it is more difficult to do so for mixed gender bands. Less than 10% of songs were performed by mixed gender groups which provides limited data for estimating language changes.

$p > .2$. This casts doubt on the possibility that language shifts are driven by an increased number female artists over time.

Instead, results are more consistent with a shift in male artists' language (Figure 4). For male artists, words related to competence become more associated with women, $\beta = -0.06$, $SE = 0.01$, $t(2{,}639.25) = -4.07$, $p < .001$. Female artists' language shows less evidence of change, $\beta = 0.02$, $SE = 0.02$, $t(2{,}030.18) = 0.93$, $p = .35$, likely because it was less biased against women to begin with. Female artists are more likely to associate women with competence overall, $\beta = -0.09$, $SE = 0.01$, $t(5{,}066.11) = -8.08$, $p < .001$, and the Gender × Time interaction, $\beta = 0.04$, $SE = 0.01$, $t(4{,}995.62) = 3.35$, $p < .001$ indicates that the gender difference has decreased over time as male artists started associating women more with competence. The fact that the confidence intervals do not overlap in the most recent time period, however, suggests that artists of both genders remain biased (i.e., lyrics still associate men more with competence).

## Robustness Tests

We conducted several robustness checks to test whether alternative explanations can explain the results. In all cases, conclusions remained the same.

### *Words Used*

One could wonder whether the results might be driven by the particular words used. We test this a few ways. First, we only consider words which have occurred in every time point. This guarantees similar predictors (i.e., words) are compared over the years. Conclusions remain the same. Words related to competence become less associated with men, $\beta = -0.03$, $SE = 0.01$, $t(2{,}149.00) = -2.07$, $p = .04$, with a slight reversal in recent years, $\beta = -0.13$, $SE = 0.06$, $t(2{,}148.00) = -2.43$, $p = .02$, $\beta^2 = 0.11$, $SE = 0.06$, $t(2{,}148.00) = 1.94$, $p = .05$.

Second, following prior research (Mikolov et al., 2013), we drop words with frequency smaller than five in each time bucket and repeat our analyses. Conclusions remain similar. Words related to competence become directionally less associated with men, $\beta = -0.02$, $SE = 0.01$, $t(3{,}049.02) = -1.61$, $p = .11$, with a slight reversal in recent years, $\beta = -0.20$, $SE = 0.05$, $t(3{,}011.54) = -4.08$, $p < .001$, $\beta^2 = 0.19$, $SE = 0.05$, $t(2{,}983.51) = 3.78$, $p < .001$.

Third, we include proper names. Rather than talk about men or women using words like he or she, some songs contain people's names (e.g., James, Jack, Angie, or Jenny) instead. To test whether this might impact the results, we used the Named Entity Recognition algorithm in Python to detect people's names and used the gender-guesser Python package to assign a female or male gender to each name. Then, we replaced names with their corresponding gender pronoun and re-ran the models. Results remain the same. Competence words are more associated with males, for example, but less so over time—linear: $\beta = -0.03$, $SE = 0.01$, $t(5{,}908.86) = -2.91$, $p = .004$; linear and quadratic: $\beta = -0.21$, $SE = 0.04$, $t(5{,}742.53) = -4.90$, $p < .001$, $\beta^2 = 0.17$, $SE = 0.04$, $t(5{,}691.28) = 4.28$, $p < .001$.

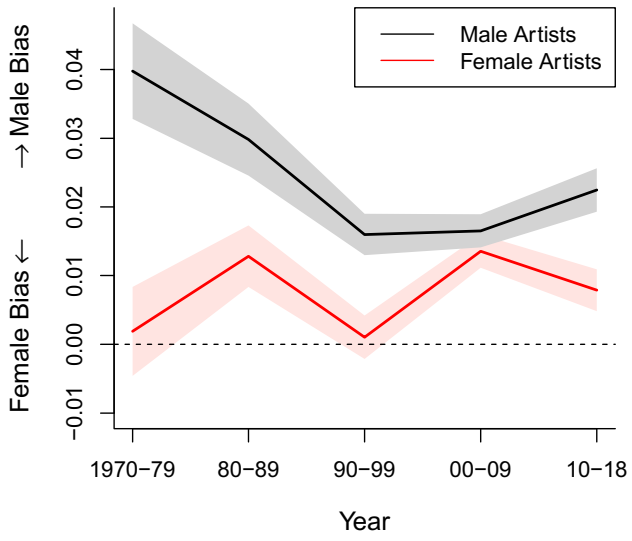
### *Songs Used*

Alternatively, one could wonder whether the results are somehow driven by the genre weighting used. If the dataset included twice as many rock songs as pop songs, for example, then rock music would have a disproportionate impact on the results. We address this in two ways.

To ensure equal representation from each genre, our main analyses use under-sampling, a common machine learning practice when datasets include imbalanced class labels. Assume data from three genres $g_1$, $g_2$, and $g_3$, where $g_1$ has the smallest number of songs. To generate a balanced sample for the cross-genre analyses, we randomly select $\|g_1\|$ songs from genres $g_2$ and $g_3$ and perform the analyses on that sample. We repeat this sampling and analysis 100 times, averaging across the runs. Substantial number of dance and rap songs do not enter the data until 1980 and 1990, respectively, so we only include them (and sample accordingly) after those time points.

That said, one could argue that certain genres are less popular and so equal weighting does not actually reflect reality. To test this possibility, we re-run the model weighting the genres based on popularity. Billboard Hot 100 charts are the only consistent audience consumption data available from the 1960s to today, so we rely on genre popularity scores collected from Billboard by Beckwith (http://thedataface.com/2016/09/culture/genre-lifecycles), indicating the percentage of songs in Hot 100 charts from each genre in each time period. We weight things accordingly. In 2000, for example, imagine country, rock, and pop songs occupy 25%, 25%, and 50% of the Billboard charts, respectively. If there are 1,000 songs from each genre in the dataset, we select 1,000 pop songs, and 500 songs from rock and country genres (total of 2,000 songs). The number of songs for 1965–1969 time bucket did not meet our threshold of half a million words and was dropped for the genre popularity analysis.

**Figure 4**
*Gender Bias in Lyrics Over Time for Female and Male Music Artists*




*Note.* Dashed line represents equal association with either gender and values greater (less) than 0 indicate the trait is more associated with men (women). Grey regions represent 95% confidence intervals around the estimates for each period. See the online article for the color version of this figure.

Conclusions remain the same. Words related to competence, $\beta = -0.04$, $SE = 0.01$, $t(5{,}711.85) = -3.37$, $p < .001$, become less associated with men over time, though things have slightly reversed as of late, $\beta = -0.05$, $SE = 0.01$, $t(5{,}731.16) = -3.85$, $p < .001$, $\beta^2 = 0.03$, $SE = 0.01$, $t(5{,}514.23) = 2.45$, $p = 0.01$.[5]

### *Complexity*

One could also wonder whether the results simply reflect the increasing complexity of song lyrics. This argument would suggest that while song lyrics in the 1970s were rather simplistic (e.g., "let it be"), today's lyrics may be much more complex and varied and this would somehow lead to our effect.

But while such an argument might suggest that the increased complexity would lead to an increased distance between men and competence words, and women and competence words, it says less about why the *relative* distance between genders and competence words would decrease. Said another way, it would suggest that the distance between both male and female words and competence would increase, but complexity alone has difficulty explaining why the increase would be greater for one gender than the other.

Further, results persist when controlling for language complexity over time. To control for the possibility that lyrics may have become more complex, we measure how the distance amongst words within each of the gender word lists (e.g., words related to men like "man" and "he") has changed. This captures whether the similarity of even related words ("man" and "he," for example, or "woman" and "she") has shifted over time. We then take the average of the similarity among words related to men (e.g., "man," "he," and "his") and similarity among words related to women (e.g., "woman," "she," and "her") and normalize our main competence measure by this average. The conclusions remain the same. Competence words have become more associated with women over time, though they remain more associated with men, $\beta = -0.03$, $SE = 0.01$, $t(5{,}827.89) = -2.95$, $p = .003$. Further, these gains have leveled off and even reversed as of late, $\beta = -0.21$, $SE = 0.04$, $t(5{,}673.88) = -4.93$, $p < .001$, $\beta^2 = 0.17$, $SE = 0.04$, $t(5{,}625.28) = 4.30$, $p < .001$.

## Potential Consequences

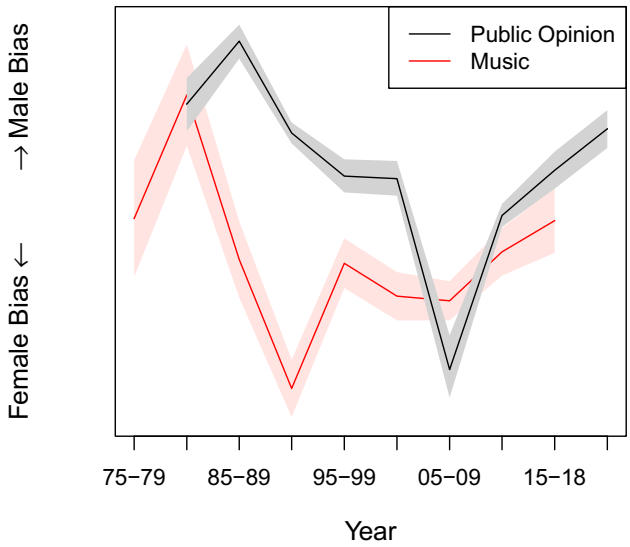
One might wonder whether these lyrical shifts have any consequences. Might they change social stereotypes (i.e., how the population at large sees women and men), for example, or do they simply reflect changes that have already occurred? Said another way, do shifts in lyrics *precede* shifts in societal stereotypes (suggesting they cause or are a leading indicator of societal change), or do they tend to *follow* them, indicating that music merely reflects what has already occurred in society at large?

Inferring causality from observational data is difficult, but to begin to explore these possibilities, we leverage prior work aggregating all available public opinion polls from 1946 to 2018 measuring gender stereotypes related to competence (Eagly et al., 2019). These polls asked respondents whether certain characteristics (e.g., competence) was more true of women, men, or equally true of both and the number of respondents that choose women, divided by the number that selected either women or men, captures gender stereotypes over time (Eagly et al., 2019).

Results indicate that gender bias in lyrics is strongly correlated with societal stereotypes, $r(7) = .62$, $p = .07$, Figure 5.

**Figure 5**

*Public Opinion Polls and Gender Bias in Lyrics Over Time*




*Note.* Values for public opinion polls and lyrics are scaled so that they are comparable and can be visualized in one plot. Larger numbers mean more association with men and smaller numbers mean more association with women. See the online article for the color version of this figure.

Importantly, this relationship does not simply reflect two time series moving in the same direction. Lyrical changes are most strongly related to *subsequent* public opinion (i.e., opinion polls that directly follow the time period of lyrics). Examining lyrics and *simultaneous* public opinion (i.e., opinion polls that occur in the same time period as the lyrics) or *preceding* public opinion (i.e., opinion polls that appear in the time period before the lyrics) show weaker or reversed relationships, that is, $r(6) = .26$, $p = .5$ and $r(5) = -.60$, $p = .2$, respectively.

These results are consistent with the notion that changes in lyrics may contribute to, or be a leading indicator of, shifts in stereotypes held by the population, but other factors may also be at work. Some third variable (e.g., the content of television shows, books, or news) could be driving both the content of lyrics and public opinion. Such variables would have to impact song lyrics more quickly than public opinion to explain the pattern of results, but we cannot rule out this possibility. That said, at the very least, these results suggest that lyrical changes are a leading indicator of shifts in public opinion. This suggests that one can look to the content of lyrics as a guide to what may soon in culture more generally. Future work should examine these possibilities in greater detail.

## Discussion

Researchers and cultural critics alike have long debated whether popular music is misogynistic. But questions of scale (e.g.,

[5] Alternatively, one could wonder whether, given that more popular songs may be more likely to have their lyrics available, the dataset, especially in earlier years, may include more popular songs. That said, because more popular songs are listened to more anyway, they should have a greater impact on the population.

examining only one genre over a few years) and measurement have limited the ability to draw strong conclusions. Consequently, it remains unclear whether lyrics are biased against women, whether this has shifted across time, what may have driven any shifts, and how such shifts relate to public opinion.

Analyzing over a quarter of a million songs over more than 50 years sheds light on these questions. First, while there is little evidence of explicit gender bias (e.g., greater violence toward women), other approaches suggest a more complex picture. Lyrics have become less biased against women over time, but bias remains. Women are more likely to be associated with competence and intelligence than they were in the past, but these concepts are still more associated with men. Lyrics have become less gendered more broadly, but they remain gendered. Further, reductions in bias seem to have slowed and may even have reversed in some cases. The results not only provide empirical evidence, but also allow for quantitative comparison of where (i.e., which genres and artists) and when (i.e., time periods) biases may be larger.

Second, ancillary analyses suggest potential consequences of lyrical changes (e.g., on social stereotypes). Lyrical shifts are strongly correlated with subsequent changes in public perceptions. These results suggest that lyrics are at least a leading indicator of broader cultural changes. Given their widespread reach, song lyrics can diffuse perceptions of different groups or identities, and potentially impact the social stereotypes help by a population of listeners. This, in turn, may impact downstream attitudes and behaviors.

Third, rather than being driven by more female artists, these results appear to be driven by male artists shifting their language over time. Further, they are driven by both the introduction of new genres (e.g., rap) and changes in older ones (e.g., country).

## Implications and Future Research

These findings have several implications. First, they speak to the ongoing debate about gender bias in music. They not only provide empirical evidence, but allow for quantitative comparison of where (i.e., which genres) and when (i.e., time periods) such biases may be larger. They suggest that while some things are moving in the right direction, bias persists. Further, they suggest tools researchers in this area can use to quantify features of interest.

Second, even outside of music, the results highlight the importance of considering subtler measures of bias. Across disciplines, there is increased interest in more implicit measures of attitudes and biases (Payne et al., 2019) and language provides another useful approach (Kesebir, 2017). Like a fingerprint, words reflect things about the people, organizations, cultures, and contexts that generate them (Berger & Packard, 2022; Pennebaker, 2011). Consequently, language can be a useful barometer in a wide range of domains. Future work might examine whether the language used in films, for example, has changed over time, and whether there have been shifts in the way women and minorities are discussed. Language reflects social stereotypes, and, as a result, the text produced in different time periods can be a valuable method to understand social opinions over time.

Third, word embeddings and other natural language processing techniques provide a powerful toolkit to study people and culture. Researchers have long been interested in quantifying cultural dynamics (Kashima, 2008), but measurement has been a key challenge. Natural language processing, however, provides a reliable method of extracting features, and doing so at scale (Jackson et al., 2019, 2022). While more and more researchers are starting to use dictionary methods (e.g., Linguistic Inquiry and Word Count, Pennebaker et al., 2015) recent innovations from computer science and statistics like word embeddings and topic modeling (e.g., Latent Dirichlet Allocation, Blei et al., 2003) are just starting to receive more attention.

These emerging approaches have the potential to shed light on a range of interesting questions (Berger et al., 2023; Berger & Packard, 2022; Boghrati et al., 2023). Take the ongoing debate about the universality of emotions. While some (e.g., Ekman, 1992) have argued that there are a discrete number of universal emotions that are consistent across cultures (e.g., anger, sadness, and disgust), others have suggested that the meaning of emotions are more malleable (Jack et al., 2012). By examining language corpora from different cultures, researchers could use word embeddings to test whether the same words really have similar semantic meanings (see Jackson et al., 2019). Whether anger is used to talk about the same things in Spanish, for example, as it is in English.

Researchers studying goals, motivation, and self-regulation may also find word embeddings useful. Work on desire and desire regulation (Hofmann et al., 2012), for example, notes that "the majority of research on self-regulation occurs in the laboratory" and that "little is known about what types of urges are felt strongly (or only weakly), what urges conflict with other goals, and how successfully people resist their urges" (p. 582). Word embeddings, topic modeling, and similar approaches can allow researchers to examine these questions in more naturalistic settings. Social media, blog posts, and even the Google Books dataset could be used to examine the urges people have, how they conflict, and even how these urges may have varied over time. What people want and feel like they should do today (e.g., balance work and family or consume less social media), for example, is likely very different today than it was 20 years ago.

Related approaches may also be valuable for looking *within* cultural items. Researchers have long theorized about the nature of narrative. Books, movies, and songs, but also online content, speeches, and even academic papers usually have beginnings, middles, and ends, through which there is some sort of narrative flow. Why are some narratives more engaging and impactful than others?

One possibility is how the plot or narrative develops or unfolds. Some narratives move very quickly, while others plod along at a slower pace. Some repeatedly tread new ground while others circle back to the same themes again and again. Natural language processing can be used to measure the geometry of narrative and whether it shapes success (Laurino Dos Santos & Berger, 2022; Toubia et al., 2021).

Word embeddings examine the semantic relationship between words, but related approaches have applied these ideas to larger chunks of linguistic information. Just as word2vec represents words as a vector in a multidimensional space, techniques like Doc2vec (Le & Mikolov, 2014) and USE (Cer et al., 2018), do something similar for sentences or larger chunks of text. This can allow researchers to test, for example, whether narratives that move at a faster pace (i.e., adjoining chunks are more distant from one another) are more or less successful (Laurino Dos Santos & Berger, 2022; Toubia et al., 2021).

Research could also examine the relationship *between* cultural items. Whether certain songs, baby names, or other cultural items

become popular depends not only on the features of those items themselves, but how similar those items are to the larger cultural context in which they are embedded. Songs whose lyrics are more atypical, or more differentiated from their genre, for example, are more popular on the Billboard charts (Berger & Packard, 2018). Similar results might hold for movies or even academic papers. Papers whose themes are quite differentiated from the journals they are published in, for example, may be more novel and thus cited more. Alternatively, one could argue that if papers are too different, they may be cited less. Topic modeling approaches (e.g., Blei et al., 2003) may be even more useful than word embeddings here, because they can capture underlying thematic content of cultural items.

More generally, these methods can be a nice complement to traditional experimental approaches. While carefully controlled experiments are nice for demonstrating causality and measuring potential processes, given small samples and a limited range of stimuli, they are often existence proofs, showing what can happen rather than providing a sense of what actually does, or how frequently it occurs. Consequently, the methods described here, as well as other natural language processing technique, can provide a useful complement. By allowing researchers to parse large and diverse field datasets, and quantify features of interest, they increase external validity and help theory construction, validation, and reproducibility across the social sciences.

## Limitations and Constraints on Generality

As with any paper, this work has some limitations. One could wonder whether the dataset is representative of the music people are actually exposed to. The fact that the results hold whether we equally weight each genre, or weight them by popularity, casts doubt on the notion that genre weighting is driving the results. At a song level, by collecting the Billboard charts each year, we certainly have a representative sample of popular songs. We also sampled less popular songs, but it is harder to be certain about whether that sample is representative because there is no complete list of songs out there. Given we collected all available sources on songs that included lyrics, genre, and year information, however, and the dataset is orders of magnitude larger than what has been used in prior work, hopefully, it is closer to being representative. Further, given that the population is exposed to popular songs much more frequently than unpopular ones, we believe we have a reasonable sample of what the population was exposed to.

That said, it is worth noting that we focused on English language songs, and song charts from the United States. Future work might examine other cultures, languages, and cultural products.

Finally, while automated textual analysis has some benefits, it also has some clear weaknesses. While it provides consistency across texts, and scalability across large datasets, we do not mean to suggest that automated methods are somehow better than more manual ones. Close reading allows researchers to pick up nuances that automated methods might miss. Indeed, before zeroing in on which aspects to focus on in this investigation, we manually read through dozens of songs to get a sense of how language was used and how it discussed men and women. In contrasting the differences between more automated and more manual approaches, Hart (2009) uses the metaphor of trying to understand a city by walking the streets or viewing it from a helicopter. Both provide different but useful pictures of what is going on. Consequently, automated and manual methods can be complementary, and often provide different viewpoints to the same problem.

## Conclusion

In conclusion, word embeddings and other computational linguistic techniques provide a powerful toolkit to study stereotypes, biases, and culture more generally. Natural language processing provides a reliable method of quantifying things and doing so at scale. These emerging approaches have the potential to shed light on a range of interesting questions, both in cultural analytics, and more broadly.

## References


Adams, T. M., & Fuller, D. B. (2006). The words have changed but the ideology remains the same: Misogynistic lyrics in rap music. *Journal of Black Studies*, *36*(6), 938–957. https://doi.org/10.1177/0021934704274072

Anderson, C. A., & Bushman, B. J. (2002). The effects of media violence on society. *Science*, *295*(5564), 2377–2379. https://doi.org/10.1126/science.1070765

Anderson, C. A., Carnagey, N. L., & Eubanks, J. (2003). Exposure to violent media: The effects of songs with violent lyrics on aggressive thoughts and feelings. *Journal of Personality and Social Psychology*, *84*(5), 960–971. https://doi.org/10.1037/0022-3514.84.5.960

Armstrong, E. G. (2001). Gangsta misogyny: A content analysis of the portrayals of violence against women in rap music, 1987–1993. *Journal of Criminal Justice and Popular Culture*, *8*(2), 96–126.

Berger, J., Kim, Y. D., & Meyer, R. (2021). What makes content engaging? How emotional dynamics shape success. *Journal of Consumer Research*, *48*(2), 235–250. https://doi.org/https://doi.org/10.1093/jcr/ucab010

Berger, J., Moe, W. W., & Schweidel, D. A. (2023). EXPRESS: What holds attention? Linguistic drivers of engagement. *Journal of Marketing*, Article 002224292311528. https://doi.org/https://doi.org/10.1177/00222429231152880

Berger, J., & Packard, G. (2018). Are atypical things more popular? *Psychological Science*, *29*(7), 1178–1184. https://doi.org/10.1177/0956797618759465

Berger, J., & Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, *77*(4), 525–537. https://doi.org/10.1037/amp0000882

Berkowitz, L. (1993). *Aggression: Its causes, consequences, and control*. Mcgraw-Hill.

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

Boghrati, R., Berger, J., & Packard, G. (2023). Style, content, and the success of ideas. *Journal of Consumer Psychology*, *109*, 20. https://doi.org/https://doi.org/10.1002/jcpy.1346

Bushman, B. J., & Anderson, C. A. (2002). Violent video games and hostile expectations: A test of the general aggression model. *Personality and Social Psychology Bulletin*, *28*(12), 1679–1686. https://doi.org/10.1177/014616702237649

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224. https://doi.org/10.1093/qje/qjz008

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., & Tar, C. (2018). *Universal sentence encoder*. ArXiv Preprint ArXiv:1803.11175.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show

robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045–1060. https://doi.org/10.1037/a0016239

Cooper, V. W. (1985). Women in popular music: A quantitative analysis of feminine images over time. *Sex Roles*, *13*(9–10), 499–506. https://doi.org/10.1007/BF00287756

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, *119*(1), 7–22. https://doi.org/10.1037/pspa0000188

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. ArXiv Preprint ArXiv:1810.04805.

Durkheim, E. (2009). *Sociology and philosophy (Routledge revivals)*. Routledge.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2019). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315. https://doi.org/10.1037/amp0000494

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Evans, K. M. (2014). *"One more Drinkin' Song": A longitudinal content analysis of country music lyrics between the years 1994 and 2013*. Brigham Young University-Provo.

Fischer, P., & Greitemeyer, T. (2006). Music and aggression: The impact of sexual-aggressive song lyrics on aggression-related thoughts, emotions, and behavior toward the same and the opposite sex. *Personality and Social Psychology Bulletin*, *32*(9), 1165–1176. https://doi.org/10.1177/0146167206288670

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Flynn, M. A., Craig, C. M., Anderson, C. N., & Holody, K. J. (2016). Objectification in popular music lyrics: An examination of gender and genre differences. *Sex Roles*, *75*(3–4), 164–176. https://doi.org/10.1007/s11199-016-0592-3

FreeDB. (2018). http://www.freedb.org/en/

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. https://doi.org/10.1037/a0022530

Greitemeyer, T., Hollingdale, J., & Traut-Mattausch, E. (2015). Changing the track in music and misogyny: Listening to music with pro-equality lyrics improves attitudes and behavior toward women. *Psychology of Popular Media Culture*, *4*(1), 56–67. https://doi.org/10.1037/a0030689

Gruber, J., Mendle, J., Lindquist, K. A., Schmader, T., Clark, L. A., Bliss-Moreau, E., Akinola, M., Atlas, L., Barch, D. M., Barrett, L. F., Borelli, J. L., Brannon, T. N., Bunge, S. A., Campos, B., Cantlon, J., Carter, R., Carter-Sowell, A. R., Chen, S., Craske, M. G., … Williams, L. A. (2021). The future of women in psychological science. *Perspectives on Psychological Science*, *16*(3), 483–516. https://doi.org/10.1177/1745691620952789

Harding, D., & Nett, E. (1984). Women and rock music. *Atlantis: Critical Studies in Gender, Culture and Social Justice*, *10*(1), 60–76.

Hart, R. P. (2009). *Campaign talk: Why elections are good for us*. Princeton University Press.

Herd, D. (2009). Changing images of violence in rap music lyrics: 1979–1997. *Journal of Public Health Policy*, *30*(4), 395–406. https://doi.org/10.1057/jphp.2009.36

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*(6), 1318–1335. https://doi.org/10.1037/a0026545

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Kaggle. (2017). *US Baby Names, Version 2*. https://www.kaggle.com/kaggle/us-baby-names

Kashima, Y. (2008). A social psychology of cultural dynamics: Examining how cultures are formed, maintained, and transformed. *Social and Personality Psychology Compass*, *2*(1), 107–120. https://doi.org/10.1111/j.1751-9004.2007.00063.x

Kesebir, S. (2017). Word order denotes relevance differences: The case of conjoined phrases with lexical gender. *Journal of Personality and Social Psychology*, *113*(2), 262–279. https://doi.org/10.1037/pspi0000094

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Laurino Dos Santos, H., & Berger, J. (2022). The speed of stories: Semantic progression and narrative success. *Journal of Experimental Psychology: General*, *151*(8), 1833–1842. https://doi.org/10.1037/xge0001171

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents. In International conference on machine learning*, pp. 1188–1196.

Lennings, H. I. B., & Warburton, W. A. (2011). The effect of auditory versus visual violent media exposure on aggressive behaviour: The role of song lyrics, video clips and musical tone. *Journal of Experimental Social Psychology*, *47*(4), 794–799. https://doi.org/10.1016/j.jesp.2011.02.006

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028. https://doi.org/10.1038/s41562-020-0918-6

Mahieux, T. B., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). *The million song dataset*. ISMIR-11.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Proceedings of the 26th international conference on neural information processing systems*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

Mishra, A., Mishra, H., & Rathee, S. (2019). *Examining the presence of gender bias in customer reviews using word embedding*. ArXiv Preprint ArXiv:1902.00496.

Morling, B., & Lamoreaux, M. (2008). Measuring culture outside the head: A meta-analysis of individualism—collectivism in cultural products.

*Personality and Social Psychology Review*, *12*(3), 199–221. https://doi.org/10.1177/1088868308318260

Moscovici, S. (2001). *Social representations: Essays in social psychology*. Nyu Press.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, *51*(1), 178–196. https://doi.org/10.1002/ejsp.2724

Packard, G., & Berger, J. (2020). Thinking of you: How second-person pronouns shape cultural success. *Psychological Science*, *31*(4), 397–407. https://doi.org/https://doi.org/10.1177/0956797620902380

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, *116*(24), 11693–11698. https://doi.org/10.1073/pnas.1818816116

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury Press/Bloomsbury Publishing.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

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

Sahlgren, M., & Lenci, A. (2016). *The effects of data size and frequency range on distributional semantic models*. ArXiv Preprint ArXiv:1609.08293.

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, Article e156. https://doi.org/10.7717/peerj-cs.156

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*(1), 21–35. https://doi.org/10.1037/0022-3514.74.1.21

Smith, S. L., Choueiti, M., Pieper, K., Clark, H., Case, A., & Villanueva, S. (2019). Inclusion in the recording studio? Gender and race/ethnicity of artists, songwriters and producers across 700 popular songs from 2012–2018. *USC Annenberg Inclusion Initiative*. https://assets.uscannenberg.org/docs/aii-inclusion-recording-studio-2019.pdf

Squire, M. (2019). *Which way to the wheat field? Women of the radical right on Facebook. Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2011695118. https://doi.org/10.1073/pnas.2011695118

Received September 9, 2022
Revision received February 14, 2023
Accepted February 21, 2023 ■