Bad Seeds: Evaluating Lexical Methods for Bias Measurement

Maria Antoniak
Cornell University
maa343@cornell.edu

David Mimno
Cornell University
mimno@cornell.edu

Abstract

A common factor in bias measurement methods is the use of hand-curated seed lexicons, but there remains little guidance for their selection. We gather seeds used in prior work, documenting their common sources and rationales, and in case studies of three English-language corpora, we enumerate the different types of social biases and linguistic features that, once encoded in the seeds, can affect subsequent bias measurements. Seeds developed in one context are often re-used in other contexts, but documentation and evaluation remain necessary precursors to relying on seeds for sensitive measurements.

1 Introduction

There has been increasing concern in the NLP community over bias and stereotypes contained in models and how these biases can trickle *downstream* to practical applications, such as serving job advertisements. In particular, there has been much recent scrutiny of word representations, with many studies finding harmful associations encoded in embedding models. Combating such biases requires *measuring* the bias encoded in a model so that researchers can establish improvements, and many variants of embedding-based measurement techniques have been proposed (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019).

These measurements have had the additional *upstream* benefit of providing computational social science and digital humanities scholars with a new means of quantifying bias in datasets of social, political, or literary interest. Researchers increasingly use embeddings (Garg et al., 2018; Knoche et al., 2019a; Hoyle et al., 2019) and other lexicon-based methods (Saez-Trumper et al., 2013; Fast et al., 2016; Rudinger et al., 2017) to provide quantitative answers to otherwise elusive political and social questions about the biases in a corpus and its authors. This work often involves comparing bias measurements across different corpora, which requires reliable, fine-grained measurements.

Target Concept	Highlighted Seeds
Unpleasant	divorce, jail, poverty, cancer, ...
African American	Tanisha, Tia, Lakisha, Latoya, ...
Domestic Work	mom, mum, ...
Ugliness	fat, chubby, obese, fatty, overweight, disformed, disfigured, wrinkle, wrinkled, ...

Table 1: Examples of real seed terms used in recent work to measure biases in corpora.

While there is a wide range of bias measurement methods, every one of them relies on lexicons of seed terms to specify stereotypes or dimensions of interest. But the rationale for choosing specific seeds is often unclear; sometimes seeds are crowd-sourced, sometimes hand-selected by researchers, and sometimes drawn from prior work in the social sciences. The impact of the seeds is not well-understood, and many previous seed sets have serious limitations. As shown in Table 1, the seeds used for bias measurement can themselves exhibit cultural and cognitive biases (e.g., reductive definitions), and in addition, linguistic features of the seeds (e.g., frequency) can affect bias measurements (Ethayarajh et al., 2019). Though they are often re-used, the suitability of these seeds to novel corpora is uncertain, and while evaluations sometimes include permutation tests, distinct sets of seeds are rarely compared.

We use a mixture of literature survey, qualitative analysis of seed terms, and analytic methods to explore the use of seed sets for bias measurement through two overarching research questions. (1) We explore *how seeds are selected and from which*

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1889–1904
August 1–6, 2021.

sources they are drawn to better understand rationales and assumptions underlying common seed sets. (2) We explore *which features of seeds can cause instability*, including both social biases and linguistic dimensions in our analysis.

Our work provides the following contributions. **Documentation:** We document and test 178 seed sets used in prior work, and we release this documentation as a resource for the research community.[1] **Analysis:** We provide a systematic framework for understanding the different sources of instability in seed sets that can affect bias measurements. We compare the gathered seeds to larger sets of artificially created seed sets, and we investigate the reliability of seed terms used for two popular embedding-based bias measurement methods in case studies on three datasets. **Recommendations:** With this larger perspective, we discuss how seed sets should be examined versus how these sets are popularly considered and what kind of documentation best practices should be followed. Seeds are a brittle but unavoidable element of current bias measurement algorithms, with weaknesses that need probing even for embedding-based measurements.

2 Background and Related Work

The term "bias" has many definitions, from a value-neutral meaning in statistics to a more normative meaning in socio-cultural studies. In the bias measurement literature in NLP, lack of precise definitions and problem specifications (Blodgett et al., 2020) has led to many of the errors we explore in this paper. In general, "bias" in NLP most often represents *harmful prejudices* (Caliskan et al., 2017) whose spurious and undesirable influence can affect model outputs. While these downstream effects have inspired work on "removing" bias from embedding models (Bolukbasi et al., 2016), there have also been critiques of these efforts (Gonen and Goldberg, 2019), and we do not focus on this use case in our study. Instead, we focus on bias measurement as a tool used in diverse settings to make comparisons across specific corpora of interest.

Unsupervised methods for bias measurement have included pointwise mutual information (Rudinger et al., 2017), normalized frequencies and cosine similarity of TF-IDF weighted word vectors (Saez-Trumper et al., 2013), generative models (Joseph et al., 2017; Hoyle et al., 2019), and a combination of odds ratios, embeddings, and crowd-sourcing (Fast et al., 2016). All of these methods rely on sets of seed terms. While much recent NLP work has focused on contextual embeddings, most recent bias-detection work has focused on vocabulary-based embeddings and word representations. Researchers have increasingly used embedding-based methods to measure biases and draw comparisons in training corpora of social interest (Kim et al., 2014; Hamilton et al., 2016; Kulkarni et al., 2016; Phillips et al., 2017; Kozlowski et al., 2019). For example, Bhatia et al. (2018) train embedding models on news sources to compare trait associations for political candidates. We believe that our results should extend to contextual embedding methods (Zhao et al., 2019; Sedoc and Ungar, 2019), but vocabulary-based embeddings are easier to analyze.

We discuss several recent studies that include analysis of seed sets (Kozlowski et al., 2019; Ethayarajh et al., 2019; Sedoc and Ungar, 2019) in §8.

3 Data Description

Training Corpora. Our dataset choices are guided by our focus on the *upstream* use case, where embeddings are trained on relatively small, special-purpose collections to answer social and humanist questions about the training corpus. The scope of these datasets fits the use case of a social scientist interested in measuring bias during a small time window, across specific genres, or in a particular set of authors. Table 2 shows an overview of the data, and more details are in the Appendix.

Our datasets include: **New York Times** articles from April 15th-June 30th, 2016; high quality **WikiText** articles, using the full WikiText-103 training set (Merity et al., 2016); and **Goodreads** book reviews for the *romance* and *history and biography* genres, sampled from the UCSD Book Graph (Wan and McAuley, 2018; Wan et al., 2019). For added validity, we also replicate existing studies, using a pre-trained model on a large Google News corpus (Mikolov et al., 2013).

For each dataset, we lowercase all text, parse and obtain POS tags using *spaCy* (Honnibal et al., 2020), tokenize the text into unigrams, and filter words that occur fewer than 10 times in the training dataset. Lowercasing controls for the varying levels of capitalization used in the gathered seeds. We leave analysis of bigram seeds to future work and rely on unigrams as a simplifying assumption.

[1] Seeds and documentation are available at https://github.com/maria-antoniak/bad-seeds

Dataset	Total Documents	Total Words	Vocabulary Size	Mean Document Length
NYT	8,888 articles	7,244,457 words	162,998 unique words	815 words
WikiText	28,472 articles	99,197,146 words	546,828 unique words	3,484 words
Goodreads (Romance)	197,000 reviews	24,856,924 words	214,572 unique words	126 words
Goodreads (History/Biography)	136,000 reviews	14,324,947 words	163,171 unique words	105 words

Table 2: Summary statistics for our test datasets. In contrast to the large, generic datasets often used for downstream applications, these datasets are small and culturally specific.

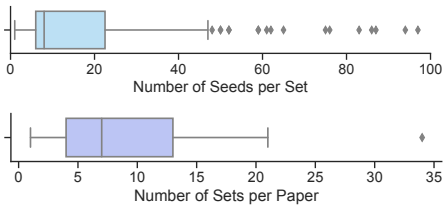


Figure 1: Overview of the gathered seed sets, showing quartiles and medians. Outliers are truncated on the plot showing the number of seeds per set; the maximum number is 1,460 seeds.

Corpus-Derived	7/18 papers
Re-Used	7/18 papers
Borrowed from Social Sciences	6/18 papers
Curated	5/18 papers
Adapted from Lexical Resources	3/18 papers
Crowd-Sourced	2/18 papers
Population-Derived	2/18 papers

Table 3: Overview of the surveyed seed sources.

Gathered Seeds Sets. We gather 178 seed sets used in a representative sample of 18 highly-cited prior works on bias measurement. Seeds include both embedding-based and non-embedding-based bias detection methods as there is often crossover and re-use of seed sets. Because we use word embedding models trained on unigrams, we do not include bigram seeds in our analysis, and in each experiment, we omit words that were not present in our training set. While these choices could be seen as limitations, we see them as realistic applications of seeds to constrained datasets, reflecting the scenario in which biases are compared across specific corpora. Figure 1 overviews the seed sets, examples used in the paper are documented in the Appendix, and the full collection is shared in the supplementary materials and is available online.

4 How Are Seeds Selected?

How do researchers select seeds, and from which sources are they popularly drawn? We explore this question using the gathered seed sets from prior works on unsupervised bias detection. The origins of these seeds and the rationales for using them are not always explained by researchers, but in cases where we were able to determine a source or rationale, we group them into the following categories. Table 3 overviews the source frequencies. We emphasize that each source comes with risks and benefits; there is no one correct method to selecting seeds, but awareness of pros and cons can help guide decisions and evaluation methods.

Borrowed from Social Sciences. Seed sets are often borrowed from prior work in psychology and other social sciences, usually in an effort to either replicate results or build confidence from previously validated work. For example, Caliskan et al. (2017) validate prompts from the Implicit Association Test (Greenwald et al., 1998), while Garg et al. (2018) and Hoyle et al. (2019) use personality traits from Williams and Bennett (1975); Williams and Best (1977, 1990). Sometimes the seeds appeal for validity via highly cited resources, like LIWC (Pennebaker et al., 2001), despite critiques about unreliability (Panger, 2016; Forscher et al., 2017). Borrowing seeds does not absolve researchers from examining and validating seeds.

Crowd-Sourced. Custom seed sets can be created through crowd-based annotation. Fast et al. (2016) use Mechanical Turk to validate the inclusion of terms in their seed sets; the final terms are then included in packaged code for researchers and practitioners. Kozlowski et al. (2019) use Mechanical Turk to gather ratings of items scaled along gender, race, and class. Crowd-sourcing can aid in gathering contemporary associations and stereotypes. However, controlling for crowd demographics can be difficult, and crowd-sourcing can result in alarming errors, in which popular stereotypes are hard-coded into the seeds (as in Table 1).

Population-Derived. Some seed sets are derived from government-collected population datasets. Popular sources include U.S. census data (Bolukbasi et al., 2016; Caliskan et al., 2017), the U.S. Bu-

reau of Labor Statistics (Caliskan et al., 2017), and the U.S. Social Security Administration (Garg et al., 2018). These sources are usually used to gather names and occupations common to certain demographic groups. These sources tend to be U.S.-centric, though the training data for the embedding does not always match (e.g., large Wikipedia datasets are not guaranteed to have U.S. authors). Reliance on these sources is particularly vulnerable to reductive definitions of the target concepts—e.g., gender (Keyes, 2017)—and assumes a level of trust and representation in the data collection that might not exist evenly across groups.

Adapted from Lexical Resources. Some seed sets are drawn from existing dictionaries, lexicons, and other public resources, such as SemEval tasks (Zhao et al., 2018) and ConceptNet (Fast et al., 2016). Pre-packaged sentiment lexicons are a popular source (Saez-Trumper et al., 2013; Sweeney and Najafian, 2019); these lexicons include the Affective Norms for English Words (ANEW) (Bradley and Lang, 1999) and negative/positive sentiment words from Hu and Liu (2004). These seeds have the advantage of previous rounds of validation, but this does not guarantee validity for new domains.

Corpus-Derived. Quantitative methods can be used to extract seed terms from a corpus of interest. For example, Saez-Trumper et al. (2013) use sorted lists of named entities extracted from a target dataset to create seed sets for personas of interest. Similarly, Sweeney and Najafian (2019) extract high frequency identity terms from a Wikipedia corpus. These methods have the advantage of ensuring high frequency terms in the target dataset, but they pose similar risks to crowd-sourcing; unless an extra round of cleaning and curation is completed by the researchers, terms with unintended effects can be included in the seed sets.

Curated. Seed sets are sometimes hand-selected by the authors, usually after close reading of the corpus of interest. For example, Rudinger et al. (2017) hand-select a set of seed terms that correspond to a set of demographic categories of interest, and Joseph et al. (2017) hand-select a set of identity seeds based on their frequency in a Twitter dataset. Often, even when papers rely on other seed sources, manual curation is included as a step in the seed creation process. Hand-curation can result in high precision seeds, but this method relies on the authors' correction for their own social biases.

Re-Used. Finally, many papers rely on prior bias measurement research for seed terms. The most popular sources in our survey include early papers on bias in embeddings such as Bolukbasi et al. (2016) and Caliskan et al. (2017). This repetition means that the seeds are tested on many different datasets, but they should not be trusted without validation; there can be mismatches in frequency and contextual meaning between datasets.

5 Bias Measurement Algorithms

In the *upstream* use case, locally trained word embeddings remain state of the art because fine-tuned pre-trained contextual models might introduce extrinsic information, and it is not feasible to pre-train contemporary contextual embeddings on such small collections. Here, we focus on two popular seed-based methods to detect bias in word embeddings. Bolukbasi et al. (2016) and Caliskan et al. (2017) both introduce embedding-based methods for bias detection that rely on sets of seed words. Each of these methods requires two sets of seed words, $\mathcal{X}$ and $\mathcal{Y}$, and one additionally requires matched *pairs* of seed words $\{(X_1, Y_1), (X_2, Y_2), ...\}$.

WEAT. Given a set of embedding vectors w, the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) defines a vector based on the difference between the mean vector of the two target sets, and then measures the cosine similarity of a set of attribute words to that vector. The strength of the association between the target sets $\mathcal{X}$ and $\mathcal{Y}$, and the sets of attributes, $\mathcal{A}$, and $\mathcal{B}$, is given by

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

where $s(w, \mathcal{A}, \mathcal{B})$ is equal to the difference in average cosine similarities between a query w and each term in $\mathcal{A}$ and w and each term in $\mathcal{B}$. To test whether the resulting difference $s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B})$ is significant, this result is compared to the same function applied to randomly permuted sets drawn from $\mathcal{X}$ and $\mathcal{Y}$. Caliskan et al. (2017) use WEAT to measure stereotypical associations between sets of targets and attributes, where, for example, the target terms might be *arts* and *science* terms, and the attribute terms might be *male* and *female* terms.

PCA. The principal component analysis (PCA) method tests how much variability there is in the difference vectors between pairs of word vectors

(Bolukbasi et al., 2016). If the vector difference between pairs of seed terms can be approximated well by a single constant vector c, then this vector represents a *bias subspace*. In this case, the subspace is simply a one dimensional vector, though this process could be extended to more dimensions. For each *pair* of embedding vectors corresponding to one seed word from set $\mathcal{X}$ and one from set $\mathcal{Y}$, Bolukbasi et al. (2016) calculate the mean vector of those two vectors and then include the two resulting half vectors from that mean to the two seed vectors as columns in the input matrix.

6 Quantifying Variation from Seeds

To quantify how large an effect seed features can have on bias measurements, we calculate a set of metrics for both PCA and WEAT methods that summarizes how well the bias subspace represents the target seeds. For each dataset, we use the popular skip-gram with negative sampling (SGNS) algorithm to train a word2vec model. We use the *gensim* package for training (Řehůřek and Sojka, 2010). We use a window size of 5, a minimum word count of 10, and a vector size of 100 for all experiments. We repeat this process across 20 bootstrapped samples of each dataset.

For PCA, we calculate the difference vector between the embedding vectors for each pair of words in the two seed sets. For each set of paired seed sets, we run PCA and plot the percent of variance explained by each component. For the gathered seeds, we only use pairings documented in prior work. We perform a manual confirmation that the first component g indeed represents the bias subspace by ranking all the words in the vocabulary by their cosine similarity to g.

For WEAT, we hold the attribute terms constant, where $\mathcal{A} = [\text{“}good\text{”}]$ and $\mathcal{B} = [\text{“}bad\text{”}]$, while our generated seed sets take the place of the target terms $\mathcal{X}$ and $\mathcal{Y}$. Holding the attribute terms constant is a simplifying assumption; our goal is not to test all possible attribute terms but to show that significant variation is possible. We then calculate the WEAT test statistic and significance.

Coherence. In addition to the PCA explained variance and WEAT test statistic, we also measure the *coherence* of each pairing of seed sets after being mapped to the bias subspace. Ideally, when we project all the words in the vocabulary onto the subspace, the two sets would be drawn as far apart as possible. We rank all words by cosine similarity to the bias subspace, and we measure the absolute difference in mean rank of the paired seed sets:

$$\text{Coherence}(X_1, Y_2) = |\overline{R_1} - \overline{R_2}|,$$

where X_1 and Y_2 are seed sets and $\overline{R_1}$ and $\overline{R_2}$ are their mean ranks in the bias subspace. Finally, we normalize the scores to a [0, 1] range. Higher coherence scores indicate that the seed sets have very different mean ranks, i.e., the seeds are separated by more of the vocabulary. For example, in Figure 4, ordered seeds (a) produce a subspace with greater coherence (sets are further apart in the bias subspace) than shuffled seeds (b).

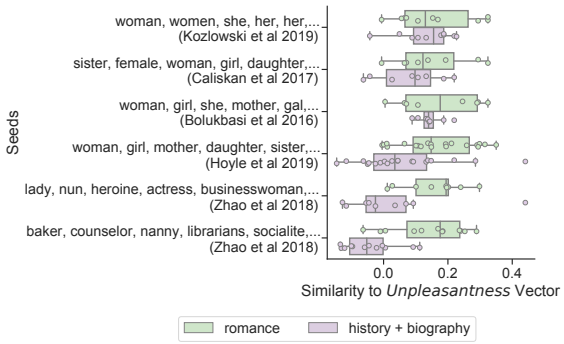


Figure 2: Bias measurements depend on seeds. We calculate the cosine similarities between different *female* seed sets and an averaged *upleasantness* vector from two embedding models. Results are consistent across seeds for romance review embeddings, but vary widely between sets for history and biography. We find similar variation even for a pretrained Google News model.

Generated Seed Sets. In order to control for frequency and POS when measuring instabilities due to semantic similarity and word order, we generate a large collection of artificial, randomized seed sets. We select a target term at random from the model's vocabulary, filtered by POS. Each seed set consists of this target term and its four nearest neighbors, ranked by cosine similarity. We repeat this process for each of the models trained on the bootstrapped samples of the corpus. We choose seed sets that are semantically similar (rather than randomly selecting seeds) because we expect that seed sets of realistic research interest would be coherent. We emphasize that researchers have used bias measurement methods for increasingly creative purposes, moving beyond gender and race, and similar bias measurement techniques can be used for aspect detection and other seed-based tasks. Example seeds are shown in Table 4.

Coherence	Generated Seed Set A	Generated Seed Set B
1.000	distinctions, similarities, friction, parallels, similarity	murder, rape, manslaughter, felony, assault
1.000	mile, miles, yards, yard, feet	example, instance, purposes, explanation, shorthand
1.000	shop, restaurant, kitchen, cafe, store	sports, soccer, football, competitions, basketball
...	...	...
0.711	ambush, bombardment, escalation, altercation, militiamen	corruption, terrorism, graft, bribery, abuses
0.689	entrance, terrace, subway, cafe, lawn	courtside, bamboo, freeway, shorts, sailboat
0.552	sticks, onions, tops, banana, mozzarella	potatoes, onions, lemon, herbs, meats

Coherence	Gathered Seed Set A	Gathered Seed Set B
0.933	CAREER: executive, management, professional...	FAMILY: home, parents, children, family, cousins...
0.910	ASIAN: asian, asian, asian, asia, china...	CAUCASIAN: caucasian, caucasian, white, america...
0.909	FEMALE: sister, mother, aunt, grandmother...	MALE: brother, father, uncle, grandfather, son...
...	...	...
0.375	FEMALE: countrywoman, sororal, witches...	MALE: countryman, fraternal, wizards, manservant...
0.110	NAMES ASIAN: cho, wong, tang, huang, chu...	NAMES CHINESE: chung, liu, wong, huang, ng...
0.050	NAMES BLACK: harris, robinson, howard...	NAMES WHITE: harris, nelson, robinson...

Table 4: When two seed sets are more semantically distinct they are more distinguishable in the resulting geometric subspace. The top table shows pairs of artificially generated seed sets, ranked by their coherence for WEAT in the NYT dataset. The bottom table shows pairs of seed sets gathered from published papers, ranked by their coherence for WEAT in the WikiText dataset. Scores are averaged across 20 bootstrapped samples of the training data, and values are rounded; no coherence scores are exactly 1.0. Higher coherence scores indicate that the seeds pairs were projected farther apart in the bias subspace.

7 Seed Choice Affects Bias Measurement

Before moving to specific seed features, we present some general results showing the instability of measurements using seeds. Figure 2 shows a motivating example, in which we imagine a digital humanities scholar interested in measuring whether women are portrayed more negatively in different genres of book reviews. As in the WEAT test, each seed is plotted according to its cosine similarity to an averaged *unpleasantness* vector (Caliskan et al., 2017). For some sets, no significant difference is visible, while for other sets, there are much larger differences, causing the researcher to draw different conclusions when comparing biases across datasets.

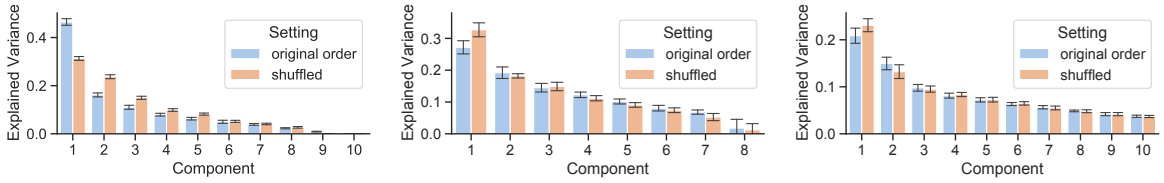
Table 4 shows both the generated and gathered seed sets ordered by their coherence after using the WEAT method to discover a bias subspace. These examples highlight factors contributing to lower coherence (e.g., similarity of the seed sets) which we discuss in §8. They also highlight the general difficulty in constructing seed sets; e.g., as noted by Garg et al. (2018), the final row demonstrates that some U.S. racial categories are not distinguishable from available census data. Similar challenges arise when seeds do not occur in the target dataset, which is often true for names. The wide variation in coherence scores, especially for the generated seeds which are less likely to contain overlapping terms, indicates that different seed sets can have widely differing "success" for bias measurement.

8 Factors Causing Instability

Sometimes seeds can reflect the curator's (or crowd's) personal biases. Instabilities can also arise from the organization of the seeds and seemingly innocuous linguistic features. We describe a series of distinct sources of instability that can be encoded in seed sets and discuss the implications of each. We rely on a combination of literature review, qualitative close reading of example seeds, and quantitative tests of seed features. We iterated through the seeds, flagging problematic sets, and then manually clustered and labeled the factors that could cause instability.

Our identified factors can be categorized as **definitional factors** (reductive definitions, inclusion of confounding concepts), **lexical factors** (frequency, POS of individual seeds), and **set factors** (number and order of seeds, similarity of seed sets).

Reductive Definitions. The seeds can be reductive and essentializing, codifying life experiences into traditional categories. Using names as placeholders for concepts like race (Nguyen et al., 2014; Sen and Wasow, 2016) or reducing gender to a binary with two extremes (Bolukbasi et al., 2016; Caliskan et al., 2017) can create a distorted view of the source data. Sometimes these are simplifying assumptions, made in an effort to measure biases that would otherwise go unexamined. However, these decisions run the risk of further entrenching these category definitions—e.g., see discussions



(a) Gender Pairs (b) Social Class Pairs (c) Chinese-Hispanic Name Pairs

Figure 3: We replicate previous gender bias results and experiment on other ordered pairs, using the NYT dataset. The first PCA component dominates for ordered gender pairs but not for shuffled gender pairs (a), while shuffling can produce a component that explains more variance for class (b) and pleasantness (c) pairs. We find similar instabilities using the pretrained model used in Bolukbasi et al. (2016). Error bars show standard deviation over the 20 bootstrapped models. Seeds pairs are listed in the Appendix.

herself	0.50	likelihood	0.36	outcomes	0.26
ms	0.49	eurozone	0.34	son	0.26
her	0.49	incentive	0.34	father	0.26
she	0.41	downturn	0.31	mother	0.26
pregnant	0.40	setback	0.30	aunt	0.25
pitching	-0.36	photographed	-0.39	potentially	-0.19
baseball	-0.36	tales	-0.41	male	-0.19
syndergaard	-0.38	hood	-0.42	hood	-0.29
himself	-0.39	garcia	-0.45	garcia	-0.29
his	-0.42	danced	-0.59	md	-0.39

(a) Gender Pairs (b) Random Pairs (c) Shuffled Gender Pairs

Figure 4: Ranking word vectors by cosine similarity with the top principle component vector for the original gender seed pairs (a) appears to identify female and male gendered words much better than random (b). But shuffling the pairing of seed words (c) maintains correlation with gender but to a less clear degree. Results are shown for the NYT corpus with a frequency threshold of 100 and bootstrap resampling.

in Keyes (2017); Larson (2017) for the mistakes and harms that can be caused by mapping names to genders—and these trade-offs should be evaluated and documented. More broadly, recent work has critiqued NLP and ML bias research for not successfully connecting with the literature in sociology and critical race studies (Hanna et al., 2020; Blodgett et al., 2020). Engaging with this literature would provide a better foundation for decision-making about seed sets and provide context for future researchers.

Imprecise Definitions. If the target concept is not well-defined, the resulting seed terms can be too broad and include multiple concepts, risking the creation of confounded or circular arguments. Similarly, the unexamined use of pre-existing sets and over-reliance on the category labels from prior work can result in a series of errors. The seeds can contain confounding terms (e.g., in Table 1, *unpleasant* contains "cancer" which in some datasets might be more prevalent for certain demographic groups) or terms from the target group (e.g., *domestic work* includes the gendered terms "mom" and "mum"). Similarly, the seeds can manifest cultural stigmas: for example, including "fat" and "wrinkled" in an *ugliness* category (Fast et al., 2016) results in a seed set that itself contains stereotypes.

These stigmas are harmful and can interact with other demographic features like gender or age (Puhl and Heuer, 2009), and unless their inclusion is intentional, they can accidentally inflate measurements towards certain groups. Predicting all such errors is impossible, and there can be cases where researchers intentionally include such terms (e.g., to capture a particular stereotype)—but this should be a conscious decision by each researcher using the seeds, and at a minimum, researchers should clearly define their target concepts.

Lexical Factors. Prior work examining seeds has shown that the frequency and part of speech of seeds can affect the resulting bias measurements. Ethayarajh et al. (2019) show that the WEAT test requires that the paired seeds occur at similar frequencies and that seed sets can be manipulated to produce certain measurements. Brunet et al. (2019) explore the effects of perturbing the training corpus, finding that (1) second-order neighbors to the seeds can have a strong impact on the bias measurement effect size and (2) effects are stronger for rarer words. Using contextual embeddings, Sedoc and Ungar (2019) show that different classes of words (e.g., names vs. pronouns) can result in different bias subspaces and that sometimes these subspaces represent an unintended dimension (e.g., age instead of gender).

Set Size and Alignment. The number of seeds included in each set can affect the resulting bias subspace; Kozlowski et al. (2019) find small increases in performance when using more seed pairs. The alignment of the seeds in matched sets (i.e., the ordering or pairing of seeds in one set with seeds in another set) can also affect the bias subspace. In the PCA method, each term in one seed set is explicitly linked to a single term in the other seed set. The specific alignment between paired words matters; altering the pairing can result in dramatically different results, even for cases like gender, which is marked in English. However, we observe conscious pairings of seeds only for obvious cases, and sometimes "obvious" pairings produce subspaces that explain less variance.

We replicate a study previously carried out on embeddings trained on internet-scale collections (Bolukbasi et al., 2016) using both a large, pretrained embedding and the relatively small NYT dataset. Figure 3 shows how much variance is explained by the first ten principal components of three difference matrices. When we use the original paired male-female seed words from Bolukbasi et al. (2016) (e.g., *man-woman*, *he-she*), we see a single dominant first component, suggesting a strong male-female axis. As previously reported, the variances fall off gradually when the seeds are a set of random words. When we shuffle the order of the seed words, the drop off is steeper than for random pairs, but there is no longer a single dominant principal component.

Similarly, Figure 4 shows that when we used the ordered gender pairs, the ranked words roughly divide into groups correlated with gender, while if we use shuffled pairs, the lists of high and low ranked words are not as easily distinguishable as masculine or feminine. We find an opposite effect *social class* pairs (Kozlowski et al., 2019); when we shuffle, we find a subspace that explains more variance than the explicitly ordered pairs (e.g., "richest"-"poorest"). We find similar differences when testing some seed sets that lack intuitive pairings, e.g., the matched *pleasantness* and *unpleasantness* seeds (Caliskan et al., 2017) and the matched *Christianity* and *Islam* seeds (Garg et al., 2018).

Order does not always affect the subspace—e.g, we found no significant difference when shuffling sets of names—but we have shown that it *can* affect the subspace, and so to build confidence in measurements, testing is required.

Figure 5: Identifying bias is less effective when set pairs are similar. Generated seeds are frequency-controlled nouns from the WikiText dataset. We highlight two sets of gathered seeds; both target similar racial categories but the name-based sets are more similar and explain less variance. We find similar trends for WEAT, coherence, and the other corpora and POS.

Set Similarity. By sampling random seed sets we find that it is more difficult to represent the variance of seed sets that are too close together. Figure 5 shows that *set similarity* (cosine similarity between the set mean vectors) is significantly correlated with explained variance for generated sets (Pearson $r = -0.67$, $p < 0.05$). We highlight two comparisons between gathered sets intended to measure racial bias that explain different degrees of variance. Synthetic pairings generally explain more variance than pairings of gathered sets of equal similarity, although for gathered sets we cannot control for POS and frequency. Table 4 shows the generated seed sets ranked by coherence, where higher scores indicate that the bias subspace was able to separate the seed sets. Similar seed sets and sets with duplicates (e.g., the pairing in the table in which both generated sets contain food terms) have low coherence scores.

9 Conclusion: Biases All the Way Down

Almost all recent work on bias measurement relies on sets of seed terms to ground cultural concepts in language. If we do not pay attention to the seeds, these methods will lack foundation and the claims they support will be left open to criticism and dismissal. *Seeds and their rationales need to be tested and documented, rather than hidden in code or copied without examination.*

Some of the risks discussed in this paper may seem obvious in retrospect, but our literature survey suggests there are widely varying levels of evaluation and documentation. Rationales for picking

sources or seeds are not always explained, or the reader is left to assume that prior work has adequately validated the seeds. Tests for frequency, semantic similarity, and other features are rare or non-existent, and clear definitions and discussion of limitations are often missing. Permutation tests are sometimes used, but these do not account for seeds outside of those already selected. Significantly different results can be found using alternative seeds sets for the same target concept, and fine-grained comparisons require validation on multiple sets.

We faced a number of challenges in gathering 178 seed sets from prior work. Sometimes seeds are shared online at an undocumented location and sometimes hard-coded into code repositories; this can significantly obscure the seeds from public view, which is troubling for tools intended for wide use on sensitive topics. Documentation is often scattered across locations, and in more than one case, we found contradictions between different sources for a single project. In one case, we were unable to find the full list of seeds used in the paper, and in several cases, it was unclear which seed sets were used for which experiments. While some authors went to commendable lengths to document their materials, there is a need for more consistent and transparent documentation.

We recommend that researchers carefully **trace the origins of seed sets**, with attention to the risks associated with the origin type. We also recommend that researchers **examine seed features**. POS, frequency, semantic similarity, and pairing order can significantly affect the results of bias measurements. Seeds should be both examined manually and tested as shown in §8; importantly, they should be compared to alternative seeds with different attributes, as in §7. To assist this we release a compilation of 178 seed sets from prior work. These tests are particularly important when comparing biases across datasets. Finally, researchers should **document all seeds** and the rationales underlying their design, including concept definitions. We add to recent calls for better documentation and problem specification in machine learning (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019; Blodgett et al., 2020) and in studies of social biases in technology (Olteanu et al., 2019). Specifically, when the seeds intentionally encode harmful stereotypes or slurs, it can be beneficial to include a trigger warning or not highlight the seeds in the paper; however, full seed lists should always be accessible, not hard-coded, with unique labels matched to experiments.

Ultimately, our goal is not to *eliminate* a problem but to *illuminate* it:[2] to help practitioners think through the potential risks posed by seed sets used for bias detection. We encourage thoughtful, critical studies, but we observe a trend in which seed sets are used in new research and applications simply because they have been used in prior published work, without additional vetting. Research precedents can take on a life of their own and we have a responsibility to explore and document possible sources of error. We believe that seed sets can be useful and are probably unavoidable, but that no technical tool can absolve researchers from the duty to choose seeds carefully and intentionally.

Acknowledgements

Thank you to our anonymous reviewers whose comments substantially influenced and improved this paper. Thank you to Rishi Bommasani, Forrest Davis, Os Keyes, Lauren Kilgour, Rosamund Thalken, Marten van Schijndel, Melanie Walsh, and Gregory Yauney for their many helpful suggestions. This work was funded through NSF grant #1652536.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Sudeep Bhatia, Geoffrey P Goodwin, and Lukasz Walasek. 2018. Trait associations for Hillary Clinton and Donald Trump in news media: A computational analysis. *Social Psychological and Personality Science*, 9(2):123–130.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to

[2] "All problems can be illuminated; not all problems can be solved."—Ursula Franklin (quoted by M. Meredith via Olteanu et al. (2019) in http://bb9.berlinbiennale.de/all-problems-can-be-illuminated-not-all-problems-can-be-solved/)

homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report. The Center for Research in Psychophysiology, University of Florida.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pages 803–811.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA. Association for Computing Machinery.

Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. 2017. A meta-analysis of change in implicit bias. *Psychological Bulletin*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, PMLR*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Kenneth Joseph, Wei Wei, and Kathleen M Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from twitter. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1362–1374. ACM.

Os Keyes. 2017. Stop mapping names to gender. `https://ironholds.org/names-gender/`. Accessed: 2021-05-26.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2019a. Identifying biases in politically biased wikis through word embeddings. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 253–257, New York, NY, USA. ACM.

Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2019b. Identifying biases in politically biased wikis through word embeddings. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 253–257. ACM.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1).

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Galen Panger. 2016. Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. *Information, Communication & Society*, 19(8):1108–1126.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Lawrence Phillips, Kyle Shaffer, Dustin Arendt, Nathan Hodas, and Svitlana Volkova. 2017. Intrinsic and extrinsic evaluation of spatiotemporal text representations in Twitter streams. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 201–210.

Rebecca M Puhl and Chelsea A Heuer. 2009. The stigma of obesity: a review and update. *Obesity*, 17(5):941–964.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1679–1684. ACM.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

John E Williams and Susan M Bennett. 1975. The definition of sex stereotypes via the adjective check list. *Sex Roles*, 1(4).

John E Williams and Deborah L Best. 1977. Sex stereotypes and trait favorability on the adjective check list. *Educational and Psychological Measurement*.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev*. Sage Publications, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Datasets

New York Times. This dataset contains 165,900 paragraphs from articles published between April 15th and June 30th, 2016.[3] The articles are drawn from all sections of the English language news, including Movies, Sports, Technology, World, U.S., Arts, Business, Books, NY Region, Health, Science, and Fashion. This dataset is small in comparison to the large training datasets used for downstream features; its scope fits the use case of a social scientist interested in measuring bias during a small time window at a particular publication.

WikiText. The WikiText training corpus contains the texts of 28,000 manually verified high-quality articles from Wikipedia.org (Merity et al., 2016). Lists have been removed, along with HTML errors, math, and code. We use the full training dataset, WikiText-103.[4] This dataset is much larger than the NYT dataset but is still of focused interest in a particular online community (Wikipedia authors).

Goodreads. We sample 500 Goodreads book reviews for books in the *romance* and *history and biography* genres, removing books with fewer than 500 reviews and reviews containing fewer than 20 characters. We use the provided genre samples from the UCSD Book Graph (Wan and McAuley, 2018; Wan et al., 2019).[5]

Google News. For some of our experiments, as a comparison for the smaller datasets, we use a model pre-trained on part of the Google News dataset.[6] This is a popular model, used in Bolukbasi et al. (2016) and many other studies. This data originates from an internal Google dataset (Mikolov et al., 2013), and we could not find a comprehensive description of the data beyond its vocabulary size: 3 million unique words and 100 billion tokens.

[3] https://www.kaggle.com/nzalake52/new-york-times-articles

[4] https://blog.einstein.ai/the-wikitext-long-term-dependency\-language-modeling-dataset/

[5] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home

[6] https://code.google.com/archive/p/word2vec/

A.2 Seed Terms

Because of the Appendix page limit, we cannot list here all the seed sets gathered from prior work. Instead, the full seed sets in addition to the rationales and sources used for their curation are released as a supplementary JSON file. After publication, the seeds will also be documented at a public website. Below, we list all the seeds used as examples (in figures or text) in the main paper. The seed IDs correspond to a matching ID field in the supplementary JSON file.

Table 1

- **Seeds ID:**
  unpleasant-Caliskan_et_al_2017
  Used In: Caliskan et al. (2017)
  Seeds: [abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison]

- **Seeds ID:**
  african_american_names-Caliskan_et_al_2017
  Used In: Caliskan et al. (2017)
  Seeds: [Alonzo, Jamel, Theo, Alphonse, Jerome, Leroy, Torrance, Darnell, Lamar, Lionel, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Wardell, Nichelle, Shereen, Ebony, Latisha, Shaniqua, Jasmine, Tanisha, Tia, Lakisha, Latoya, Yolanda, Malika, Yvette]

- **Seeds ID:**
  domestic_work-Fast_et_al_2016
  Used In: Fast et al. (2016)
  Seeds: [chore, mom, vacuum, scrubbing, cook, washing, baking, wash, morning, meal, house, chef, laundry, bake, organizing, cooking, spotless, mum, washer, remodeling, parent, job, nanny, kitchen, dishwasher, cleaning, family, cleaner, bathroom, errand, sitter, housekeeper, serve, housekeeping, tidy, cleaned, housework, scrub, organize, home, clean]

- **Seeds ID:**
  ugliness-Fast_et_al_2016
  Used In: Fast et al. (2016)
  Seeds: [despise, balding, slimy, acne,

grotesque, degrading, horrible, fat, diseased, repulsive, awful, nasty, brutish, grotesquely, distasteful, unworthy, scruffy, chubby, gross, insulting, crooked, revolting, unappealing, hairy, pathetic, cockroach, abnormally, unsightly, crippled, lousy, wrinkled, freakish, disfigured, disgusting, pudgy, tacky, obese, disgust, degrade, horrid, deformed, hideous, bloated, ugly, scum, demeaning, pig, obnoxious, blob, wart, disgraceful, fatty, bald, overweight, disgusted, unattractive, wrinkle, filthy, loathsome]

Table 4

- **Used In:** Caliskan et al. (2017)
  Seeds ID 1:
  career-Caliskan_et_al_2017
  Seeds ID 2:
  family-Caliskan_et_al_2017
  Seeds 1: [executive, management, professional, corporation, salary, office, business, career]
  Seeds 2: [home, parents, children, family, cousins, marriage, wedding, relatives]

- **Used In:** Manzini et al. (2019)
  Seeds ID 1:
  asian-Manzini_et_al_2019
  Seeds ID 2:
  caucasian-Manzini_et_al_2019
  Seeds 1: [asian, asian, asian, asia, china, asia]
  Seeds 2: [caucasian, caucasian, white, america, america, europe]

- **Used In:** Caliskan et al. (2017)
  Seeds ID 1:
  female_2-Caliskan_et_al_2017
  Seeds ID 2:
  male_2-Caliskan_et_al_2017
  Seeds 1: [sister, mother, aunt, grandmother, daughter, she, hers, her]
  Seeds 2: [brother, father, uncle, grandfather, son, he, his, him]

- **Used In:** Zhao et al. (2018)
  Seeds ID 1:
  female_definition_words_1-Zhao_et_al_2018
  Seeds ID 2:
  male_definition_words_1-Zhao_et_al_2018
  Seeds 1: [countrywoman, sororal, witches, maidservant, mothers, diva, actress, spinster, mama, duchesses, barwoman, countrywomen, dowry, hostesses, airwomen, menopause, clitoris, princess, governesses, abbess, women, widow, ladies, sorceresses, madam, brides, baroness, housewives, godesses, niece, widows, lady, sister... (see Supplementary Materials for full list)]
  Seeds 2: [countryman, fraternal, wizards, manservant, fathers, divo, actor, bachelor, papa, dukes, barman, countrymen, brideprice, hosts, airmen, andropause, penis, prince, governors, abbot, men, widower, gentlemen, sorcerers, sir, bridegrooms, baron, househusbands, gods, nephew, widowers, lord, brother, (see Supplementary Materials for full list)]

- **Used In:** Garg et al. (2018)
  Seeds ID 1:
  names_asian-Garg_et_al_2018
  Seeds ID 2:
  names_chinese-Garg_et_al_2018
  Seeds 1: [cho, wong, tang, huang, chu, chung, ng, wu, liu, chen, lin, yang, kim, chang, shah, wang, li, khan, singh, hong]
  Seeds 2: [chung, liu, wong, huang, ng, hu, chu, chen, lin, liang, wang, wu, yang, tang, chang, hong, li]

- **Used In:** Garg et al. (2018)
  Seeds ID 1:
  names_black-Garg_et_al_2018
  Seeds ID 2:
  names_white-Garg_et_al_2018
  Seeds 1: [harris, robinson, howard, thompson, moore, wright, anderson, clark, jackson, taylor, scott, davis, allen, adams, lewis, williams, jones, wilson, martin, johnson]
  Seeds 2: [harris, nelson, robinson, thompson, moore, wright, anderson, clark, jackson, taylor, scott, davis, allen, adams, lewis, williams, jones, wilson, martin, johnson]

Figure 2

- **Seeds ID:**
  female-Kozlowski_et_al_2019
  Seeds: [woman, women, she, her, her, hers, girl, girls, female, feminine]

- **Seeds ID:**
  female_1-Caliskan_et_al_2017
  Seeds: [sister, female, woman, girl, daughter, she, hers, her]

- **Seeds ID:**
  definitional_female-Bolukbasi_et_al_2016

Seeds: [woman, girl, she, mother, daughter, gal, female, her, herself, Mary]

- **Seeds ID:** female_singular-Hoyle_et_al_2019
  Seeds: [woman, girl, mother, daughter, sister, wife, aunt, niece, empress, queen, princess, duchess, lady, dame, waitress, actress, goddess, policewoman, postwoman, heroine, witch, stewardess, she]

- **Seeds ID:** female_definition_words_2-Zhao_et_al_2018
  Seeds: [lady, saleswoman, noblewoman, hostess, coquette, nun, heroine, actress, chairwoman, businesswoman, spokeswoman, waitress, councilwoman, stateswoman, policewoman, countrywomen, horsewoman, headmistress, governess, widow, witch, fiancee]

- **Seeds ID:** female_stereotype_words-Zhao_et_al_2018
  Seeds: [baker, counselor, nanny, librarians, socialite, assistant, tailor, dancer, hairdresser, cashier, secretary, clerk, stenographer, optometrist, housekeeper, bookkeeper, homemaker, nurse, stylist, receptionist]

Figure 3 (a)

- **Used In:** Bolukbasi et al. (2016)
- **Seeds 1 ID**: definitional_female-Bolukbasi_et_al_2016
- **Seeds 2 ID**: definitional_male-Bolukbasi_et_al_2016
- **Seeds 1:** [she, her, woman, Mary, herself, daughter, mother, gal, girl, female]
- **Seeds 2:** [he, his, man, John, himself, son, father, guy, boy, male]
- **Seeds 1 Shuffled:** [herself, woman, daughter, Mary, her, girl, mother, she, female, gal]
- **Seeds 2 Shuffled:** [man, his, he, son, guy, himself, father, boy, male, John]

Figure 3 (b)

- **Used In:** Kozlowski et al. (2019)
- **Seeds 1 ID**: upperclass-Kozlowski_et_al_2019
- **Seeds 2 ID**: lowerclass-Kozlowski_et_al_2019
- **Seeds 1:** [rich, richer, richest, affluence, affluent, expensive, luxury, opulent]
- **Seeds 2:** [poor, poorer, poorest, poverty, impoverished, inexpensive, cheap, needy]
- **Seeds 1 Shuffled:** [richer, opulent, luxury, affluent, rich, affluence, richest, expensive]
- **Seeds 2 Shuffled:** [poorer, impoverished, poorest, cheap, needy, poverty, inexpensive, poor]

Figure 3 (c)

- **Used In:** Garg et al. (2018)
- **Seeds 1 ID**: names_chinese-Garg_et_al_2018
- **Seeds 2 ID**: names_hispanic-Garg_et_al_2018
- **Seeds 1:** [chung, liu, wong, huang, ng, hu, chu, chen, lin, liang, wang, wu, yang, tang, chang, hong, li]
- **Seeds 2:** [ruiz, alvarez, vargas, castillo, gomez, soto, gonzalez, sanchez, rivera, mendoza, martinez, torres, rodriguez, perez, lopez, medina, diaz, garcia, castro, cruz]
- **Seeds 1 Shuffled:** [tang, chang, chu, yang, wu, hong, huang, wong, hu, liu, lin, chen, liang, chung, li, ng, wang]
- **Seeds 2 Shuffled:** [ruiz, rodriguez, diaz, perez, lopez, vargas, alvarez, garcia, cruz, torres, gonzalez, soto, martinez, medina, rivera, castillo, castro, mendoza, sanchez, gomez]

Figure 4 (a)

- **Used In:** Bolukbasi et al. (2016)
- **Seeds 1 ID**: definitional_female-Bolukbasi_et_al_2016
- **Seeds 2 ID**: definitional_male-Bolukbasi_et_al_2016
- **Seeds 1:** [she, her, woman, Mary, herself, daughter, mother, gal, girl, female]
- **Seeds 2:** [he, his, man, John, himself, son, father, guy, boy, male]

Figure 4 (b)

- **Used In:** N/A (random seeds)
- **Seeds 1 ID**: N/A
- **Seeds 2 ID**: N/A
- **Seeds 1:** [negatives, vel, theirs, canoe, meet, bilingual, mor, facets, fari, lily]
- **Seeds 2:** [chun, brush, dictates, caesar, fewest, breitbart, rod, heaped, julianna, longest]

Figure 4 (c)

- **Used In:** Bolukbasi et al. (2016)
- **Seeds 1 ID**: definitional_female-Bolukbasi_et_al_2016
- **Seeds 2 ID**: definitional_male-Bolukbasi_et_al_2016
- **Shuffled Seeds 1:** [female, she, woman, gal, her, daughter, girl, herself, mother, Mary]
- **Shuffled Seeds 2:** [John, man, son, father, male, himself, guy, he, his]

Figure 5 (Black vs White Names)

- **Used In:** Knoche et al. (2019b)
- **Seeds 1 ID**: white_names-Knoche_et_al_2019
- **Seeds 2 ID**: black_names-Knoche_et_al_2019
- **Seeds 1:** [adam, chip, harry, josh, roger, alan, frank, ian, justin, ryan, andrew, fred, jack, matthew, stephen, brad, greg, jed, paul, todd, brandon, hank, jonathan, peter, wilbur, amanda, courtney, heather, melanie, sara, amber, crystal, katie, meredith, shannon, betsy, donna, kristin, nancy, stephanie, bobbie-sue, ellen, lauren, peggy, sue-ellen, colleen, emily, megan, rachel, wendy, brendan, geoffrey, brett, jay, neil, anne, carrie, jill, laurie, kristen, sarah]
- **Seeds 2:** [alonzo, jamel, lerone, percell, theo, alphonse, jerome, leroy, rasaan, torrance, darnell, lamar, lionel, rashaun, tyree, deion, lamont, malik, terrence, tyrone, everol, lavon, marcellus, terryl, wardell, aiesha, lashelle, nichelle, shereen, temeka, ebony, latisha, shaniqua, tameisha, teretha, jasmine, latonya, shanise, tanisha, tia, lakisha, latoya, sharise, tashika, yolanda, lashandra, malika, shavonn, tawanda, yvette, hakim, jermaine, kareem, jamal, rasheed, aisha, keisha, kenya, tamika]

Figure 5 (Black vs White Roles)

- **Used In:** Manzini et al. (2019)
- **Seeds 1 ID**: black_roles-Manzini_et_al_2019
- **Seeds 2 ID**: caucasian_roles-Manzini_et_al_2019
- **Seeds 1:** [slave, musician, runner, criminal, homeless]
- **Seeds 2:** [manager, executive, redneck, hillbilly, leader, farmer]