Assessing the Reliability of Word Embedding Gender Bias Measures

Yupei Du, Qixiang Fang and Dong Nguyen
Utrecht University
Utrecht, the Netherlands
{y.du,q.fang,d.p.nguyen}@uu.nl

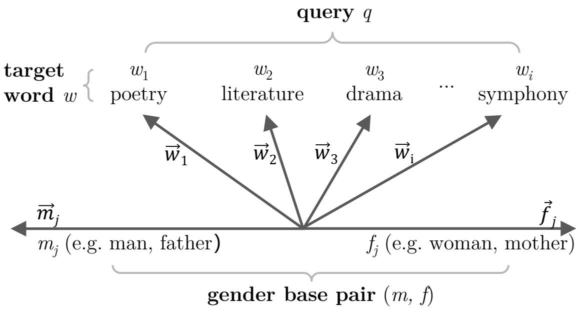
Abstract

Various measures have been proposed to quantify human-like social biases in word embeddings. However, bias scores based on these measures can suffer from measurement error. One indication of measurement quality is *reliability*, concerning the extent to which a measure produces consistent results. In this paper, we assess three types of reliability of word embedding gender bias measures, namely test-retest reliability, inter-rater consistency and internal consistency. Specifically, we investigate the consistency of bias scores across different choices of random seeds, scoring rules and words. Furthermore, we analyse the effects of various factors on these measures' reliability scores. Our findings inform better design of word embedding gender bias measures. Moreover, we urge researchers to be more critical about the application of such measures.[1]

1 Introduction

Despite their success in various applications, word embeddings have been shown to exhibit a range of human-like social biases. For example, Caliskan et al. (2017) find that both GloVe (Pennington et al., 2014) and skip-gram (Mikolov et al., 2013) embeddings associate pleasant terms (e.g. *love* and *peace*) more with European-American names than with African-American names, and that they associate career words (e.g. *profession* and *business*) more with male names than with female names.

Various measures have been proposed to quantify such biases in word embeddings (Ethayarajh et al., 2019; Zhou et al., 2019; Manzini et al., 2019). These measures allow us to assess biases in word embeddings and the performance of bias-mitigation methods (Bolukbasi et al., 2016; Zhao et al., 2018). They also enable us to study social biases in a new way complementary to traditional qualitative and experimental approaches (Garg et al., 2018; Chaloner and Maldonado, 2019).



$$\text{bias}_{w_i}^{m_j, f_j} = \text{scoring rule}(\vec{w}_i, \vec{m}_j, \vec{f}_j)$$
$$\text{bias}_q = \underset{i,j}{\text{aggregate}}(\text{bias}_{w_i}^{m_j, f_j})$$

Figure 1: Measuring word embedding gender biases in the concept of *arts*. The arrows are hypothetical low-dimensional projections of word embeddings.

A key challenge in developing bias measures is that social biases are abstract concepts that cannot be measured directly but have to be inferred from some observable data. This renders the resulting bias scores more prone to measurement errors. Therefore, it is important to carefully assess these measures' measurement quality. Jacobs and Wallach (2021) also highlight similar measurement issues in the context of fairness research.

In this paper, we focus on one aspect of measurement quality: **reliability**. It concerns the extent to which a measure produces consistent results. In particular, we investigate the reliability of word embedding **gender bias** measures.

Figure 1 illustrates how gender biases in word embeddings are typically measured, with measuring gender biases in the concept of *arts* as an example. To calculate the gender bias score of a **target word** w (e.g. *poetry*, a word that relates to the concept of interest), we need to specify a **gender base pair** (m, f) (a gendered word pair, e.g. *father/mother*) and **a scoring rule**. A scoring rule is a function that takes the embedding vectors of w,

[1] Our code is available at https://github.com/nlpsoc/reliability_bias.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10012–10034
November 7–11, 2021. ©2021 Association for Computational Linguistics

m and f as input and returns a bias score $bias_w^{m,f}$ as output. In practice, often an **ensemble of gender base pairs** are used. In this case, we aggregate (e.g. average) all the bias scores w.r.t. different gender base pairs to obtain an overall bias score for a target word. Furthermore, multiple conceptually related target words (e.g. *poetry*, *drama*) may be specified to form a **query** q (e.g. *arts*). By aggregating individual bias scores of these target words, we can compute an overall bias score for a concept.

Clearly, the choice of target words, gender base pairs and scoring rules may influence the resulting bias scores. Reliable measurements of word embedding gender biases thus require target words, gender base pairs and scoring rules that produce consistent results. In this work, by drawing from measurement theory, we propose a comprehensive approach (§4) to evaluate three types of reliability for these different components:

- First, we assess the consistency of bias scores associated with different target words, gender base pairs and scoring rules, over different random seeds used in word embedding training (i.e. **test-retest reliability**; §5.2).
- Second, we assess the consistency of bias scores associated with different target words and gender base pairs, across different scoring rules (i.e. **inter-rater consistency**; §5.3).
- Third, we assess the consistency of bias scores over 1) different target words within a query and 2) different gender base pairs (i.e. **internal consistency**; §5.4).

Furthermore, we use multilevel regression to model the effects of various factors (e.g. word properties, embedding algorithms, training corpora) on the reliability scores of target words (§5.5).

Our experiments show that word embedding gender bias scores are mostly consistent across different random seeds (i.e. high test-retest reliability) and across target words within the same query (i.e. high internal consistency). However, different scoring rules generally fail to agree with one another (i.e. low inter-rater consistency). Moreover, word embedding algorithms have a large influence on the reliability of bias scores.

Contributions First, we connect measurement theory to word embedding bias measures. Specifically, we propose a reliability evaluation framework for word embedding (gender) bias measures. Second, we provide a comprehensive assessment of the reliability of word embedding gender bias measures. Based on our findings, we urge researchers to be more critical about applying such measures.

2 Related Work

Measuring gender biases in word embeddings has been receiving a growing amount of research interest in NLP. Various gender bias measures have been proposed. They are based on different techniques, such as linear gender subspace identification (Bolukbasi et al., 2016; Vargas and Cotterell, 2020; Manzini et al., 2019), psychological tests (Ethayarajh et al., 2019; Caliskan et al., 2017), inference from nearest neighbours (Gonen and Goldberg, 2019) and regression (Sweeney and Najafian, 2019; Badilla et al., 2020).

However, recent studies have raised concerns over the reliability of such measures. Zhang et al. (2020) show that gender bias scores easily vary in their direction and magnitude when different forms (e.g. capitalisation) of target words or different gender base pairs are used. Similarly, Antoniak and Mimno (2021) look into 178 different gender base pairs from previous works and find that the choice of gender base pairs can greatly impact bias measurements. They therefore urge future work to examine and document the choices of gender base pairs. Moreover, D'Amour et al. (2020) find that underspecification of models can lead to unstable contextualised word embedding bias scores. These findings call for a more systematic evaluation of the reliability of word embedding gender bias measures, which is the goal of our study.

Such measures' lack of reliability may partly stem from the fact that word embeddings themselves are often unstable, sensitive to choices of, for instance, word embedding algorithms (Wendlandt et al., 2018; Antoniak and Mimno, 2018; Hellrich et al., 2019), hyper-parameters (Levy et al., 2015; Mimno and Thompson, 2017; Hellrich et al., 2019) and even random seeds (Wendlandt et al., 2018; Hellrich and Hahn, 2016; Bloem et al., 2019) during word embedding training.

Various word (embedding) attributes have been found to contribute to the instability of word embeddings, including part-of-speech tags (Wendlandt et al., 2018; Pierrejean and Tanguy, 2018), word frequency (Hellrich and Hahn, 2016; Pierrejean and Tanguy, 2018) and word ambiguity (Wendlandt et al., 2018; Hellrich and Hahn, 2016).

3 Preliminaries

In this section, we first review three popular scoring rules used for measuring word embedding gender biases (§3.1). Then, we introduce the conceptual framework of reliability and motivate its use in word embedding gender bias measurements (§3.2).

3.1 Scoring Rules

Following Zhang et al. (2020), we focus on three popular scoring rules: DB/WA, RIPA and NBM.[2]

DB/WA DB/WA (Direct Bias / Word Association) is one of the most commonly used scoring rules in previous work (Bolukbasi et al., 2016; Caliskan et al., 2017). Given a gender base pair (m, f), the DB/WA score of a target word w is

$$\text{DB/WA}_w^{(m,f)} = \cos(\vec{w}, \vec{m}) - \cos(\vec{w}, \vec{f}),$$

where $\vec{*}$ is the corresponding word vector of $*$, and $\cos(x, y)$ refers to the cosine similarity of x and y.

RIPA Another scoring rule based on vector similarity is Relational Inner Product Association (RIPA; Ethayarajh et al., 2019). The main difference between DB/WA and RIPA is that RIPA performs normalisation at the gender base pair level instead of at the word level. Formally,

$$\text{RIPA}_w^{(m,f)} = \vec{w} \cdot \frac{\vec{m} - \vec{f}}{||\vec{m} - \vec{f}||},$$

where $||*||$ refers to the L2 norm of $*$.

NBM Unlike DB/WA and RIPA, which are based on vector similarities, NBM (Neighbourhood Bias Metric) is based on a word's k nearest neighbours (Gonen and Goldberg, 2019). Specifically,

$$\text{NBM}_w^{(m,f)} = \frac{|masculine(w)| - |feminine(w)|}{k},$$

where $|masculine(w)|$ and $|feminine(w)|$ are the number of words in w's k nearest neighbours biased towards the respective gender based on their DB/WA scores. Following Zhang et al. (2020) and Gonen and Goldberg (2019), we use $k = 100$.

3.2 Reliability

In measurement theory, reliability is the extent to which a measure produces consistent results over a variety of measurement conditions, in which basically the same results should be obtained (Drost, 2011). In this work, we focus on three important types of reliability: test-retest reliability, inter-rater consistency and internal consistency.

Test-retest reliability concerns the consistency of measurements across different measurement occasions (assuming no substantial change in the true value, Weir, 2005). For example, if gender bias scores vary substantially across different measurement occasions (e.g. different random seeds during embedding training; different random data samples), they should be considered to have low test-retest reliability. In this case, derived conclusions from these scores are likely to be untrustworthy.

Inter-rater consistency is the degree to which different raters produce consistent measurements (Shrout and Fleiss, 1979). For example, consider scoring rules as the raters of word embedding bias scores. In this case, if different scoring rules measure gender biases in a similar way, they should produce bias scores that tend to agree with one another in both signs and normalised magnitude.

Internal consistency is defined as the agreement among multiple components that make up a measure of a single construct (Cronbach, 1951). In the example from Figure 1, we specify a query consisting of various *arts*-related target words to measure gender biases of the concept *arts*. We then compute individual bias scores for all target words before aggregating them to obtain an overall bias score. If the bias scores of target words are distinct from one another (i.e. low or negative correlation), the query has low internal consistency. In this case, one should question whether these target words measure the *arts* concept in comparable ways.

4 Estimating Reliability of Word Embedding Gender Bias Measures

In this section, we propose an evaluation framework to assess the reliability of word embedding gender bias measures. Respectively, we present our operational definitions for test-retest reliability (§4.1), inter-rater consistency (§4.2) and internal consistency (§4.3). See Table 1 for an overview.

Notation Suppose we have s scoring rules, g gender base pairs and t target words. We train k word

[2] This terminology is also adopted from Zhang et al. (2020).

	Test-retest reliability		Inter-rater consistency		Internal consistency	
Component	Target words	Gender base pairs	Target words	Gender base pairs	Gender base pair ensemble	Queries
Source of variations	Different random seeds		Different scoring rules		Individual gender base pairs	Individual target words

Table 1: Operational definitions for different types of reliability.

embedding models with the same hyper-parameters but k different random seeds. For each scoring rule, we calculate the bias score of each target word w.r.t each gender base pair, on each word embedding model. As a result, we get a four-dimensional bias scores matrix B of $\mathbb{R}^{s\times g\times t\times k}$. For calculating inter-rater consistency and internal consistency, we average the gender bias scores derived from the k word embedding models. Averaging these embedding models can partial out the influence of random seeds, and therefore lead to more accurate estimation of other types of reliability. In this way, we get another bias matrix B' of $\mathbb{R}^{s\times g\times t}$.

4.1 Test-retest Reliability

We measure test-retest reliability as the consistency of bias scores associated with each target word, gender base pair and scoring rule across different random seeds[3]. Specifically, we focus on two questions: First, are the bias scores of a single target word (averaged across gender base pairs) consistent across random seeds (i.e. **test-retest reliability of target words**)? Second, are the average bias scores of all target words w.r.t one gender base pair consistent across random seeds (i.e. **test-retest reliability of gender base pairs**)? We explore both questions for each scoring rule separately.

To calculate test-retest reliability for each target word, we use a bias score matrix of size $\mathbb{R}^{g\times k}$ by slicing B. We then use intra-class correlation (ICC) to estimate test-retest reliability.

ICC is a popular family of estimators for both test-retest reliability and inter-rater consistency. There are different forms of ICC estimators, each of which can involve distinct assumptions and can therefore lead to very different interpretations (Koo and Li, 2016). Shrout and Fleiss (1979) define 6 forms of ICC and present them as "ICC" with 2 additional numbers in parentheses (e.g., ICC(2,1) and ICC(3,1)). The first number refers to the **model** and can take on three possible values (1, 2 or 3). The second number refers to the **intended use** of raters/measurements in an application and can take on two values (1 or k). See Appendix C for a detailed description of these value options. We adopt ICC(2,1) as the estimator for test-retest reliability:

$$\mathcal{I}^{(2,1)} = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{t}(MS_C - MS_E)} \quad (1)$$

where MS_E, MS_R, and MS_C are the mean square of error, of rows and of columns, respectively[4].

Similarly, for the test-retest reliability of each gender base pair, we slice B to get a bias score matrix of size $\mathbb{R}^{t\times k}$. We then calculate its ICC value using Equation 1 (by substituting t with g).

4.2 Inter-rater Consistency

We define inter-rater consistency as the consistency of bias scores across different scoring rules. We again investigate two questions: First, are the bias scores of a single target word (averaged across all gender base pairs) consistent across scoring rules (i.e. **inter-rater consistency of a target word**)? Second, are the average bias scores of all target words w.r.t. a single gender base pair consistent across scoring rules (i.e. **inter-rater consistency of a gender base pair**)?

To calculate the inter-rater consistency of each target word, we use a bias score matrix of size $\mathbb{R}^{g\times s}$ by slicing and transposing B'. Following Koo and Li (2016), we adopt ICC(3, 1) as the estimator:

$$\mathcal{I}^{(3,1)} = \frac{MS_R - MS_E}{MS_R + (s-1)MS_E}. \quad (2)$$

Similarly, for the second question, we can get a bias score matrix size of $\mathbb{R}^{t\times s}$ for each gender base pair and calculate the inter-rater consistency of the gender base pair via Equation 2.

[3] One may argue that we can use pre-trained word embeddings. However, deterministic alternatives are not free of reliability issues. Instead, they are of "fixed randomness" (Hellrich and Hahn, 2016).

[4] See our code for the exact implementation of the different reliability estimators

4.3 Internal Consistency

We investigate the internal consistency of both queries and the ensemble of gender base pairs. We thus focus on two questions: First, are the bias scores of different target words within a query consistent (i.e. **internal consistency of a query**)? Second, are the average bias scores of all target words consistent across different gender base pairs (i.e. **internal consistency of the ensemble of gender base pairs**)? We examine both questions for each scoring rule separately.

To calculate the internal consistency of a query consisting of t target words, we first slice and transpose B' to get a bias score matrix size of $\mathbb{R}^{g \times t}$. We then use Cronbach's alpha (Cronbach, 1951) as the estimator of internal consistency. Cronbach's alpha is the most common estimator for internal consistency, which assesses how closely related a set of test items are as a group (e.g. different target words of the same query). Specifically,

$$\alpha = \frac{t}{t-1}\left(1 - \frac{\sum_{i=1}^{t} \sigma_i^2}{\sigma_X^2}\right), \tag{3}$$

where σ_i^2 is the variance of the bias scores of target word i in the query w.r.t. different gender base pairs. σ_X^2 is the sum of $\sum_{i=1}^{t} \sigma_i^2$ and all covariances of bias scores between target words.

We calculate the internal consistency of the ensemble of gender base pairs in a similar way, by creating a bias score matrix size of $\mathbb{R}^{t \times g}$ and applying Equation 3 (substituting t with g).

5 Experiments

5.1 Experimental Setup

Training Embeddings We select three corpora with different characteristics to train word embeddings. Two are from subReddits: r/AskScience ($\sim$ 158 million tokens) and r/AskHistorians ($\sim$ 137 million tokens, also used by Antoniak and Mimno 2018)[5]. The third is the training set of WikiText-103 ($\sim$ 527 million tokens, Merity et al., 2016), consisting of high-quality Wikipedia articles.

We use two popular word embedding algorithms: Skip-Gram with Negative Sampling (SGNS; Mikolov et al. 2013) and GloVe (Pennington et al., 2014). For both algorithms, we set the number of embedding dimensions to 300. For all other hyper-parameters, we use the default values of previous implementations.[6] For each corpus-algorithm pair, we train $k = 32$ word embedding models using different random seeds.

Gender Base Pairs We collect and lower-case all 23 gender base pairs from Bolukbasi et al. (2016) and Garg et al. (2018).

Target Word Lists For the assessment of test-retest reliability and inter-rater consistency, we include three word lists used in previous word embedding bias studies: 1) 320 occupation words from Bolukbasi et al. (2016) (OCC16), 2) 76 occupation words from Garg et al. (2018) (OCC18) and 3) 230 adjectives from Garg et al. (2018) (ADJ). However, these three lists are very specific (i.e. only concerning occupation words and adjectives) and thus unlikely applicable to other (future) research where different biases are of interest and different target words might be used (e.g. measuring gender biases of a whole corpus). Therefore, we also consider two additional, larger target word lists: 4) the top 10,000 most frequent words of Google's trillion word corpus (Google10K)[7] and 5) the full vocabulary of each corpus (Full).

Queries For the assessment of internal consistency, we examine six gender bias related queries from Caliskan et al. (2017): *math*, *arts*, *arts_2*, *career*, *science*, and *family*, each consisting of eight target words. Note that target word lists are different from queries. The former does not necessarily consist of conceptually related words.

5.2 Results: Test-retest Reliability

Figure 2 shows the distribution of test-retest reliability scores of target words and gender base pairs across target word lists and scoring rules. Here, word embeddings are trained with SGNS on WikiText-103. Similar results are found for other corpora and algorithms (see Appendix D.1).

[5] We use all posts and replies from 2014, retrieved from https://archive.org/download/2015_reddit_comments_corpus/reddit_data/2014/.

[6] For SGNS, we use Gensim 3.8.3 (Řehůřek and Sojka, 2010) with a window size of 5, a minimum word count of 5 and 5 iterations. For GloVe, we use the official implementation https://github.com/stanfordnlp/GloVe, with a window size of 15, a minimum word count of 5 and 15 iterations. Because we do not fine-tune hyper-parameters, our results do not necessarily indicate which algorithm itself is of better reliability. To investigate the potential impact of this decision, we also include an explorative study on the influence of hyper-parameters in Appendix E.

[7] https://github.com/first20hours/google-10000-english.

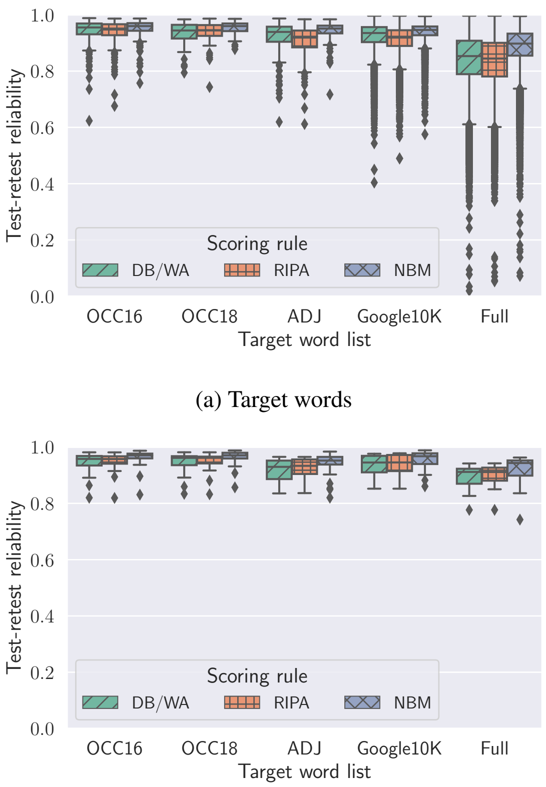
First, we observe that the majority of target words and gender base pairs have acceptable test-retest reliability with ICC values greater than 0.6, regardless of the scoring rule used.[8] Nevertheless, quite some target words and gender base pairs fall below the lower whiskers of the box-plots (indicating low test-retest reliability).

Moreover, comparing with Google10K, which consists of frequent words, a higher proportion of words in the full vocabulary have very low test-retest reliability. For example, 0.01% of the target words in Google10K have a test-retest reliability lower than 0.50 for word embeddings trained with SGNS on WikiText-103. In contrast, for the full vocabulary this is 0.1%, approximately 10 times that of Google10K. These results suggest that we should be careful when making word lists that consist of infrequent words (e.g. when studying less common concepts). If we do need to use infrequent words, we should check their test-retest reliability before deriving further conclusions.

5.3 Results: Inter-rater Consistency

Figure 3 shows the distributions of inter-rater consistency scores of both target words and gender base pairs across different corpora (word embeddings trained by GloVe). More (similar) results on different algorithms are in Appendix D.2.

We observe that the inter-rater consistency of the majority of both target words and gender base pairs are rather low. This finding suggests that different scoring rules may measure very different aspects of word embedding gender biases, and hence their resulting bias scores differ substantially. More closely, we observe that for target words, the bias scores are the least similar between RIPA and NBM (Pearson's r: 0.836, $p < 0.05$), while they are much more similar between DB/WA and RIPA (Pearson's r: 0.923, $p < 0.05$), and between DB/WA and NBM (Pearson's r: 0.897, $p < 0.05$). A possible reason is that DB/WA and RIPA scores are both based on cosine similarities, and that NBM scores are based on DB/WA scores of the closest neighbours. In contrast, RIPA and NBM scores are computed in less comparable ways. Nevertheless, future studies are needed to further investigate the differences among scoring rules.



(a) Target words

(b) Gender base pairs

Figure 2: Distribution of the test-retest reliability scores across target word lists and scoring rules (based on SGNS and WikiText-103). Most target words and gender base pairs have acceptable test-retest reliability across the scoring rules. Nevertheless, some outliers with low test-retest reliability exist, especially for target word lists with infrequent words.

5.4 Results: Internal Consistency

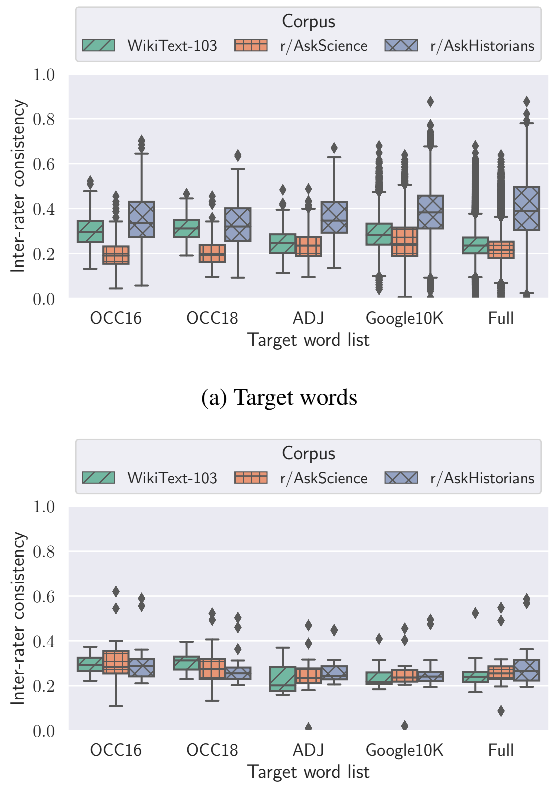
Figure 4 presents the distribution of the internal consistency scores of every query and the ensemble of all gender base pairs across corpora. Each boxplot contains six scores from the combinations of embedding algorithms and scoring rules. We make four observations:

First, the internal consistency of most queries and the ensemble of gender base pairs are acceptable (Cronbach's alpha values ≥ 0.7).[9] This indicates that most target words in the same query likely measure gender bias of the same concept. Bias scores of a target word are also generally consistent across gender base pairs.

Second, however, the patterns of internal consistency vary substantially across queries. For example, on the *WikiText-103* corpus, the internal

[8] Generally, ICC values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.9 are considered to have poor, moderate, good, and excellent reliability, respectively (Koo and Li, 2016). This also holds for ICC values of inter-rater consistency (Section 5.3).

[9] Cronbach's alpha values greater than 0.7 are considered acceptable. See Cicchetti (1994).



(a) Target words

(b) Gender base pairs

Figure 3: Distribution of inter-rater consistency scores across target word lists and corpora (based on GloVe). For both target words and gender base pairs, we observe generally low consistency in bias scores across the three scoring rules, regardless of the corpus used.

consistency scores of *family* are much higher and less varied than the scores of *math*.

Third, the internal consistency of a query and the ensemble of gender base pairs seems dependent on specific corpora. For instance, the internal consistency scores of *math* are high and have a low variance on the corpus *r/AskScience*, but they are low and have a very high variance on *r/AskHistorians*.

Fourth, the high variance of scores for some queries (e.g. *math* on *r/AskHistorians*) suggests that a query's internal consistency may depend also on word embedding algorithms and scoring rules.

5.5 Factors Influencing the Reliability of Gender Base Pairs and Target Words

In this section, we investigate factors influencing the test-retest reliability and inter-rater consistency of both gender base pairs and target words. Because we only have a small number of gender base pairs, we qualitatively inspect them using visualisations. For (the large number of) target words, we use regression to model the effects of the factors.

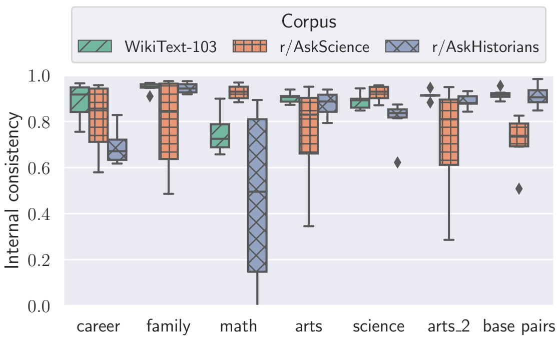


Figure 4: Distribution of internal consistency scores of gender bias related queries (e.g., *career*) and the ensemble of gender base pairs (*base pairs*) across corpora. Each query and the ensemble of gender base pairs has six reliability scores across different combinations of embedding algorithms and scoring rules. While many queries and the ensemble of gender base pairs show overall acceptable internal consistency, the specific scores can still highly depend on the specific query, corpus, scoring rule and even word embedding algorithm used.

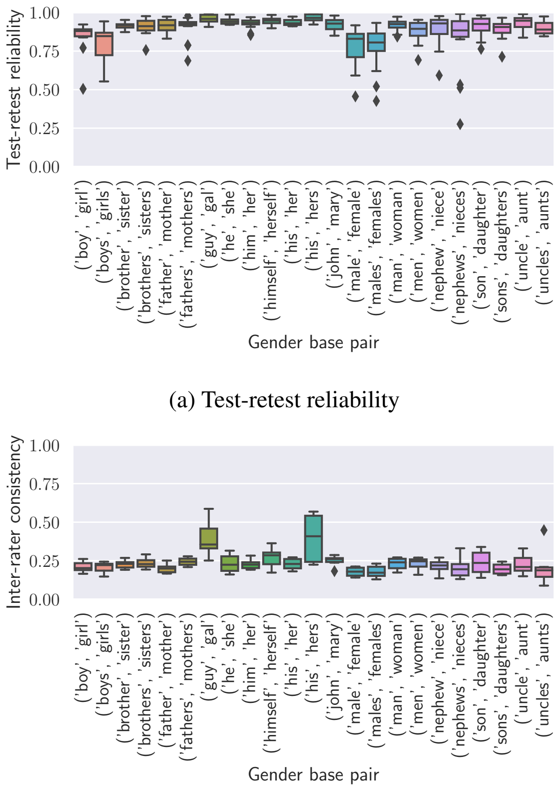
Gender Base Pairs: Visualisation The distributions of test-retest reliability and inter-rater consistency of gender base pairs (on full vocabularies) are shown in Figure 5. We make two observations:

First, gender base pairs in singular form are usually of higher test-retest reliability (e.g. *boy*∼*girl* versus *boys*∼*girls*), which is consistent with findings by Zhang et al. (2020). The median difference in test-retest reliability between singular and plural gender base pairs is statistically significant ($t(7) = 2.45, p < .05$). In contrast, such a statistical difference is not found for inter-rater consistency ($t(7) = -0.13, p > .05$).

Second, gender base pairs of higher test-retest reliability also tend to be of higher inter-rater consistency, evidenced by the moderate correlation between the median test-retest reliability scores and the median inter-rater consistency scores of gender base pairs ($r = 0.644$, $p < .05$).

Target Words: Regression Analyses We use multilevel regression to study potential influencing factors of the test-retest reliability and inter-rater consistency of target words.[10] Comparing with OLS regression and its variants, multilevel models allow for dependent observations. Therefore, they suit our data better where reliability scores are

[10] We offer a more detailed description of multilevel regression, the used features and PoS estimates in Appendix F.



(a) Test-retest reliability

(b) Inter-rater consistency

Figure 5: Test-retest reliability and inter-rater consistency of different gender base pairs on full vocabularies. Each gender base pair has multiple reliability scores across combinations of embedding algorithms and corpora (as well as scoring rules for test-retest reliability). Gender base pairs in singular form tend to have higher test-retest reliability. Also, gender base pairs with higher test-retest reliability are more likely to score higher in inter-rater consistency.

nested within groups (e.g. different training algorithms and corpora of embeddings) and are thus correlated. Multilevel models have a further advantage that they estimate not only the effects of fixed factors (i.e. standard features) but also the amount of variance explained by each grouping factor.

We collect a range of word-level features as fixed factors, mostly inspired by previous studies (Wendlandt et al., 2018; Pierrejean and Tanguy, 2018; Hellrich and Hahn, 2016). These include 1) word-intrinsic features: log number of WordNet synsets (log #senses) and the most common Part-of-Speech tag (PoS) in the Brown corpus (Francis and Kucera, 1979), as in Wendlandt et al. (2018); 2) corpus-related features: log frequencies of words in the training corpus (log freq) and their squares ($\log^2$ freq); 3) embedding-related features: cosine

	Test-retest		Inter-rater	
	Estimate	ΔR^2	Estimate	ΔR^2
$\text{SR}_{\text{DB/WA}}$	reference	0.0040	-	-
SR_{RIPA}	**-0.0102**	-	-	-
SR_{NBM}	-0.0001	-	-	-
log freq	**0.0062**	0.0275	**0.0241**	0
$\log^2$ freq	**0.0023**	0.0051	**0.0008**	0
log #senses	**-0.0012**	0.0002	**-0.0007**	0
PoS	-	0.0093	-	0.0022
NN Sim	**-0.0024**	0.0120	**0.0011**	0.0038
L2 norm	**-0.0118**	0.0456	**-0.0051**	0
ES	**0.0457**	0.1020	**0.0186**	0
R^2_{fixed}	0.3261	-	0.0581	-
R^2_{corpus}	0.0113	-	0.0271	-
$R^2_{\text{algorithm}}$	0.1802	-	0.4307	-
R^2_{total}	0.5177	-	0.5160	-

Table 2: Results of multilevel regression on the test-retest and inter-rater reliability of target words. Estimates are standardised (bold if $p < 0.05$). SR: scoring rule. ΔR^2 is the reduction in explained variance when a factor is left out. R^2_{fixed}, R^2_{corpus}, $R^2_{\text{algorithm}}$ and R^2_{total} refer to the explained variance of fixed factors (i.e. word-level features and scoring rules), embedding training corpora, embedding training algorithms, total effects of all these three parts, respectively.

similarity to nearest neighbour (NN Sim), L2 norm (L2 Norm) and embedding stability (ES). We calculate ES as follows: for each pair of word embedding models, we first fit an orthogonal transformation Q that minimizes the Frobenius norm of their difference. The stability of a word across multiple random seeds is then calculated as the average pairwise cosine similarity of its embedding vectors after transformation by Qs. We also consider scoring rules as a fixed factor because we are interested in comparing the influence of these three scoring rules on target words' test-retest reliability. The two grouping factors are embedding algorithms and training corpora.

We summarise the results in Table 2. For test-retest reliability, the model has a satisfactory total explained variance ($R^2_{\text{model}} : 51.77\%$). Fixed factors (including scoring rules) together explain a substantial part of the variation ($R^2_{\text{fixed}} : 32.61\%$). Among these factors, embedding stability (ES) appears to be the most important one, indicated by the largest standardised effect estimate and ΔR^2. The higher the embedding stability, the higher the test-retest reliability, which is expected. L2 norm and word frequency also account for a considerable amount of variance. When the L2 norm is lower or when the frequency is higher, test-retest reliability is higher. This observation is also consistent with prior research findings. For instance, Hellrich and Hahn (2016) show that word frequency positively correlates with embedding stability when word frequency is not too high. Also, Arora et al. (2016) find that the L2 norm correlates negatively with word frequency. This finding agrees with our observation in Figure 2 as well. In contrast, the choice of scoring rules has only a minor impact on test-retest reliability (R^2 : 0.4%). Despite a statistically significant difference between DB/WA and RIPA, the difference is very small (-0.0102) and therefore unlikely important.

Among the group level factors, embedding algorithms alone explain 18.02% of the total variance. This suggests that the test-retest reliability of a target word is determined by the word embedding training to a considerable degree. In contrast, the choice of corpora is much less influential.

For inter-rater consistency, the resulting model is also of good explanatory power (R^2_{model} : 51.60%). However, it is clear that word-level features fail to explain much of the variance (5.81%). Between the two grouping factors, algorithm dominates with an R^2 score of 0.4307. This indicates that the inter-rater reliability of a target word is largely determined by the word embedding algorithm used.

Note that we also explored potential interactions between the fixed factors and how they might impact the outcome test-retest and inter-rater reliability scores. However, it turned out that interaction effects generally had small effect sizes and did not considerably improve overall model fit. We therefore excluded them from the final models.

6 Conclusion & Discussion

In this paper, we propose to leverage measurement theory to examine the reliability of word embedding bias measures. We find that bias scores are mostly consistent across different random seeds (i.e. high test-retest reliability), as well as across gender base pairs and target words within a query (i.e. high internal consistency). In contrast, the three scoring rules fail to agree with one another (i.e. low inter-rater consistency). Furthermore, our regression results suggest that the consistency of bias scores across different random seeds are mostly influenced by various word-level features as well as the word embedding algorithm used. Meanwhile, the bias scores of target words across different scoring rules are dominated by the word embedding algorithm used. We thus urge future studies to be more critical about applying such measures.

Nevertheless, our work has limitations. First, we only consider gender bias measures. Future work should apply our reliability evaluation framework to other types of bias (e.g. racial bias). Second, we focus on static word embeddings. Future work should investigate the reliability of bias measures for contextualised embeddings. Third, we do not address validity, the other crucial aspect of measurement quality. We thus call for future studies on the validity of word embedding bias measures. Fourth, Goldfarb-Tarrant et al. (2021) argue that intrinsic (word embeddings) biases sometimes fail to agree with extrinsic biases (measured in downstream tasks, e.g. coreference resolution). One potential research direction is to assess the reliability of extrinsic bias measurements as well, to shed further light on the disconnect between intrinsic and extrinsic biases. Lastly, while ICC and Cronbach's Alpha are established reliability estimators in many scientific disciplines, correct interpretation of their values is often challenging and requires both statistical and field-specific expertise (Lee et al., 2012; Streiner, 2003). Future work should address the appropriate use of these estimators and their limitations in the context of NLP research.

Acknowledgements

We thank all anonymous reviewers for their constructive and helpful feedback. We also thank Anna Wegmann for the proofreading and productive discussions. This work was partially supported by the Dutch Research Council (NWO) (grant number: VI.Veni.192.130 to D. Nguyen; grant number: VI.Vidi.195.152 to D. L. Oberski).

Ethical Statement

Intended Usage As aforementioned in §1, word embedding bias measures are often used to analyse word embedding models, to assess the effect of bias mitigation methods, and to study societal biases. Our work thus intends to evaluate the quality of these measures and their derived conclusions. Moreover, our framework can also be used to assess the reliability of bias measures which consists of target words, gender base pairs, and scoring rules that were not included in this study. In this way, our framework can contribute to the development of models that are less biased and hence potentially less harmful.

Limitations In this study, we focused on common measures of gender biases in word embeddings. Measurements of gender biases in word embeddings typically rely on manually crafted sets of target words and pairs of gendered words (i.e. gender base pairs, such as *he* vs. *she*). In our experiments we use existing lists of words and word pairs that have been frequently used in related work. However, these word pairs were constructed by taking the very narrow view of binary gender. We hope to see more work on measures of bias in embeddings that considers non-binary gender identities as well as intersectional identities.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 430–436. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jelke Bloem, Antske Fokkens, and Aurélie Herbelot. 2019. Evaluating the consistency of word embeddings from small data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 132–141, Varna, Bulgaria. INCOMA Ltd.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290. Place: US Publisher: American Psychological Association.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Ellen A. Drost. 2011. Validity and reliability in social science research. *Education Research and Perspectives*, 38(1):105–123.

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.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

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.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

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.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Johannes Hellrich, Bernd Kampe, and Udo Hahn. 2019. The influence of down-sampling strategies on SVD word embedding stability. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 18–26, Minneapolis, USA. Association for Computational Linguistics.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kyoung Min Lee, Jaebong Lee, Chin Youb Chung, Soyeon Ahn, Ki Hyuk Sung, Tae Won Kim, Hui Jong Lee, and Moon Seok Park. 2012. Pitfalls and Important Issues in Testing Reliability Using Intraclass Correlation Coefficients in Orthopaedic Research. *Clinics in Orthopedic Surgery*, 4(2):149–155.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

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. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Predicting word embeddings variability. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 154–159, New Orleans, Louisiana. Association for Computational Linguistics.

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.

P. E. Shrout and J. L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

David L. Streiner. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, 80(1):99–103.

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.

Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. 2015. Lexical comparison between Wikipedia and Twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661, Beijing, China. Association for Computational Linguistics.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2902–2913, Online. Association for Computational Linguistics.

Joseph P. Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1):231–240.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 759–769, Suzhou, China. 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.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

A Word Lists

A.1 Gender Base Pairs (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018).

boy ∼ girl, boys ∼ girls, brother ∼ sister, brothers ∼ sisters, father ∼ mother, fathers ∼ mothers, guy ∼ gal, he ∼ she, him ∼ her, himself ∼ herself, his ∼ her, his ∼ hers, john ∼ mary, male ∼ female, males ∼ females, man ∼ woman, men ∼ women, nephew ∼ niece, nephews ∼ nieces, son ∼ daughter, sons ∼ daughters, uncle ∼ aunt, uncles ∼ aunts.

A.2 Target Word Lists

OCC16 (Bolukbasi et al., 2016) accountant, acquaintance, actor, actress, adjunct_professor, administrator, adventurer, advocate, aide, alderman, alter_ego, ambassador, analyst, anthropologist, archaeologist, archbishop, architect, artist, artiste, assassin, assistant_professor, associate_dean, associate_professor, astronaut, astronomer, athlete, athletic_director, attorney, author, baker, ballerina, ballplayer, banker, barber, baron, barrister, bartender, biologist, bishop, bodyguard, bookkeeper, boss, boxer, broadcaster, broker, bureaucrat, businessman, businesswoman, butcher, butler, cab_driver, cabbie, cameraman, campaigner, captain, cardiologist, caretaker, carpenter, cartoonist, cellist, chancellor, chaplain, character, chef, chemist, choreographer, cinematographer, citizen, civil_servant, cleric, clerk, coach, collector, colonel, columnist, comedian, comic, commander, commentator, commissioner, composer, conductor, confesses, congressman, constable, consultant, cop, correspondent, councilman, councilor, counselor, critic, crooner, crusader, curator, custodian, dad, dancer, dean, dentist, deputy, dermatologist, detective, diplomat, director, disc_jockey, doctor, doctoral_student, drug_addict, drummer, economics_professor, economist, editor, educator, electrician, employee, entertainer, entrepreneur, environmentalist, envoy, epidemiologist, evangelist, farmer, fashion_designer, fighter_pilot, filmmaker, financier, firebrand, firefighter, fireman, fisherman, footballer, foreman, freelance_writer, gangster, gardener, geologist, goalkeeper, graphic_designer, guidance_counselor, guitarist, hairdresser, handyman, headmaster, historian, hitman, homemaker, hooker, housekeeper, housewife, illustrator, industrialist, infielder, inspector, instructor, interior_designer, inventor, investigator, investment_banker, janitor, jeweler, journalist, judge, jurist, laborer, landlord, lawmaker, lawyer, lecturer, legislator, librarian, lieutenant, lifeguard, lyricist, maestro, magician, magistrate, maid, major_leaguer, manager, marksman, marshal, mathematician, mechanic, mediator, medic, midfielder, minister, missionary, mobster, monk, musician, nanny, narrator, naturalist, negotiator, neurologist, neurosurgeon, novelist, nun, nurse, observer, officer, organist, painter, paralegal, parishioner, parliamentarian, pastor, pathologist, patrolman, pediatrician, performer, pharmacist, philanthropist, philosopher, photographer, photojournalist, physician, physicist, pianist, planner, plastic_surgeon, playwright, plumber, poet, policeman, politician, pollster, preacher, president, priest, principal, prisoner, professor, professor_emeritus, programmer, promoter, proprietor, prosecutor, protagonist, protege, protester, provost, psychiatrist, psychologist, publicist, pundit, rabbi, radiologist, ranger, realtor, receptionist, registered_nurse, researcher, restaurateur, sailor, saint, salesman, saxophonist, scholar, scientist, screenwriter, sculptor, secretary, senator, sergeant, servant, serviceman, sheriff_deputy, shopkeeper, singer, singer_songwriter, skipper, socialite, sociologist, soft_spoken, soldier, solicitor, solicitor_general, soloist, sportsman, sportswriter, statesman, steward, stockbroker, strategist, student, stylist, substitute, superintendent, surgeon, surveyor, swimmer, taxi_driver, teacher, technician, teenager, therapist, trader, treasurer, trooper, trucker, trumpeter, tutor, tycoon, undersecretary, understudy, valedictorian, vice_chancellor, violinist, vocalist, waiter, waitress, warden, warrior, welder, worker, wrestler, write.

OCC18 (Garg et al., 2018) janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper, accountant, physicist, gardener, dentist, weaver, blacksmith, psychologist, supervisor, mathematician, surveyor, tailor, designer, economist, mechanic, laborer, postmaster, broker, chemist, librarian, attendant, clerical, musician, porter, scientist, carpenter, sailor, instructor, sheriff, pilot, inspector, mason, baker, administrator, architect, collector, operator, surgeon, driver, painter, conductor, nurse, cook, engineer, retired, sales, lawyer, clergy, physician, farmer, clerk, manager, guard, artist, smith, official, police, doctor, professor, student, judge, teacher, author, secretary, soldier.

ADJ (Garg et al., 2018) headstrong, thankless, tactful, distrustful, quarrelsome, effeminate, fickle, talkative, dependable, resentful, sarcastic, unassuming, changeable, resourceful, persevering, forgiving, assertive, individualistic, vindictive, sophisticated, deceitful, impulsive, sociable, methodical, idealistic, thrifty, outgoing, intolerant, autocratic, conceited, inventive, dreamy, appreciative, forgetful, forceful, submissive, pessimistic, versatile, adaptable, reflective, inhibited, outspoken, quitting, unselfish, immature, painstaking, leisurely, infantile, sly, praising, cynical, irresponsible, arrogant, obliging, unkind, wary, greedy, obnoxious, irritable, discreet, frivolous, cowardly, rebellious, adventurous, enterprising, unscrupulous, poised, moody, unfriendly, optimistic, disorderly, peaceable, considerate, humorous, worrying, preoccupied, trusting, mischievous, robust, superstitious, noisy, tolerant, realistic, masculine, witty, informal, prejudiced, reckless, jolly, courageous, meek, stubborn, aloof, sentimental, complaining, unaffected, cooperative, unstable, feminine, timid, retiring, relaxed, imaginative, shrewd, conscientious, industrious, hasty, commonplace, lazy, gloomy, thoughtful, dignified, wholesome, affectionate, aggressive, awkward, energetic, tough, shy, queer, careless, restless, cautious, polished, tense, suspicious, dissatisfied, ingenious, fearful, daring, persistent, demanding, impatient, contented, selfish, rude, spontaneous, conventional, cheerful, enthusiastic, modest, ambitious, alert, defensive, mature, coarse, charming, clever, shallow, deliberate, stern, emotional, rigid, mild, cruel, artistic, hurried, sympathetic, dull, civilized, loyal, withdrawn, confident, indifferent, conservative, foolish, moderate, handsome, helpful, gentle, dominant, hostile, generous, reliable, sincere, precise, calm, healthy, attractive, progressive, confused, rational, stable, bitter, sensitive, initiative, loud, thorough, logical, intelligent, steady, formal, complicated, cool, curious, reserved, silent, honest, quick, friendly, efficient, pleasant, severe, peculiar, quiet, weak, anxious, nervous, warm, slow, dependent, wise, organized, affected, reasonable, capable, active, independent, patient, practical, serious, understanding, cold, responsible, simple, original, strong, determined, natural, kind.

A.3 Queries (Caliskan et al., 2017)

Career executive, management, professional, corporation, salary, office, business, career.

Family home, parents, children, family, cousins, marriage, wedding, relatives.

Arts poetry, art, dance, literature, novel, symphony, drama, sculpture.

Arts_2 poetry, art, shakespeare, dance, literature, novel, symphony, drama.

Math math, algebra, geometry, calculus, equations, computation, numbers, addition.

Science science, technology, physics, chemistry, einstein, nasa, experiment, astronomy.

B Environmental Setup

Running Environments All experiments except for multilevel modelling were performed on Intel Xeon E5-2699 CPUs with Python 3.7. Running these experiments took approximately 700 CPU hours. Training the word embedding models with different random seeds took up most of the time (500 CPU hours). Moreover, calculating NBM bias scores for the full vocabularies was computationally expensive (more than 100 CPU hours). However, fortunately, the required consumption has a linear relationship with the number of target words. Thus, considerably fewer computational resources will be needed to estimate the reliability of a small set of target words.

Multilevel modelling was conducted in the statistical software R (version: 4.0.1) on an Intel Core i7-8565U CPU, consuming approximately 3 CPU hours.

Data Preprocessing For Reddit data (i.e. *r/AskScience* and *r/AskHistorians*), we lowercased, removed redundant spaces/urls, and used the Spacy[11] library to tokenize each sentence. For training GloVe embeddings, we substituted "<unk>" symbols in WikiText-103 with "<raw_unk>" symbols.

C Choosing ICC Estimators

Despite ICC being a commonly used tool for estimating test-retest and inter-rater reliability, there exist distinct forms of ICC estimators. Different forms of ICC can involve distinct assumptions and can therefore lead to very different interpretations (Koo and Li, 2016).

In this work, we follow the framework proposed by Shrout and Fleiss (1979). They define 6 forms

[11] `https://github.com/explosion/spaCy`

of ICC and present them as “ICC” with 2 additional numbers in parentheses (e.g., ICC(2,1) and ICC(3,1)). The first number refers to the **model** and can take on three possible values (1, 2 or 3): 1 is a one-way random-effects model, where each subject receives a unique, random set of raters; 2 is a two-way random-effects model, where all subjects receive the same randomly chosen set of raters and the reliability results are assumed to be generalisable to unseen raters; 3 is a two-way mixed-effects model, where the selected raters are the only raters of interest and thus the reliability results are not generalisable to other raters. The second number refers to the **intended use** of raters/measurements in an application and can take on two values (1 or k). 1 refers to having only a single rater or measurement; k means using the mean of k raters or measurements.

Therefore, depending on the specific research data and goals, one of the 6 ICC forms may be used. For test-retest reliability, we use ICC(2,1) for the following three reasons. First, the raters (i.e. different random seeds) are a random sample of the population (of all possible random seeds). Second, each bias score receives the same raters (i.e. random seeds). Third, in actual research practices, researchers would normally use only one rater (one random seed) to measure word embedding biases.

For inter-rater consistency, we use ICC(3,1) based on two considerations. First, we are only interested in comparing three specific scoring rules (i.e. raters). Second, in practice, researchers would use only the result from one scoring rule (i.e. rater) to measure word embedding biases.

D Additional Figures

D.1 Test-retest Reliability

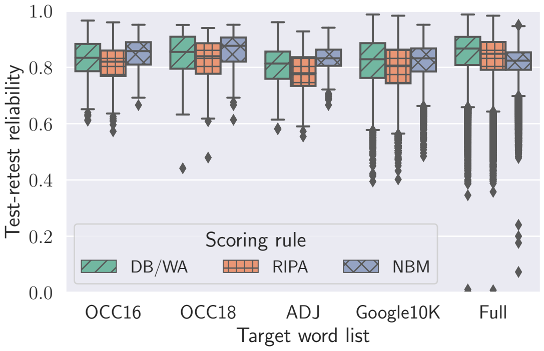


Figure 6: Test-retest reliability of target words. The word embeddings are trained with GloVe on WikiText-103.

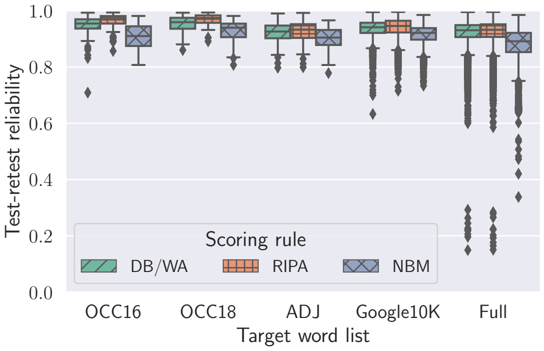


Figure 7: Test-retest reliability of target words. The word embeddings are trained with SGNS on r/AskScience.

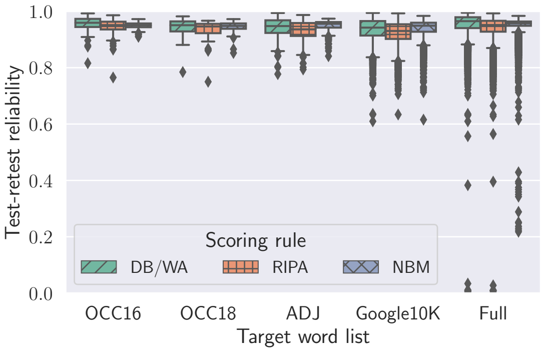


Figure 8: Test-retest reliability of target words. The word embeddings are trained with GloVe on r/AskScience.

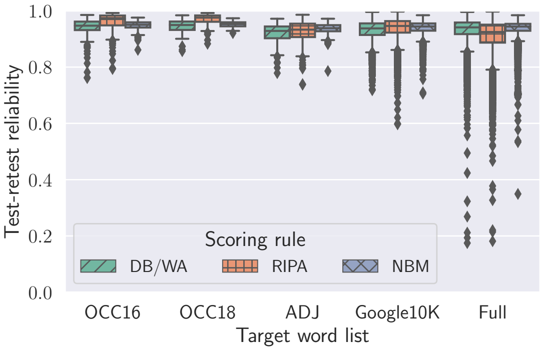


Figure 9: Test-retest reliability of target words. The word embeddings are trained with SGNS on r/AskHistorians.

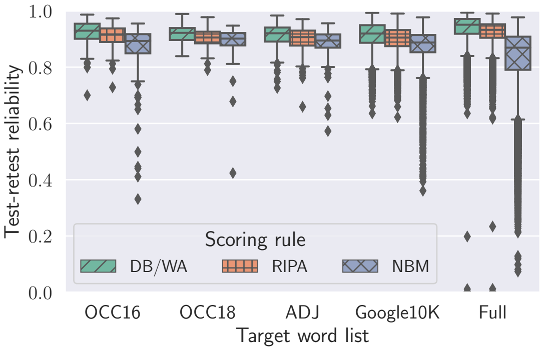


Figure 10: Test-retest reliability of target words, The word embeddings are trained with GloVe on r/AskHistorians.

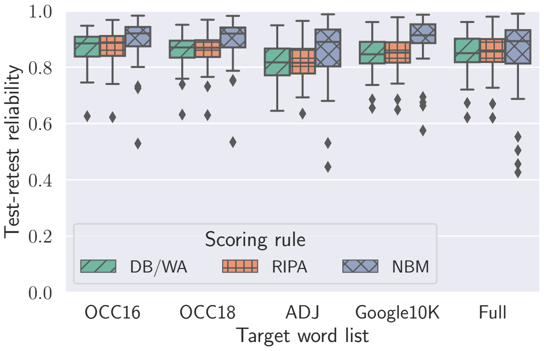


Figure 11: Test-retest reliability of gender base pairs. The word embeddings are trained with GloVe on WikiText-103.

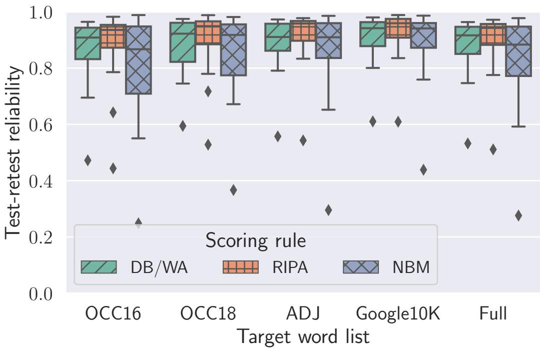


Figure 12: Test-retest reliability of gender base pairs. The word embeddings are trained with SGNS on r/AskScience.

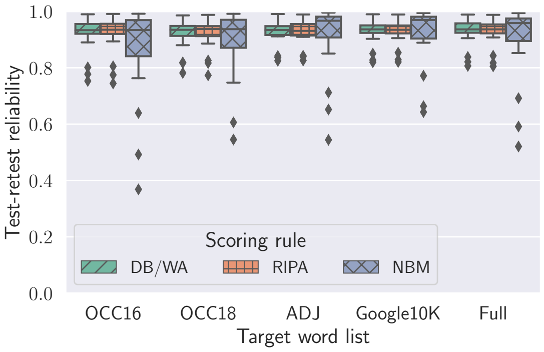


Figure 13: Test-retest reliability of gender base pairs. The word embeddings are trained with GloVe on r/AskScience.

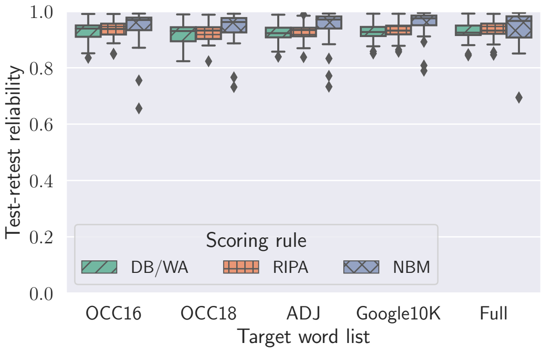


Figure 15: Test-retest reliability of gender base pairs. The word embeddings are trained with GloVe on r/AskHistorians.

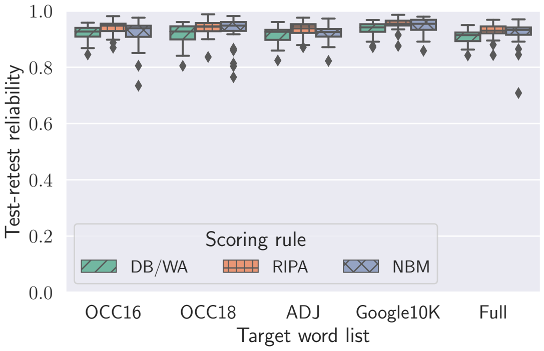


Figure 14: Test-retest reliability of gender base pairs. The word embeddings are trained with SGNS on r/AskHistorians.

D.2 Inter-rater Consistency

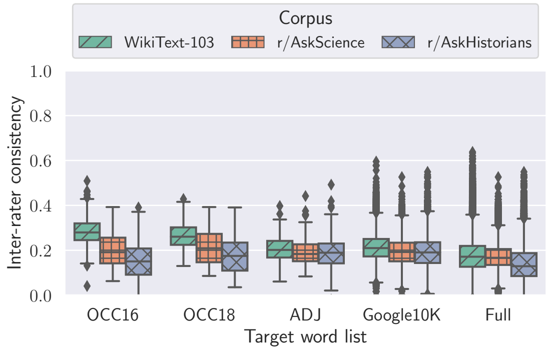


Figure 16: Inter-rater consistency of target words. The word embeddings are trained with SGNS.

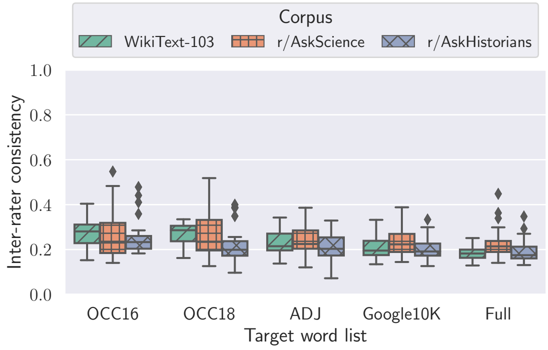


Figure 17: Inter-rater consistency of gender base pairs. The word embeddings are trained with SGNS.

E Effect of Hyper-parameters

In our study, we did not fine-tune the hyper-parameters while training the word embeddings, because the main goal of this paper is not to compare different setups of word embedding algorithms. In this section, we explore whether the results are sensitive to the choice of hyper-parameters used for training the word embeddings. In the main paper, for SGNS, we use 300 dimensions and 5 iterations. Here, we experiment with two different hyper-parameters on *r/AskHistorians*, 1) 3 iterations, and 2) 100 dimensions, respectively.

The results are shown in Figures 18 to 24. Comparing with the default hyper-parameters, we observe that although the specific values of different types of reliability change, the overall trends remain the same.

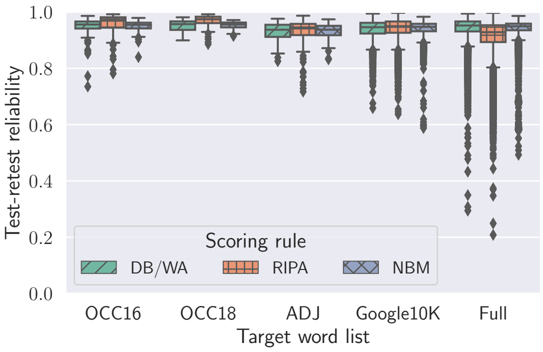


Figure 18: Test-retest reliability of target words of 100-dimensional SGNS word embeddings trained on *r/AskHistorians*.

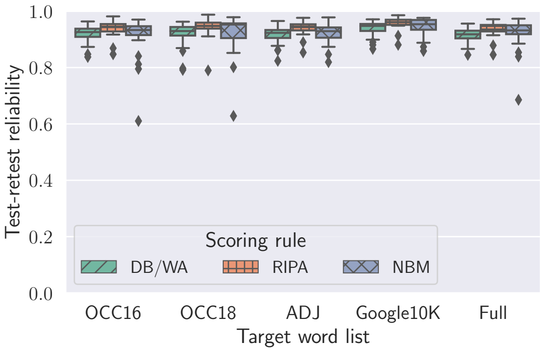


Figure 19: Test-retest reliability of gender base pairs of 100-dimensional SGNS word embeddings trained on *r/AskHistorians*.

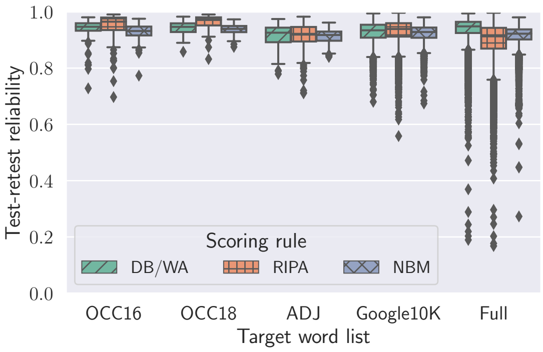


Figure 20: Test-retest reliability of target words of SGNS word embeddings trained with three iterations on *r/AskHistorians*.

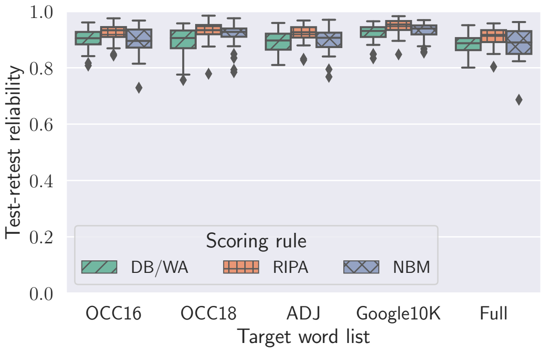


Figure 21: Test-retest reliability of gender base pairs of SGNS word embeddings trained with three iterations on *r/AskHistorians*.

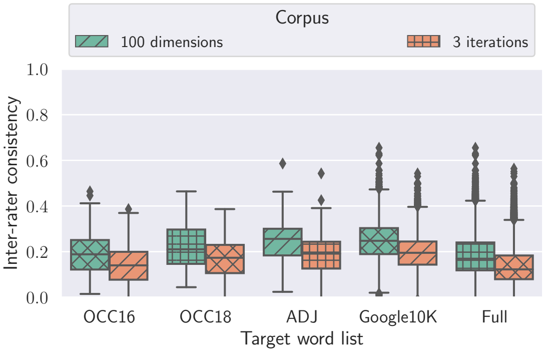


Figure 22: Inter-rater consistency of target words of SGNS word embeddings trained with three iterations or 100 dimensions on *r/AskHistorians*.

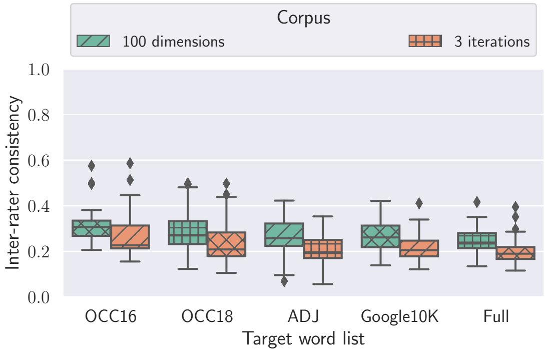


Figure 23: Inter-rater consistency of gender base pairs of SGNS word embeddings trained with three iterations or 100 dimensions on *r/AskHistorians*.

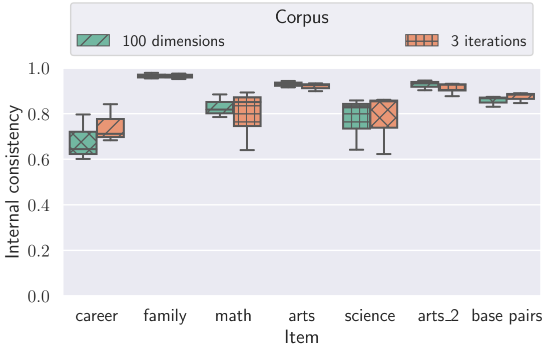


Figure 24: Internal consistency of SGNS word embeddings trained with three iterations or 100 dimensions on *r/AskHistorians*.

F Multilevel Regression

This section is a detailed description of the multilevel regression experiments in §5.5.

Multilevel Models In a similar research setup, Wendlandt et al. (2018) used variants of OLS regression to study influencing factors of word embedding stability. However, OLS regression has a strong assumption that the observations are unconditionally independent of one another. If this assumption is violated, standard errors of regression coefficients will be underestimated, leading to an overstatement of statistical significance.

In our study, our observations (reliability scores of target words) are nested within different corpora and word embedding algorithms. As a result, the residual variance is naturally partitioned into a within-corpus-algorithm component and a between-corpus-algorithm component. If the between-component is not explicitly accounted for, the observations will be correlated (thus not independent anymore) within a corpus or algorithm.

We therefore use **multilevel** regression instead of OLS regression (or its variants). Multilevel regression (or mixed-effect models, hierarchical models, among many other names) accounts for the grouping/hierarchical structure of data by explicitly modelling residuals at different levels of the data. In this way, the model not only relaxes the assumption of unconditional independent data, but also estimates group effects. The latter is especially beneficial because it allows us to quantify how much variance in the data is explained by the corpora and the algorithms, in addition to word-level features.

Formally, our multilevel regression model is:

$$\begin{aligned} y_{i(a,c)} &= X_{i(a,c)}\beta + \nu_{0a} + \mu_{0c} + \epsilon_{i(a,c)} \\ \nu_{0a} &\sim N\left(0, \sigma^2_{\nu_0}\right) \\ \mu_{0a} &\sim N\left(0, \sigma^2_{\mu_0}\right) \\ \epsilon_{0a} &\sim N\left(0, \sigma^2_{\epsilon_0}\right) \end{aligned}$$

where $y_{i(a,c)}$ refers to each reliability score of a target word i nested within an algorithm a and corpus c. $X_{i(a,c)}$ is a row vector containing observations on the word-level explanatory factors with a leading element of one. β is a column vector of parameter estimates of those factors. ν_{0a}, μ_{0a}, ϵ_{0a} are the residual error terms for algorithm a, corpus c and target word i. They are assumed to be normally distributed around zero with their own variances.

Note that we can not conduct such regression analyses for reliability scores on the level of queries, gender base pairs or scoring rules due to limited sample sizes (i.e. only 6 queries, 23 gender base pairs and 3 scoring rules).

Feature Selection We collect a wide range of features for the regression model. They can be divided into two categories: word-level features and group-level features.

We use several word-level features. First, we consider the natural logarithm of the number of synsets in WordNet (log #senses). We regard it as an approximation of the number of senses of words (Wendlandt et al., 2018). Then, we use the most common PoS tags of words (PoS). We use it to represent the syntactic roles of words in sentences (Wendlandt et al., 2018; Pierrejean and Tanguy, 2018). We call these two features word-intrinsic features because they are unrelated to the training corpora or algorithms.

We also include the natural logarithm of the frequencies of words (log freq). Frequency is found to influence the stability of word embeddings by Hellrich and Hahn (2016). Then, we consider its squared value as well (log^2 freq), because the relationship between word frequency and its corresponding embedding stability appears to be quadratic in our data. We refer to both words' frequencies and their squares as corpus-related features since they are only related to the training corpora.

Then, we use several embedding-related features. First, we consider a word's cosine similarity with its nearest neighbour (NN Sim, Pierrejean and Tanguy 2018). Intuitively, stable embeddings should have stable nearest neighbours. We also use the L2 norm of word embeddings (L2 Norm, Pierrejean and Tanguy 2018). Third, we use embedding stability (ES) as a predictor too. Previous studies (Wendlandt et al., 2018; Hellrich and Hahn, 2016; Antoniak and Mimno, 2018) usually measure the stability of word embeddings by the changes of their nearest neighbours. However, this method is insensitive to minor changes in the embedding space. Instead, we adopt a different method that detects the changes of lexical semantics (Hamilton et al., 2016; Tan et al., 2015; Kulkarni et al., 2015), to measure embedding stability in this paper. Specifically, given two word embedding models

	Test-retest		Inter-rater	
	Estimate	ΔR^2	Estimate	ΔR^2
scoring rule$_{\text{DB/WA}}$	reference	0.0040	-	-
scoring rule$_{\text{RIPA}}$	**-0.0102**		-	-
scoring rule$_{\text{NBM}}$	-0.0001		-	-
log freq	**0.0062**	0.0275	**0.0241**	0
$\log^2$ freq	**0.0023**	0.0051	**0.0008**	0
log #senses	**-0.0012**	0.0002	**-0.0007**	0
PoS$_{\text{adj}}$	reference	0.0093	reference	0.0022
PoS$_{\text{adp}}$	**-0.0096**		**-0.0195**	
PoS$_{\text{adv}}$	-0.0004		**-0.0118**	
PoS$_{\text{conj}}$	-0.0010		**-0.0446**	
PoS$_{\text{det}}$	**-0.0212**		**-0.0246**	
PoS$_{\text{noun}}$	**0.0085**		**0.0057**	
PoS$_{\text{num}}$	**-0.0025**		**-0.0886**	
PoS$_{\text{pron}}$	-0.0039		-0.0012	
PoS$_{\text{prt}}$	0.0037		**-0.0125**	
PoS$_{\text{verb}}$	0.0001		**-0.0056**	
PoS$_{\text{x}}$	**-0.0031**		**-0.0205**	
NN Sim	**-0.0024**	0.0120	**0.0011**	0.0038
L2 norm	**-0.0118**	0.0456	**-0.0051**	0
ES	**0.0457**	0.1020	**0.0186**	0
R^2_{fixed}	0.3261	-	0.0581	-
R^2_{corpus}	0.0113	-	0.0271	-
$R^2_{\text{algorithm}}$	0.1802	-	0.4307	-
R^2_{total}	0.5177	-	0.5160	-

Table 3: Results of multilevel regression on the test-retest and inter-rater reliability of target words. Estimates are standardized (bold if $p < 0.05$). ΔR^2 is reduction in explained variance when the corresponding factor is left out. R^2_{fixed}, R^2_{corpus}, $R^2_{\text{algorithm}}$ and R^2_{total} refer to the explained variance of fixed factors (i.e. word level features and scoring rules), embedding training corpus used, embedding training algorithm used, total effects of all these three parts, respectively.

W_1, W_2, we fit a transformation matrix Q , where

$$Q = \underset{Q^\top Q = I}{\arg\min} \|W_1 - QW_2\|_F.$$

Then the stability of word w is given by the cosine similarity of (w_1, Qw_2), where w_1 and w_2 are the corresponding word vectors of w in word embeddings W_1 and W_2.

Furthermore, for the regression model on test-retest reliability, we also include scoring rules as a fixed factor. Scoring rules can also be considered as a grouping factor, as the bias scores are also clustered within scoring rules. However, this approach would not yield specific effect estimates for each of the three scoring rules, but instead only a variance estimate for all the scoring rules as a whole. Because we are interested in comparing the effects of the three scoring rules on target words' test-retest reliability scores, we explicitly model scoring rules as a fixed factor.

Lastly, we include word embedding training corpora and word embedding training algorithms as the two group-level factors.

Results The full results are in Table 3. The interpretation of the table is the same as in §5.5. The main difference between this table and Table 2 is that the parameter estimates of the PoS feature are no longer omitted. We can see that this feature as a whole explains less than 1% of the variation in both the test-retest reliability and the inter-rater consistency of target words. This suggests that in the presence of other features, PoS is not a very important factor. Nevertheless, we do see some statistically significant and moderately sized parameter estimates of PoS categories. For instance, compared to adjectives, determiners tend to score on average 0.0212 lower on test-retest reliability. Numerals also score on average -0.0886 lower on inter-rater consistency than do adjectives.