PPOL 6801
Replication Report

Zhiyang Cheng (zc385) & **Muhammad Saad** (ms4689)

Introduction

We replicated the paper '*Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*', published by Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou in 2018. Although the authors run various and supplementary analyses, we replicate six outputs on embedding bias and historic trends on gender and ethnic minorities. This report includes a study and data overview, the replicated results, replication autopsy, and suggested extensions.

All code is publicly available on the authors' repository (Garg, 2018), including links to used data.

Study Overview

Garg et al. (2018) key argument is that word embeddings trained on publicly available large-scale text from the 20th and 21st centuries can be used to track socioeconomic trends for gender and ethnic minorities in the United States of America (USA). The authors compare changes in embedding bias for gender and ethnic minorities, and compare them to changes in occupation rates and historical surveys on social stereotypes. For the embedding bias, they compare how words representative of gender and ethnic minorities ("woman," "man," "Asian," "Black") and words representative of reference categories ("man", "white") appear in proximity of words that represent occupations or stereotypes-related adjectives.

Key Results

The authors show that changes in embedding bias correlate strongly with actual occupational distributions and social attitudes towards gender and ethnic minorities. The study has six main outputs, including time-series and cross-sectional graphs for occupation and embedding bias, correlation heatmaps for adjectives related embedding bias, and validation analysis. We tried to replicate these figures using the authors' code. We also do an extension of the validation exercise based on another related-database.

Data Sources

The study uses three types of data: embeddings trained on text-corpora, real-world demographic data for validation, and word lists. All data was available, except for the New York Times Annotated Database (completely unavailable) and the Word2Vec embeddings (hosted at a different platform). The authors also use Wikipedia and Common Crawl text for supplementary validation checks, which we do not replicate.

Table 1: Data Sources Overview: Study & Replication

Data Source	Study	Replication
Embeddings Data		
Corpus of Historical American English (COHA)	Google Books/Corpus of Historical American English (COHA) embeddings which are trained using SGNS. (eng-all embeddings using Histwords). These are a set of nine embeddings (Hamilton et al., 2016), each trained on a decade in the 1900s.	Same: SGNS-trained COHA embeddings (eng-all embeddings using HistWords)
Google News Word2Vec	Word2Vec vectors trained on the Google News dataset. The authors shared a Google Code Archive link (Mikolov et al., 2013) for access, which is no longer available.	Similar: Word2Vec embeddings but on a part of the Google News dataset. Available on HuggingFace (HuggingFace, 2023).
New York Times Annotated Corpus	New York Times articles to train embeddings related to Islam and Christianity (Sandhaus, 2008). Data is no longer available.	Different: Cornell Newsroom data, available from 1998 till 2017 (Grusky et al., 2018). Contains news articles and their summaries for numerous American news outlets.
Historical Trends Data		
U.S. Census occupational gender and ethnic shares	Used the Census occupational gender and ethnic rates over time to compare with embedding biases. Available through Integrated Public Use Microdata Series (Ruggles et al., 2023) and on the authors' repository under the data folder as *occupation_percentages_gender_occ1950.csv*	Same: US Census Data as available on the authors' repository
Word Lists		
Collated by the	The word lists collated by the authors are both	Same: Word lists

authors	available in their repository, as well as in the paper Appendix.	compiled by authors

Replication Results

We give an overview of the study's main outputs and how they compared with our replication.

Table 2: Overview of Study and Replication: Key Results

Study Output	Study Data	Replication Data	Replication Output
Figure 1: Comparison of gender embeddings bias and gender occupation Rate (Overall snapshot) - Appendix A	Google News Word2Vec (Google Code Archive)	Google News Word2Vec (Huggingface)	*Partial* Replication
Figure 2: Comparison of gender average embedding bias and average occupation rate (Timeseries comparison 1910-90) - Appendix B	eng-all HistWords COHA	eng-all HistWords COHA	*Perfect* Replication
Figure 3: Comparison of ethnicity (Asian) average embedding bias and average occupation rate (Timeseries comparison 1910-90) - Appendix C	eng-all HistWords COHA	eng-all HistWords COHA	*Partial* Replication
Figure 4: Pearson correlation of embeddings bias for women related adjectives across decades (Timeseries comparison 1910-90) - Appendix D	eng-all HistWords COHA	eng-all HistWords COHA	*Perfect* Replication
Figure 5: Pearson correlation of embeddings bias for Asian related adjectives across decades (Timeseries comparison 1910-90) - Appendix E	eng-all HistWords COHA	eng-all HistWords COHA	*Perfect* Replication

Figure 7. Islam bias score over time for words related to terrorism - Appendix F	New York Times Annotated Corpus (1988 -2004)	Cornell Newsroom corpus (1998 -2012)	*Extension*

Differences Between Replication and Study

Figure 1: Partial Replication

The sign on the fitted line is the same, i.e., largely the proximity of words representative of women and of words representative of particular occupations remains the same for the study and the replication. However, the fitted line in the replication is slightly flatter.

We believe that the HuggingFace Word2Vec may not have included some of the older Google News data which were used to train the Word2Vec embeddings available to the authors on Code Archive.

Figure 3 Mismatch: Partial Replication

The line graph for Asian occupation rates difference is the same, but the Asian embedding bias scores before 1950 do not match. It is worth noting that among the word lists in the data folder, there is a file named 'occupations1950_professional.txt', which corresponds to 1950, and this may mean that there may be some discrepancy with the word lists.

The codebook in this case is also not documented clearly, including mapping of different embeddings' files. We suspect that the authors likely performed some preprocessing or filtering which we missed out on.

Figure 7: Replication Extension

For this figure, the authors used the NYT Annotated Corpus, which is no longer available. Therefore, we could not replicate Figure 7. We took this as an opportunity to add a robustness check for the embeddings using separate news articles data. We used the Cornell Newsroom data, which is publicly available and contains news articles and their summaries for numerous American news outlets. There was a timeline mismatch though. The NYT data was available from 1988 till 2004. The Newsroom one is from 2001 till 2011. We used the word lists for Islam and Christianity related words compiled by the authors. We also trained the embeddings using the GLoVe algorithm, similar to the study.

Although the magnitude of Average Islam Bias is not the same, the temporal trends match. We see a peak in 2001 for both NYT and Newsroom, which coincides with 9/11 and association of muslims with terrorism. We also see a peak around 2005 for Newsroom, which matches the heightened sentiment against muslims after the 2005 London bombings.

Replication Autopsy

Three things worked well for us. Firstly, COHA-based analyses were robust, because the COHA embeddings have not changed and most of the analyses using them replicated cleanly. Secondly, the census-based validations replicated perfectly as well. The correlation between embedding bias and real-world occupation data reproduced the original findings closely. Lastly, our modified repository is also functional with changes after some debugging in the notebooks.

We did encounter some challenges as well. Firstly, the NYT-dependent results cannot be replicated. Secondly, Asian bias results were not similar. Our suspicion is that "eng-all" was mistakenly used in the original or inconsistently documented. Without exact inputs, perfect replication was a slight challenge. Overall, word embeddings are also highly sensitive to changes in training corpus and vocabulary updates, which may have affected our replication results.

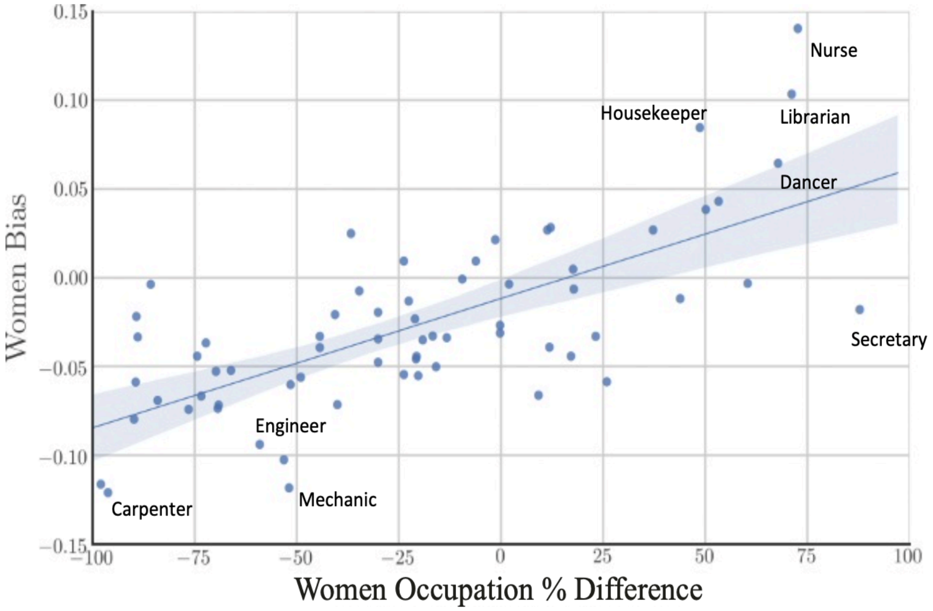
Replication Extension

We suggest extending the study in three key ways.

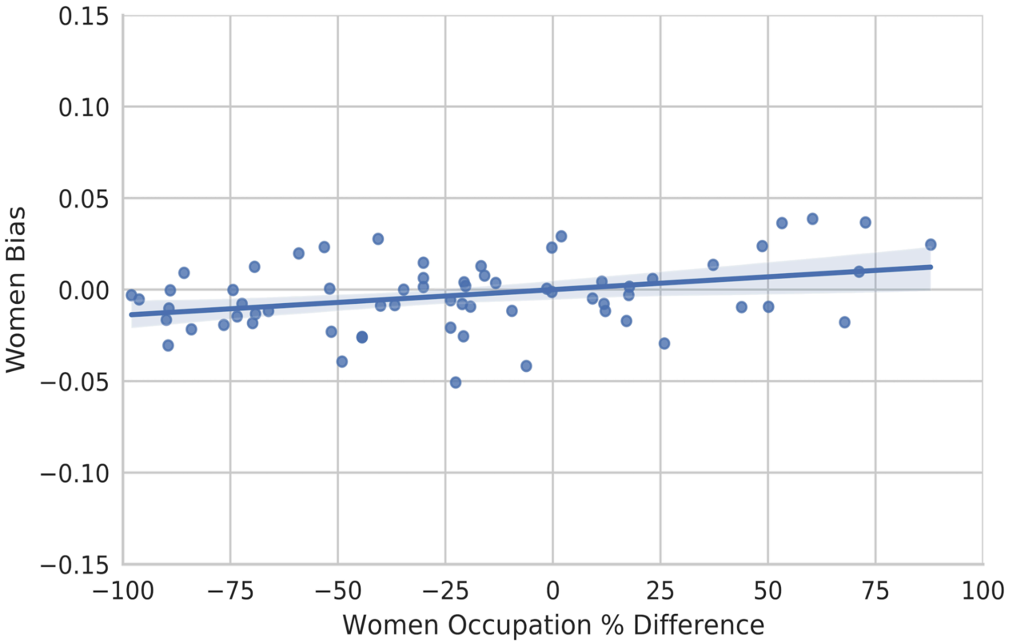
1. We would use contemporary open-source embeddings. Instead of Google News or NYT embeddings, we could use BERT-family contextual embeddings and Open-source news corpora (Newsroom, C4, RealNews), which would ensure long-term replicability. Another challenge in the study is the authors' use of compiled word lists. These word lists may not be completely exhaustive or relevant. The authors also do not expand much on how these word lists were compiled. Using BERT-related embeddings would remove the need for these word lists.
2. We would also add a human validation layer, where we could do stereotypes rating tasks on crowdworkers and adjective association surveys. This would allow that the word lists compiled are possibly more robust and exhaustive.
3. Lastly, the authors' use of word lists for names related to minorities has certain challenges. It assumes that the names across ethnicities are likely to be different. However, this may not be the case. For example, there may not be much difference between African-American and White names in the US. Therefore, the external validity of the study's methodology may be limited.

Appendices

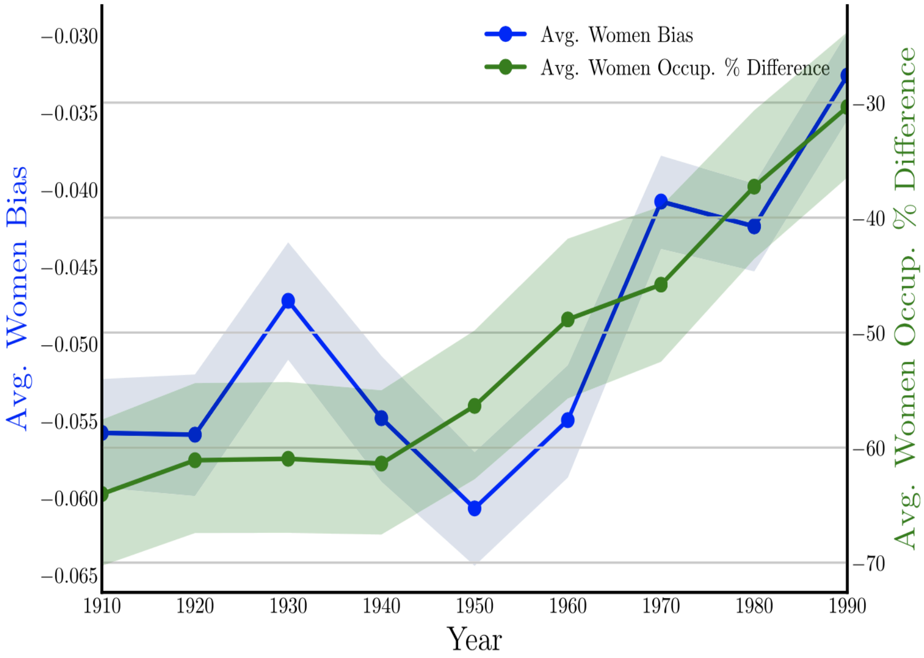
Appendix A: Figure 1 **Study**



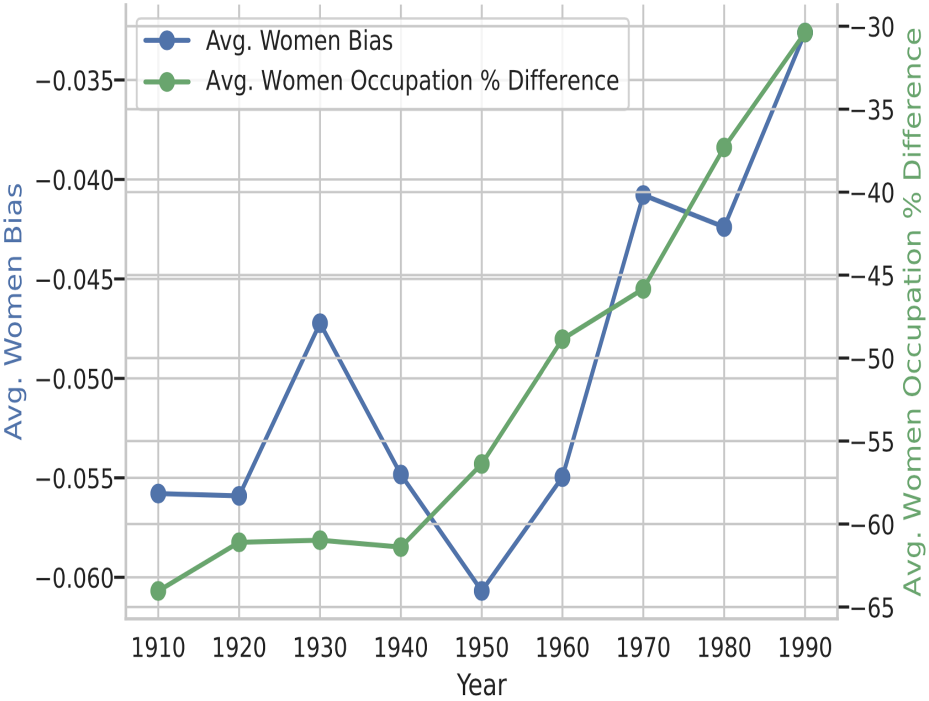
Appendix A: Figure 1 **Replication**



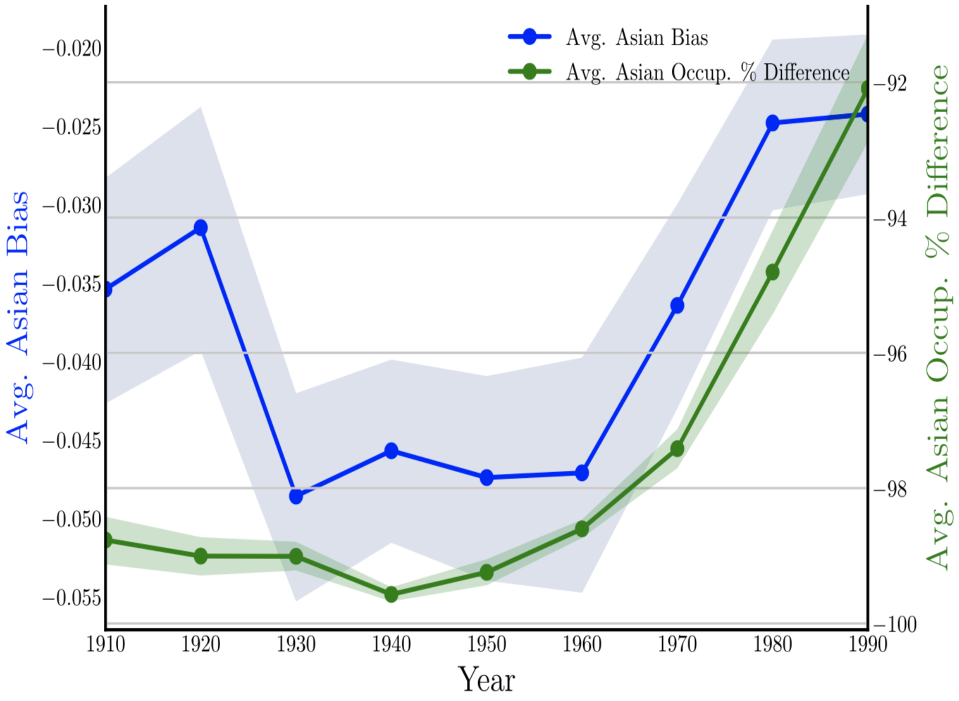
Appendix B: Figure 2 **Study**



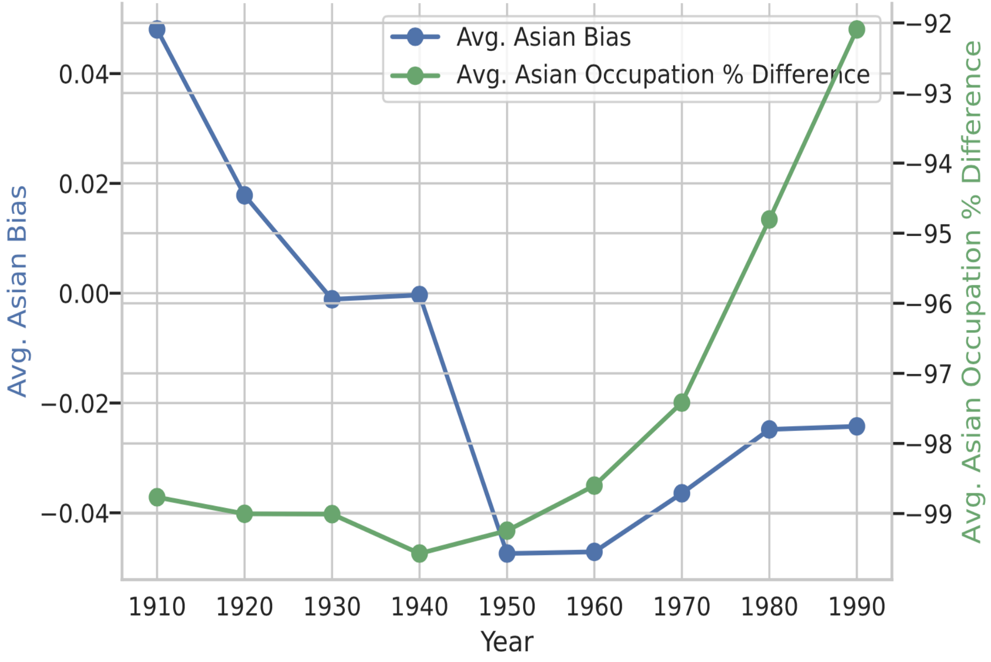
Appendix B: Figure 2 **Replication**



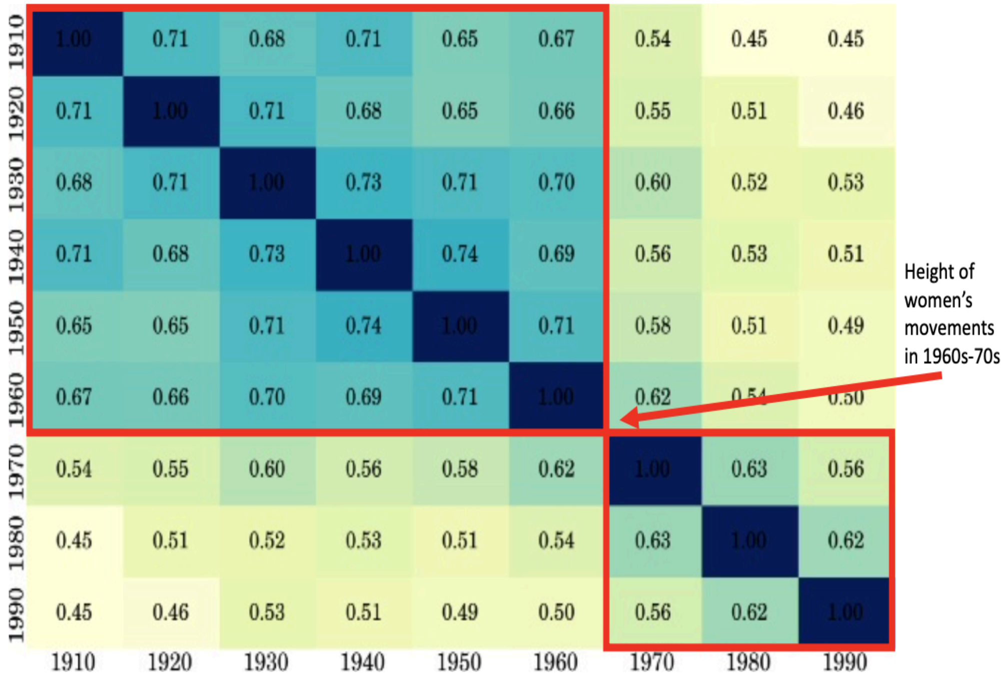
Appendix C: Figure 3 **Study**



Appendix C: Figure 3 **Replication**



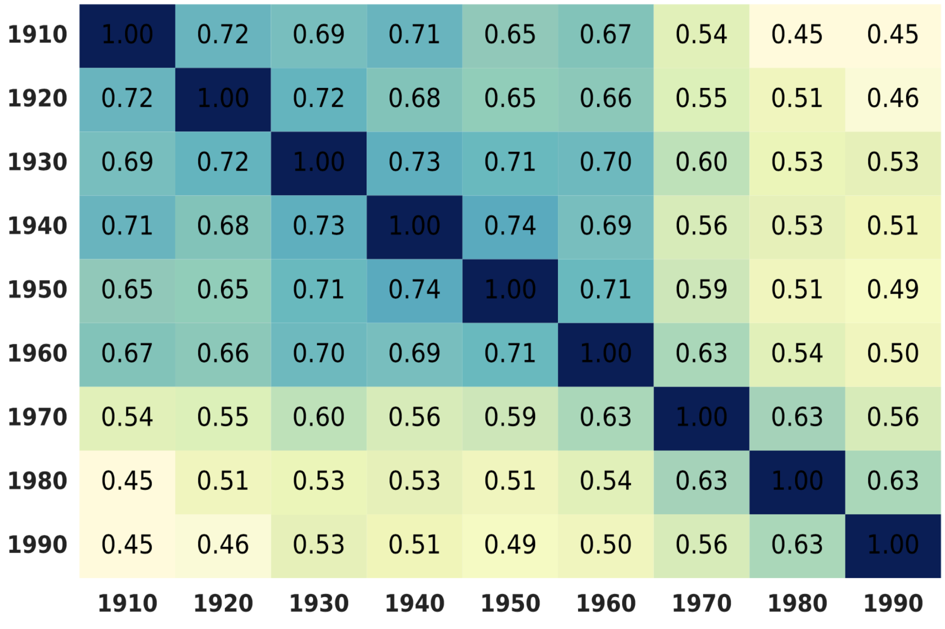
Appendix D: Figure 4 **Study**



	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.71	0.68	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.71	1.00	0.71	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.68	0.71	1.00	0.73	0.71	0.70	0.60	0.52	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.58	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.62	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.58	0.62	1.00	0.63	0.56
1980	0.45	0.51	0.52	0.53	0.51	0.54	0.63	1.00	0.62
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.62	1.00

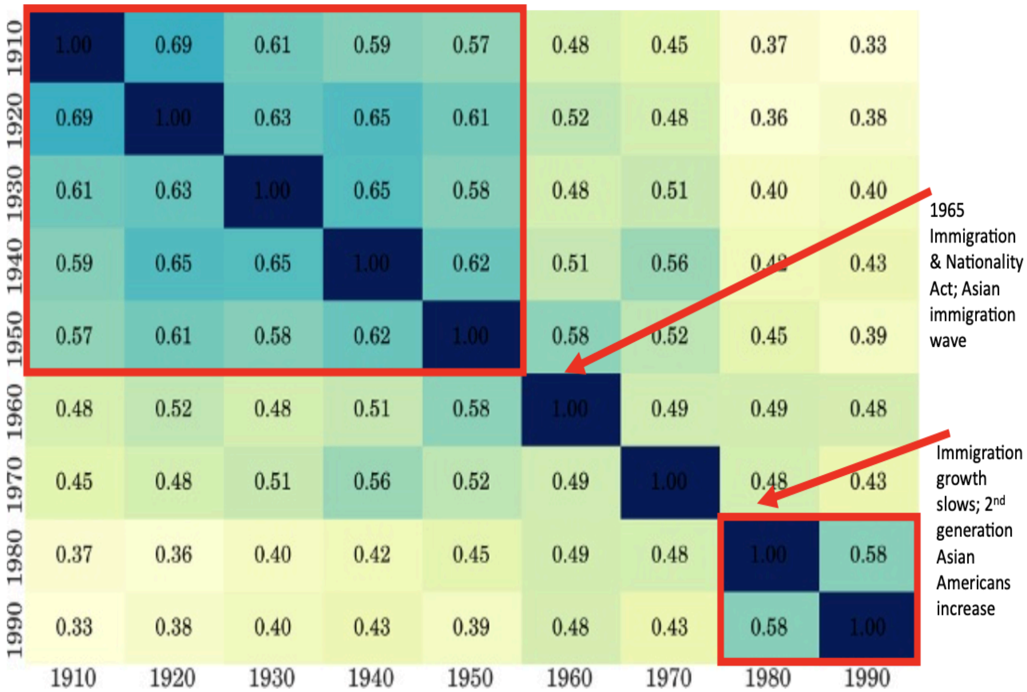
Height of women's movements in 1960s-70s

Appendix D: Figure 4 **Replication**



	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.72	0.69	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.72	1.00	0.72	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.69	0.72	1.00	0.73	0.71	0.70	0.60	0.53	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.59	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.63	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.59	0.63	1.00	0.63	0.56
1980	0.45	0.51	0.53	0.53	0.51	0.54	0.63	1.00	0.63
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.63	1.00

Appendix E: Figure 5 **Study**

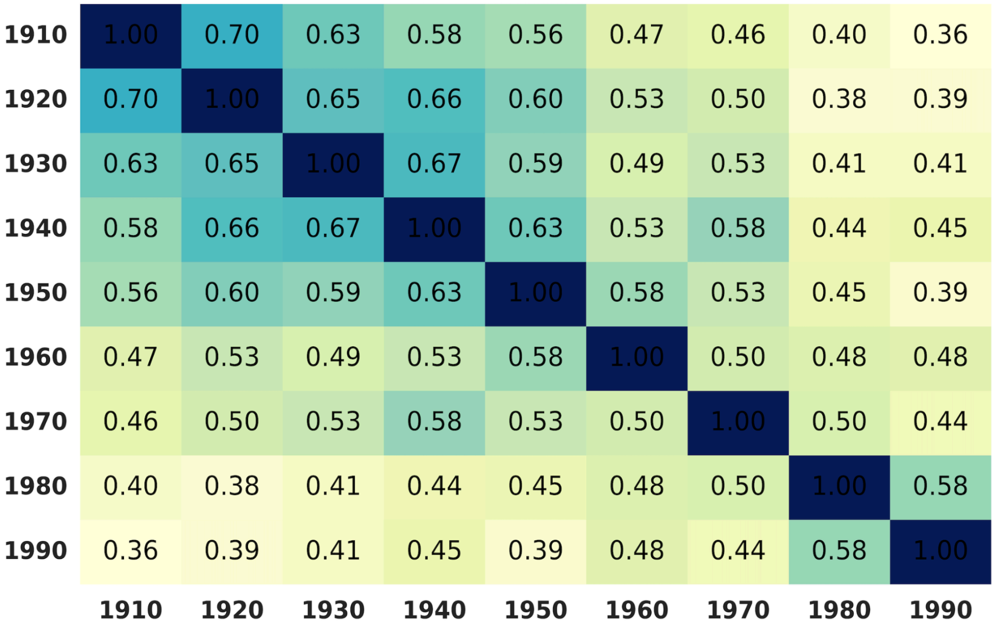


	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.69	0.61	0.59	0.57	0.48	0.45	0.37	0.33
1920	0.69	1.00	0.63	0.65	0.61	0.52	0.48	0.36	0.38
1930	0.61	0.63	1.00	0.65	0.58	0.48	0.51	0.40	0.40
1940	0.59	0.65	0.65	1.00	0.62	0.51	0.56	0.42	0.43
1950	0.57	0.61	0.58	0.62	1.00	0.58	0.52	0.45	0.39
1960	0.48	0.52	0.48	0.51	0.58	1.00	0.49	0.49	0.48
1970	0.45	0.48	0.51	0.56	0.52	0.49	1.00	0.48	0.43
1980	0.37	0.36	0.40	0.42	0.45	0.49	0.48	1.00	0.58
1990	0.33	0.38	0.40	0.43	0.39	0.48	0.43	0.58	1.00

1965 Immigration & Nationality Act; Asian immigration wave

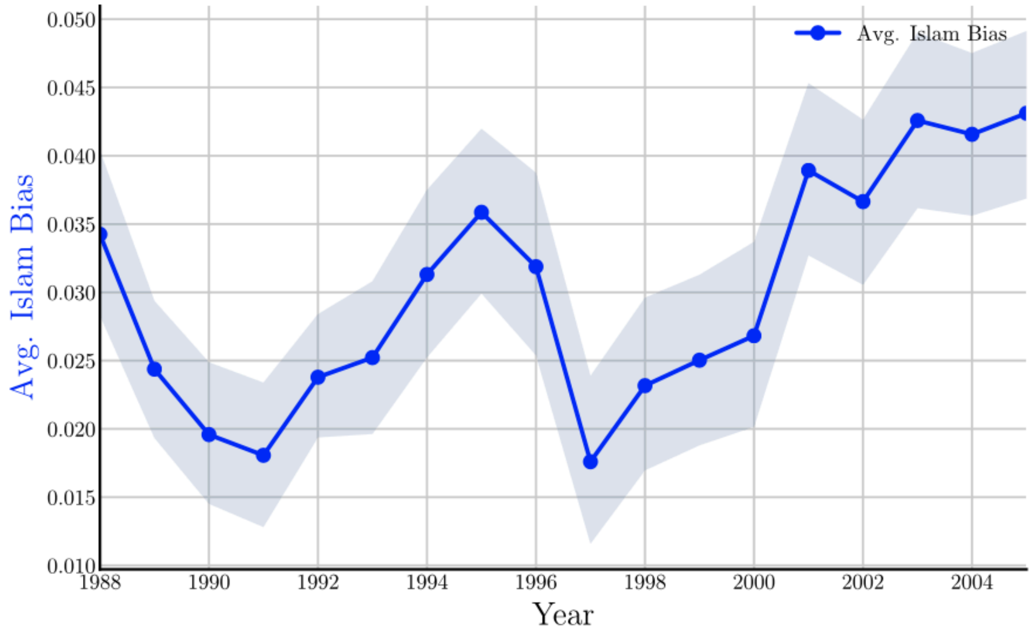
Immigration growth slows; 2nd generation Asian Americans increase

Appendix E: Figure 5 **Replication**

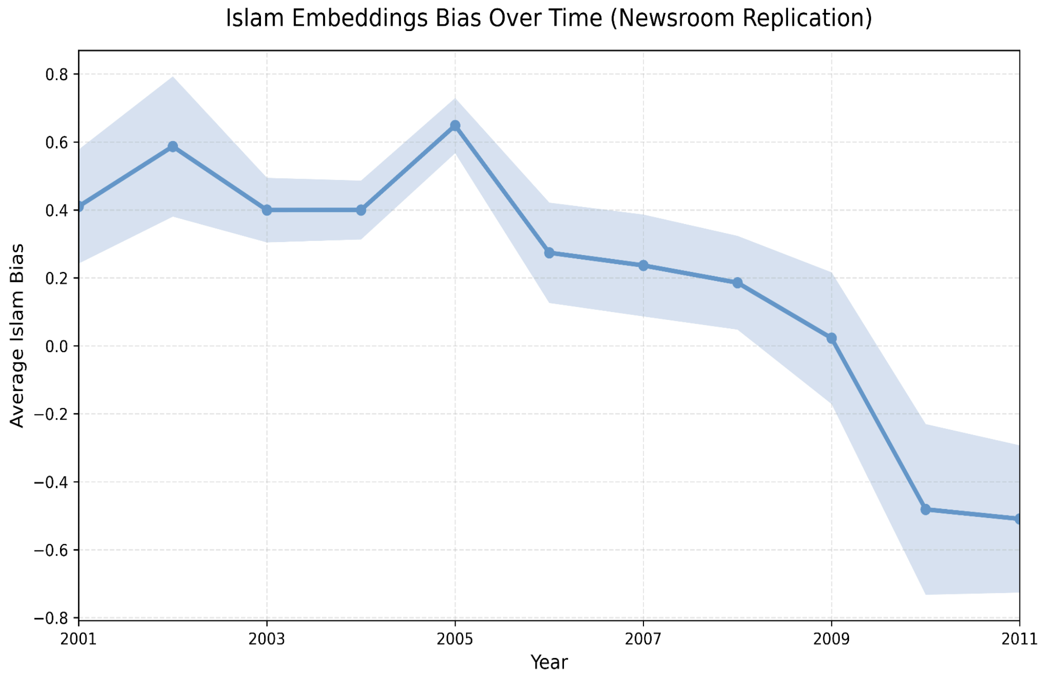


	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.70	0.63	0.58	0.56	0.47	0.46	0.40	0.36
1920	0.70	1.00	0.65	0.66	0.60	0.53	0.50	0.38	0.39
1930	0.63	0.65	1.00	0.67	0.59	0.49	0.53	0.41	0.41
1940	0.58	0.66	0.67	1.00	0.63	0.53	0.58	0.44	0.45
1950	0.56	0.60	0.59	0.63	1.00	0.58	0.53	0.45	0.39
1960	0.47	0.53	0.49	0.53	0.58	1.00	0.50	0.48	0.48
1970	0.46	0.50	0.53	0.58	0.53	0.50	1.00	0.50	0.44
1980	0.40	0.38	0.41	0.44	0.45	0.48	0.50	1.00	0.58
1990	0.36	0.39	0.41	0.45	0.39	0.48	0.44	0.58	1.00

Appendix F: Figure 7 **Study**



Appendix F: Figure 7 **Replication**



References

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

Garg, N. (2018). *EmbeddingDynamicStereotypes* [Computer software]. GitHub. https://github.com/nikhgarg/EmbeddingDynamicStereotypes

Grusky, M., Naaman, M., & Artzi, Y. (2018). *Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies* [Dataset]. Cornell NLP Group. https://lil.nlp.cornell.edu/newsroom/index.html

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *HistWords: Word embeddings for historical text* [Dataset]. Stanford NLP Group. https://nlp.stanford.edu/projects/histwords/

HuggingFace. (2023). *Word2vec-google-news-300* [Model]. HuggingFace. https://huggingface.co/fse/word2vec-google-news-300

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* [Software]. Google Code Archive. https://code.google.com/archive/p/word2vec/

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., & Sobek, M. (2023). *IPUMS USA: Version 6.0* [Dataset]. IPUMS. https://www.ipums.org/projects/ipums-usa/d010.v6.0

Sandhaus, E. (2008). *The New York Times Annotated Corpus (LDC2008T19)* [Dataset]. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2008T19