Supplementary Material

A Pilot study on the number of runs

The presented bias silhouettes for a given metric, word list and word embedding model are only approximations of the full silhouette, as the number of possible combinations is usually not feasible to evaluate. We thus conducted a pilot study to find out, how many shuffled word lists need to be evaluated to sufficiently approximate the full silhouette. To do so, we repeated the evaluation presented in the paper with different amounts of shuffled word lists. The full results can be seen in Table 1 to 3.

B Bias Silhouette plots

For reasons of brevity, we omitted some bias silhouette plots from the main paper. We therefore show a complete version of the plots per evaluated metric in Figure 1 to 3.

		GloVe						NBatch					
Metric	**Bias type**	**5 runs**	**10 runs**	**20 runs**	**40 runs**	**80 runs**	**100 runs**	**5 runs**	**10 runs**	**20 runs**	**40 runs**	**80 runs**	**100 runs**
ECT	**Gender**	0.058	0.083	0.104	0.114	0.126	0.128	0.041	0.059	0.067	0.087	0.094	0.096
	Ethnicity	0.011	0.013	0.015	0.017	0.018	0.018	0.012	0.021	0.023	0.026	0.028	0.030
	Religion	0.013	0.021	0.029	0.038	0.044	0.044	0.023	0.028	0.036	0.048	0.050	0.052
RNSB	**Gender**	0.104	0.149	0.163	0.203	0.214	0.216	0.002	0.003	0.003	0.004	0.005	0.005
	Ethnicity	0.208	0.269	0.344	0.364	0.401	0.411	0.001	0.002	0.003	0.004	0.004	0.004
	Religion	0.254	0.305	0.329	0.381	0.422	0.427	0.017	0.035	0.038	0.045	0.047	0.047
WEAT	**Gender**	0.032	0.036	0.039	0.041	0.053	0.054	0.055	0.082	0.085	0.093	0.116	0.117
	Ethnicity	0.009	0.013	0.018	0.021	0.022	0.022	0.047	0.058	0.070	0.076	0.079	0.080
	Religion	0.014	0.017	0.021	0.029	0.033	0.034	0.032	0.042	0.052	0.063	0.069	0.070

Table 1: Robustness scores for the pilot study on number of required runs. Each column represents the results for a specific number of shuffled word lists describing *bias-conveying concepts* that were combined with different metrics.

		GloVe						NBatch					
Metric	**Bias type**	**5 lists**	**10 lists**	**20 lists**	**40 lists**	**80 lists**	**100 lists**	**5 lists**	**10 lists**	**20 lists**	**40 lists**	**80 lists**	**100 lists**
ECT	**Gender**	0.037	0.050	0.057	0.066	0.069	0.069	0.087	0.105	0.118	0.122	0.128	0.137
	Ethnicity	0.170	0.178	0.207	0.230	0.254	0.257	0.085	0.108	0.116	0.118	0.123	0.130
	Religion	0.184	0.246	0.269	0.288	0.303	0.306	0.093	0.097	0.101	0.106	0.112	0.114
RNSB	**Gender**	0.068	0.090	0.109	0.141	0.145	0.149	0.013	0.020	0.031	0.038	0.043	0.044
	Ethnicity	0.123	0.183	0.231	0.252	0.274	0.274	0.004	0.005	0.009	0.011	0.013	0.013
	Religion	0.216	0.305	0.352	0.371	0.385	0.387	0.066	0.082	0.119	0.127	0.135	0.138
WEAT	**Gender**	0.034	0.037	0.042	0.045	0.050	0.050	0.094	0.155	0.185	0.228	0.246	0.256
	Ethnicity	0.093	0.109	0.148	0.195	0.212	0.213	0.095	0.131	0.179	0.214	0.253	0.256
	Religion	0.182	0.214	0.227	0.263	0.291	0.296	0.204	0.209	0.248	0.286	0.321	0.328

Table 2: Robustness scores for the pilot study on number of required runs. Each column represents the results for a specific number of shuffled word lists describing *social groups* that were combined with different metrics.

		Bias-conveying concepts						Social groups					
Metric	**Bias type**	**5 lists**	**10 lists**	**20 lists**	**40 lists**	**80 lists**	**100 lists**	**5 lists**	**10 lists**	**20 lists**	**40 lists**	**80 lists**	**100 lists**
ECT	**Gender**	0.063	0.060	0.051	0.049	0.048	0.048	0.009	0.010	0.016	0.015	0.013	0.013
	Ethnicity	0.071	0.072	0.072	0.073	0.073	0.073	0.052	0.065	0.077	0.077	0.074	0.075
	Religion	0.078	0.083	0.082	0.083	0.082	0.081	0.080	0.060	0.073	0.075	0.071	0.072
RNSB	**Gender**	0.452	0.451	0.440	0.437	0.438	0.438	0.581	0.581	0.585	0.586	0.585	0.586
	Ethnicity	0.208	0.197	0.208	0.211	0.215	0.218	0.155	0.160	0.160	0.162	0.165	0.162
	Religion	0.223	0.221	0.227	0.219	0.215	0.217	0.182	0.173	0.185	0.181	0.194	0.191
WEAT	**Gender**	0.221	0.217	0.211	0.214	0.213	0.213	0.186	0.205	0.201	0.203	0.196	0.197
	Ethnicity	0.218	0.221	0.220	0.219	0.218	0.219	0.192	0.189	0.201	0.200	0.203	0.207
	Religion	0.085	0.085	0.086	0.085	0.083	0.083	0.055	0.087	0.073	0.079	0.078	0.079

Table 3: Accuracy scores for the pilot study on number of required runs. Each column represents the results for a specific number of shuffled word lists describing either a *bias-conveying concepts* or *social group* that were combined with different metrics.

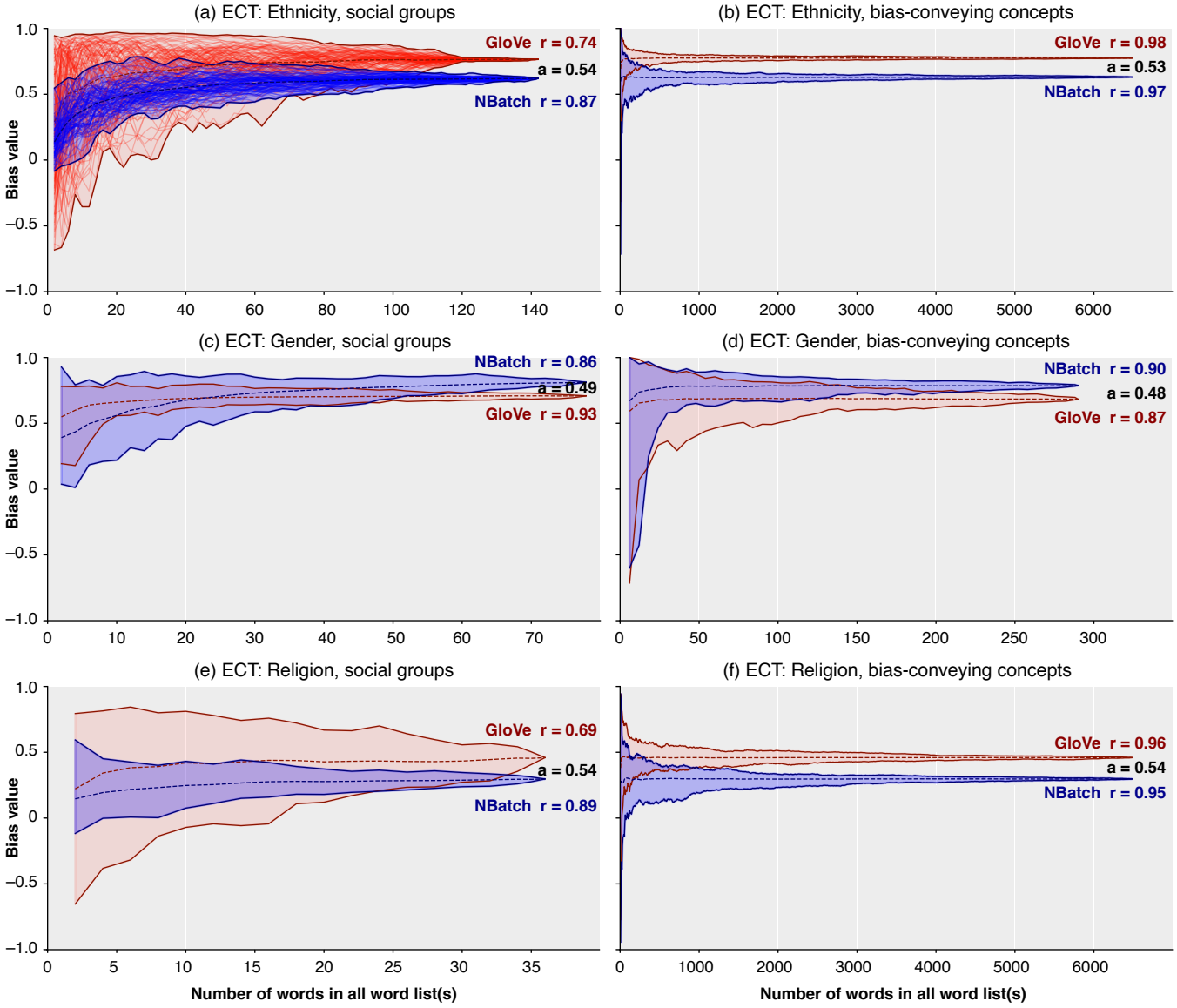

Figure 1: The bias silhouettes and the resulting robustness and accuracy scores on the two reference word embedding models for the social group word lists (left) and bias-conveying concepts (right) of the three considered types of social bias (top to bottom) under the ECT metric. Exemplarily, we show all $n = 100$ interpolated bias curves for the bias silhouettes of the ethnicity social group word lists at the top left.

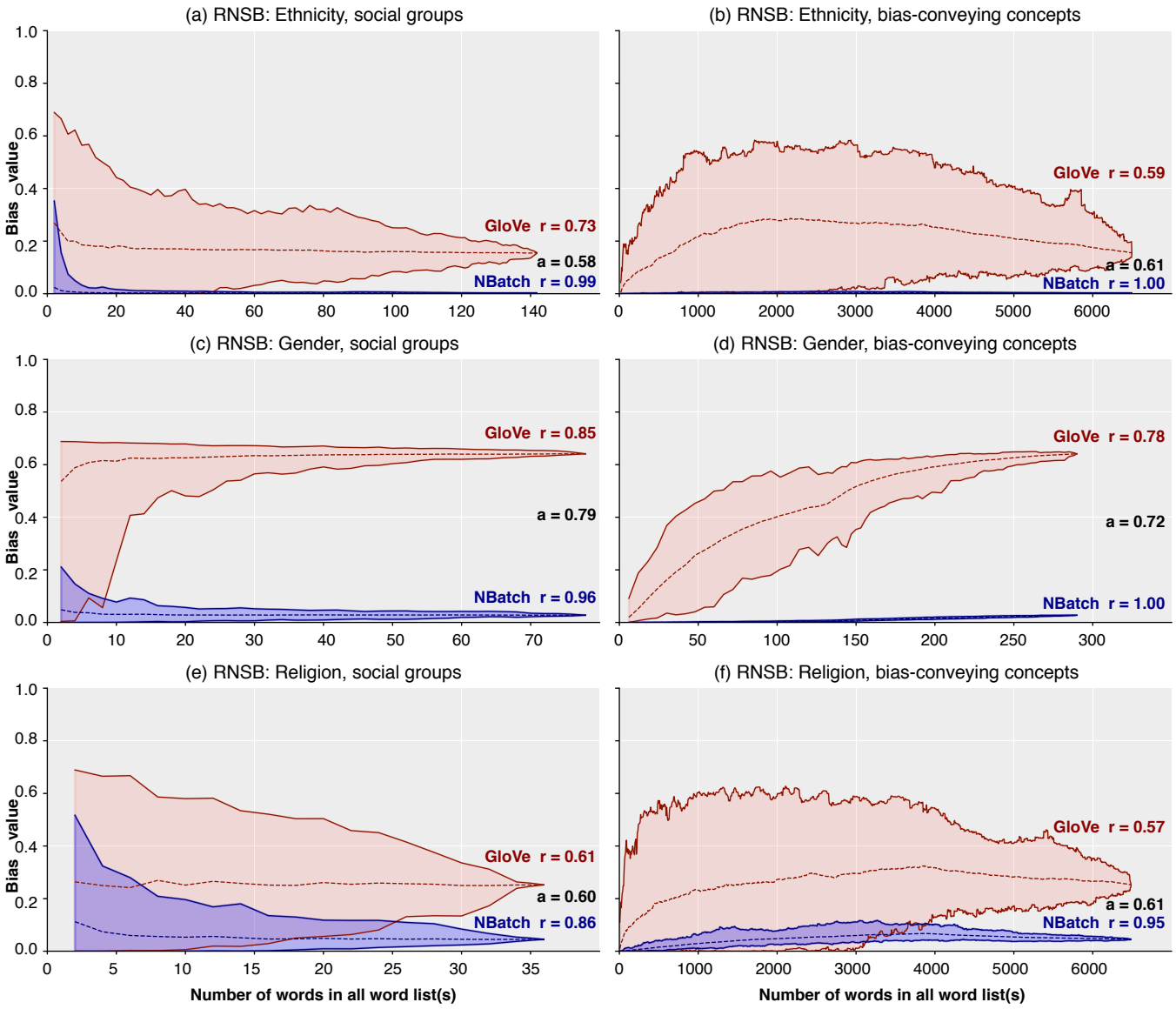

Figure 2: Bias silhouettes, robustness, and accuracy score analog to those in Figure 1 but here for the RNSB metric.

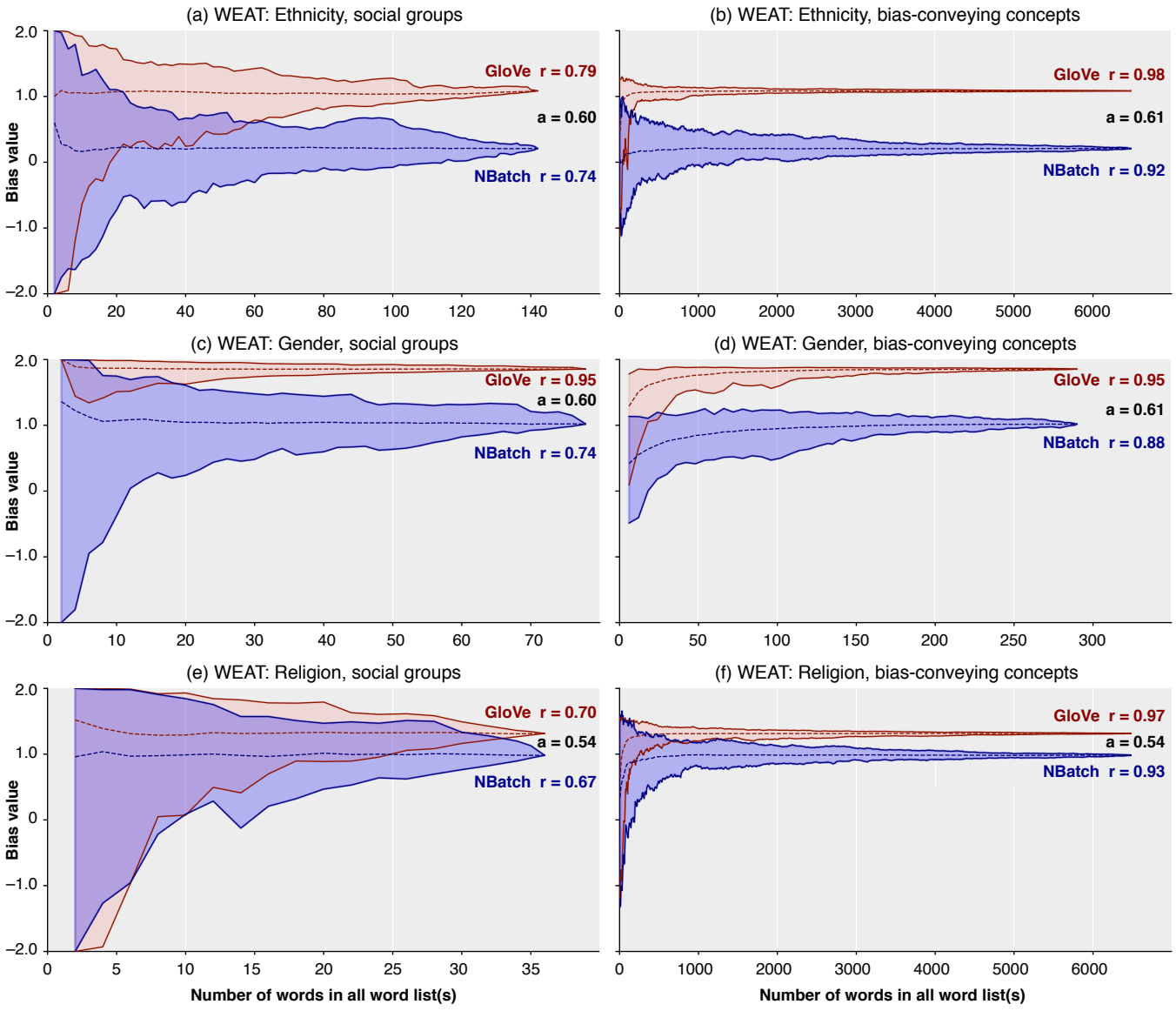

Figure 3: Bias silhouettes, robustness scores, and accuracy score analog to those in Figure 1 but here for the WEAT metric.