Exploring a Causal Link between Language and Cultural Biases

Molly Lewis
mollyllewis@gmail.com
Department of Psychology
University of Wisconsin-Madison

Gary Lupyan
lupyan@wisc.edu
Department of Psychology
University of Wisconsin-Madison

Abstract

The abstract.

Keywords: IAT, cultural biases, gender, linguistic relativity.

Introduction

Study 1: Cross-cultural gender bias in implicit behavior

We quantified the degree of gender bias in a culture using data from the Implicit Association Task ("IAT"; Greenwald, McGhee, & Schwartz, 1998). The IAT measures the strength of respondents' associations between two pairs of concepts (e.g., male-career/female-family vs. male-family/female-career). The underlying assumption of the measure is that concepts that are represented as more similar to each other in the cognitive system should be easier to pair together in a behavioral task, compared to two concepts that are relatively dissimilar. Concepts are paired in the task by assigning them to the same response keys in a 2AFC categorization task. In the critical blocks of the task, concepts are assigned to keys in a way that is either bias-congruent (i.e. Key A = male/career; Key B = female/family) or bias-incongruent (i.e. Key A = male/family; Key B = female/career). Participants are then presented with a word related to one of the four concepts and asked to classify it by responding with one of the two keys as quickly as possible. Slower reaction times in the bias-incongruent blocks relative to the bias-congruent blocks are interpretted as an implicit association between the correspondings concepts (i.e. a bias to associate male with career, and female with family).

Method

We analyzed an exisiting dataset of IAT scores collected online from a large, culturally diverse sample (Project Implicit: `https://implicit.harvard.edu/implicit/`; Nosek, Banaji, & Greenwald, 2002)[1]. Our analysis included all gender-career IAT scores collected from respondents between 2005 and 2016 who had complete data and were located in countries with more than 400 total respondents (N = 773,205). We further restricted our sample based on participants' reaction times and errors using the same criteria described in Nosek, Banjai, and Greenwald (2002, pg. 104). Our final sample included 664,359 participants from 49 countries, with a median of 1,123 participants per country.

Several measures have been used in the literature to describe the difference in reaction time between bias congruent and incongruent blocks. Here, we use the best perfomring measure, D-score, which quantifies the difference between critical blocks while also accounting for individual differences in response time (Greenwald, Nosek, & Banaji, 2003).

In addition to the implicit measure, we also analyzed an explicit measure of gender bias. After completing the IAT, participants were asked, "How strongly do you associate the following with males and females?" for both the words "career" and "family." Participants indicated their response on a Likert scale ranging from female (1) to male (7). We calculated an explicit gender bias score for each participant as the career response minus the family response, such that greater values indicated more gender bias (as for the D-score).

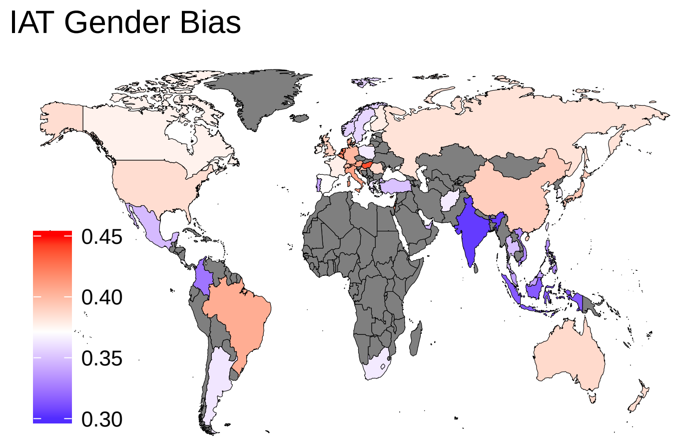

Figure 1: IAT gender bias (D-score) for the 49 countries with available data. All countries show a gender bias, with red indicating above average and blue indicating below average bias.

Results

Broadly, we replicate the patterns in the IAT literature (Nosek et al., 2002). First, participants in all countries showed a bias to associate men with career and females with family. Figure 1 shows the magnitude of the IAT gender bias (D-score) across all 49 countries (M = 0.37; SD = 0.03). Second, implicit and explicit bias measures were correlated both at the level of individual participants (r = 0.15; $p < .00001$) and at the level of countries (r = 0.31; p = 0.03).

Finally, previous work has shown a difference for women

Study 2: Cross-cultural gender bias in language

In Study 2, we ask whether participants' implicit and explicit gender biases are correlated with biases found in the seman-

[1]All analysis code can be found in an online repository: https://github.com/mllewis/IATLANG

tics of participants' native languages. To model semantics, we turn to a recently developed machine-learning method for deriving lexical semantics from text: auto-encoding neural network models. The underlying assumption of these models is that the meaning of a word can be described by the words it tends to co-occur with – an approach known as distributional semantics (Firth, 1957). Under this approach, a word like "dog" is represented more semantically similar to "hound" than "banana" because it co-occurs with words more in common with "hound" than "banana" in a large corpus of text.

Recent developments in machine learning allow the idea of distributional semantics to be implentented in a way that both takes into account many features of local language structure while remaining computationally tractable. The best known of these word embedding models is *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013). The model's output is a vector for each word reprseneting its semantics. A measure of the semantic similarity between two words can be derived by taking the distance between the word vectors (using cosine distance, for example). Similarity measures derived from these models have been shown to be highly correlated with human judgements of word similarity (e.g., Hill, Reichart, & Korhonen, 2015).

Recent work has used models like word2vec to measure the presence of social biases in the semantics of English in a way that is highly analogus to the behvioral IAT (Caliskan, Bryson, & Narayanan, 2017; henceforth *CBN*). This is done by measuring the distance in vector space between the same sets of words that are presented to participants in the IAT task. CBN demonstrate that these distance measures are highly correlated with reaction times in the behavioral IAT task, suggesting that the biases measured by the IAT are also found in the lexical semantics of natural language.

In Study 2, we use the method described by CBN to measure the biases in the semantics of the natural languages spoken in the countries of participants in Study 1. To do this, we take advantage of a set of models that have been pre-trained on a corpus of Wikipedia text in a large number of languages (Bojanowski, Grave, Joulin, & Mikolov, 2016). In Study 2a, we replicate the CBN findings with the Wikipedia corpus; In Study 2b, we show that the implicit and explicit gender biases reported in Study 1 for individual countries are correlated with the biases found in the semantics of the natural language spoken by those participants.

Study 2a: Replication of Caliskan, et al. (2017)

Method We use a word embedding model that has been pre-trained model on the corpus of English Wikipedia using the fastText algorithm (Bojanowski et al., 2016)[2]. The model contains 2,519,370 words with each word reprsented by a 300 dimension vector.

Using the Wikpedia fastText model, we calculate an effect size for each of the 10 biases reported in CBN, corresponding to behavioral IAT results existing in the literature: flowers-insects, instruments-weapons, race, gender-career gender-math, gender-science, mental-physical (originally labeled as WEAT 1-10). We calculate the bias using the same effect size metric described in CBN, a standardized difference score of the relative similarity of the target words to the target attributes (i.e. relative similarity of male to career vs. relative similarity of female to career). This measure is analagous the behavioral D-score measure in Study 1 and, like for D-score, larger values indicate a larger bias.

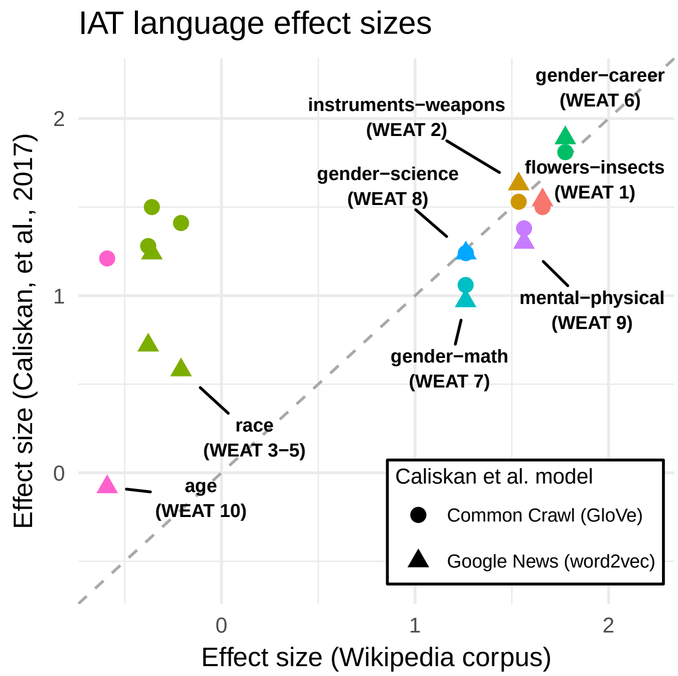

Figure 2: Effect sizes for the 10 IAT biases types (WEAT 1-10) reported in Caliskan et al. (2017; CBN). The effect sizes reported in CBN are plotted against effect sizes from the Wikipedia corpus. Point color corresponds to bias type, and point shape corresponds to the two CBN models trained on different corpora and with different algorithms.

Results Figure 2 shows the effect size measures derived from the Wikipedia corpus plotted against effect size esimtates reported by CBN from two different models (trained on the Common Crawl and Google News corpora). With the exception of biases related to race and age, effect sizes from the Wikipedia corpus are highly similar to those reported by CBN. In particular, for the gender-career IAT – the bias relevant to our current purposes – we estimate the effect size to be 1.78, while CBN estimates it to be 1.81 (Common Crawl) and 1.89 (Google News).

Study 2b: Predicting implicit bias with language IAT

With our corpus validated, we next turn to estimating the magnitude of the gender-career bias in the each of the languages spoken in the countries described in Study 1 with the

[2]Available here: https://github.com/facebookresearch/fastText/

goal of examining the relationship between behavioral gender biases and language gender biases.

Method

Results

Study 3: grammar and bias

Study 4: exploring bias more directly

Conclusion

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv Preprint arXiv:1607.04606*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, philological society. Oxford. reprinted in Palmer, F.,(ed. 1968), Selected Papers of JR Firth, Longman, Harlow.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.