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

-> why these effects matter

Chugh, D. (2004). Societal and managerial implications of implicit social cognition: Why milliseconds matter. Social Justice Research, 17, 203– 222. doi:10.1023/B:SORE.0000027410.26010.40 Rudman, L. A. (2004). Social justice in our minds, homes, and society: The nature, causes, and consequences of implicit bias. Social Justice Re- search, 17, 129–142. doi:10.1023/B:SORE.0000027406.32604.f6

correlations are small with explict (oswald 2013), but matter Greenwald (2015; https://faculty.washington.edu/agg/pdf/Greenwald,Banaji&Nosek.JPSP.2015.pdf)

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' implicit 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 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 as quickly as possible by responding with one of the two keys. Slower reaction times in the bias-incongruent blocks relative to the bias-congruent blocks are interpreted as indicating an implicit association between the corresponding concepts (i.e. a bias to associate male with career, and female with family).

Method

We analyzed an existing IAT dataset collected online by 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 = 772,467). 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 663,709 participants from 48 countries, with a median of 998 participants per Country. Note that although the respondents were from largely non-English speaking countries, the IAT was conducted in English. We do not have language background data from the participants, but we assume that most respondents from non-English speaking countries were native speakers of the dominant language of the country and L2 speakers of English.

Several measures have been used in the literature to quantify gender biases from reaction time to congruent and incongruent blocks. Here, we used the most robust measure, D-score, which quantifies the difference between critical blocks for each participant while controlling for individual differences in response time (Greenwald, Nosek, & Banaji, 2003). For each country, we calculated an effect size as the mean D-score divided by its standard deviation (Cohen's d); larger values indicate greater bias.

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 indicate a greater bias to associate males with family.

We expected that countries with greater gender equality would have participants with lower implicit and explicit gender biases. As a measure of gender equality, we used the Women's Peace and Security Index (WPS, 2017), which measures inclusion, justice, and security of women by country, with larger values indicating higher gender equality.

Results

Broadly, we replicate the three patterns of findings in the literature on the gender-career IAT (Nosek et al., 2002). First, participants overall showed a bias to associate men with career and females with family (d = 1.08). Figure 1 shows the

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

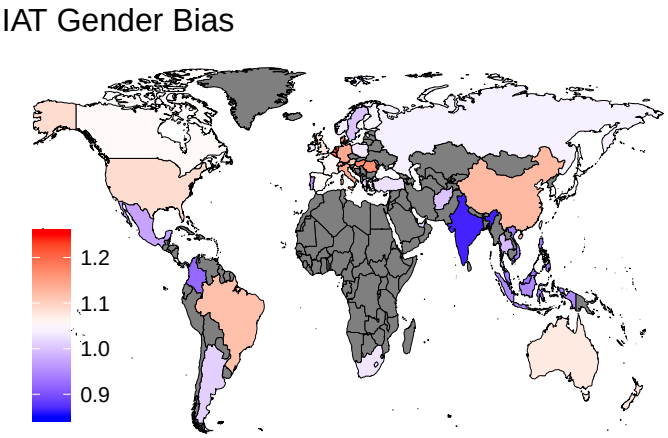

Figure 1: IAT gender bias effect size for 48 countries with available data. All countries show a gender bias, with red indicating above average and blue indicating below average bias.

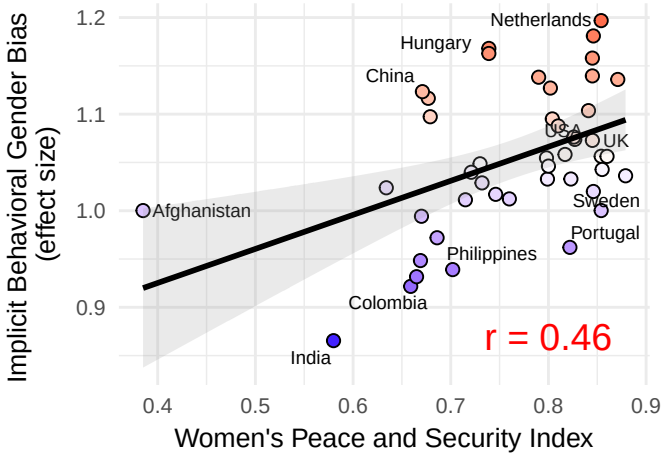

Figure 2: Magnitude of the implicit gender bias (measured by the IAT) predicted by an independent measure of gender equality, Women's Peace and Security Index (WPS). Each point corresponds to a country with notable points labeled. Contra our prediction, we find that countries with greater gender equality have larger gender implicit bias.

mean effect size for each of the 48 countries in our sample, with participants from all countries showing a gender bias (M = 1.05; SD = 0.07). Second, implicit and explicit bias measures were moderately correlated both at the level of individual participants ($r = 0.15$; $p < .00001$) and at the level of countries ($r = 0.31$; $p = 0.03$). Third, consistent with previous findings (Nosek et al., 2002), the magnitude of the implicit bias is larger for female participants compared to males (d = 0.26; $p < .00001$).

Our independent measure of gender equality—the Women's Peace and Security Index—was uncorrelated with explicit bias (r = -0.01; p = 0.96). Counter to our expectations, we found that countries such as the Netherlands, with allegedly greatest gender equality, have participants with the highest implicit gender bias according to the IAT ($r = 0.46$; p <.01; Fig. X).

Discussion

In Study 1, we replicate previously reported patterns of gender bias in the gender-career IAT literature, with roughly comparable effect sizes (c.f. Nosek, et al. (2002): overall effect: d = .72; explicit-implict correlation: r = .17; participant gender effect: d = .1). The weak correlation between explicit and implicit measures is consistent with claims that these two measures tap into different cognitive constructs (Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013).

The novel finding from Study 1 is the direction of the correlation between actual gender equality of a country (as measured by the WPS) and implicit gender bias—participantsw in countries with greater ginder equality have *greater* implicit gender bias. This robust correlation is particularly surpising given that the English was likely the second language for most of the participants in our sample.

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 semantics 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.

In Study 2, we ask whether participants' implicit and explicit gender biases are correlated with the semantic structure of different languages. For example, are the semantics of the words "woman" and "family" more similar in French than in English? To the extent that there are positive correlations between implicit and explicit gender-biases and semantics, we may speculate that the gender biases derive in part from exposure to language or else that the structure of the language reflects pre-existing gender biases. To model semantics, we turn to a machine-learning methods for deriving lexical semantics from large corpora of 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 as more similar to "hound" than "banana" because it occurs with words more in common with "hound" than "banana" where co-occurrences are effectively

defined at multiple hierarchical levels.

Recent developments in machine learning allow the idea of distributional semantics to be implemented 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 takes as input a corpus of text and outputs a vector for each word corresponding to its semantics. From these vectors, we can derive a measure of the semantic similarity between two words by taking the distance between their vectors (e.g., cosine distance). Similarity measures estimated from these models have been shown to be highly correlated with human judgements of word similarity (e.g., Hill, Reichart, & Korhonen, 2015), though more for some forms of similarity than others (Chen Dawn, Peterson, & Griffiths, 2017).

It turns out, that various human biases as measured by the IAT can be predicted from distributional semantics models like word2vec. Caliskan, Bryson, and Narayanan (2017; henceforth *CBN*) measured the distance in vector space between the same sets of words that are presented to participants in the IAT task. CBN found that these distance measures are highly correlated with reaction times in the behavioral IAT task. For example, in the career-gender IAT, CBN find a bias in the semantics of English to associate males with career and females with family, 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. While CBN only analyzed biases for models trained on English, we extend their method to compare biases across a wide number of languages. 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 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 represented by a 300 dimension vector.

Using the Wikpedia-trained model, we calculate an effect size for each of the 10 biases reported in CBN which correspond to behavioral IAT results existing in the literature: flowers/insects–pleasant/unpleasant, instruments/weapons–pleasant/unpleasant, European-American/Afro-American–pleasant/unpleasant,[3] males/females–career/family, math/arts–male/female, science/arts–male/female, mental-disease/physical-disease–permanent/temporary, and young/old–pleasant/unpleasant (labeled as WEAT 1-10 in CBN). 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 analogous the behavioral effect size measure in Study 1 and, like for the behavioral effect size, larger values indicate larger gender bias.

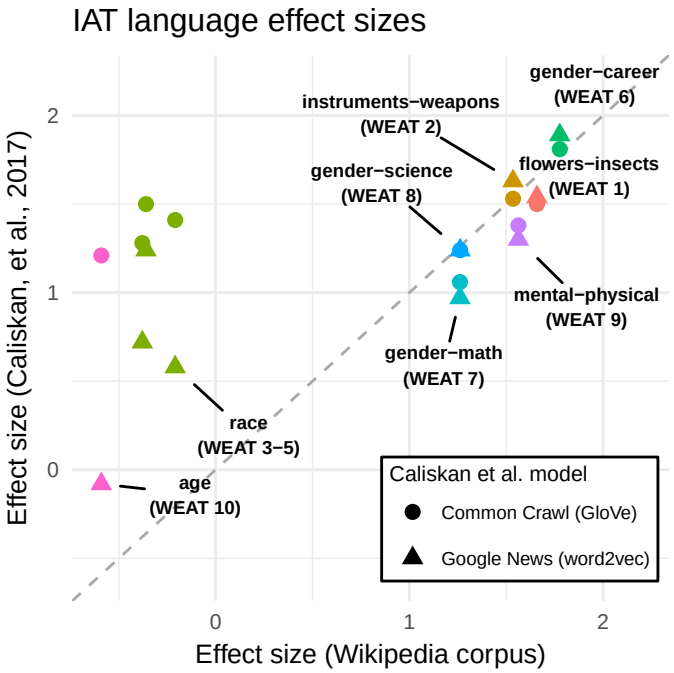

Figure 3: Effect sizes for the 10 IAT biases types (WEAT 1-10) reported in Caliskan et al. (2017; CBN). The effect sizes reported in CBN are plotted against effect sizes from the Wikipedia corpus. Point color corresponds to bias type, and point shape corresponds to the two CBN models trained on different corpora and with different algorithms.

Results Figure 2 shows the effect size measures derived from the Wikipedia corpus plotted against effect size estimates 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 comparable 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 as 1.81 (Common Crawl) and 1.89 (Google News).

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

[3] CBN test three versions of this bias.

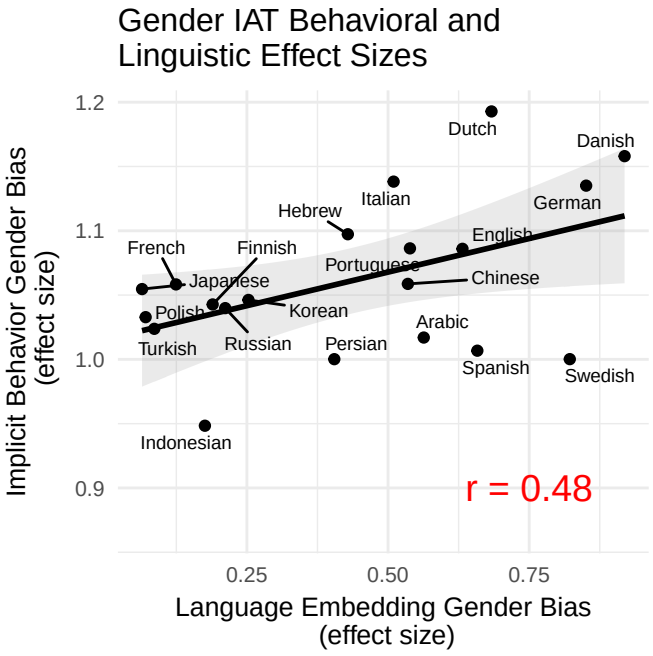

Figure 4: Gender bias effect size for each language from the behavioral IAT task (averaging across countries speaking the same primary language; Study 1) versus gender bias effect size estimated from embedding models trained on each language.

Study 2b: Predicting implicit bias with language IAT

With our corpus validated, we next turn toward examining the relationship between psychological and linguistic gender biases. In Study 2b, we estimate the magnitude of the gender-career bias in each of the languages spoken in the countries described in Study 1 and compare it with estimates of behavioral gender bias from Study 1. If language causally influences psychological gender bias, we predict these two measures should be positively correlated.

Method For each country included in Study 1, we identified the most frequently spoken language in those countries using the CIA factbook (Central Intelligence Agency, 2017). This included a total of 31 unique languages. For a sample of 20 of these languages (see Fig. 3), we had native speakers translate the set of 32 words from the gender-career IAT, with a slight modification[4]. The original gender-career IAT task (Nosek et al., 2002) used proper names to cue the male and female categories (e.g. "John," "Amy"). Because there are not direct translation equivalents of proper names across languages, we instead used a set of generic gendered words which had been previously used for a different version of the gender IAT (e.g., "male," "man," "female," "woman;" Nosek et al., 2002).

We used these translations to calculate an effect size from the models trained on Wikipedia in each language, using the same method as in Study 2a. We then compared the effect size of the linguistic gender bias to the behavioral gender bias, averaging across countries that spoke the same language and weighting by sample size.

Results Implicit IAT gender bias effect sizes were positively correlated with effect sizes of gender bias estimated from the native language embedding model (r = 0.48; p = 0.03; Fig. 3), suggesting that countries that have more gender bias encoded in their language also have a larger psychological gender bias. Explicit gender bias was not reliably correlated with language gender bias (r = 0.23; p = 0.33).

Discussion

Study 3: Grammar and Gender Bias

Study 2 suggests that psychological gender bias and linguistic gender bias are correlated, consistent with the idea that language may play a causal factor in shape gender bias in behavior. Nevertheless, Study 2 is also consistent with a second hypothesis in which the causal influence goes the opposition direction: greater psychological gender biases cause greater gender biases becoming encoded in the statistics of the language. In Study 3, we test the language-causal hypothesis more directly by examining whether there is a relationship between psychological gender bias and language along a linguistic dimension that is unlikely to be a subject of rapid change – namely, grammatical gender. While of course grammars do change, they are less malleable than the semantics of words, and thus less likely to be affected by psychological biases. We predict, therefore, that if there is some causal influence of language on psychological biases languages that encode gender grammatically will tend to have larger psychological gender biases.

Method

For each of the 31 languages represented in our sample of participants (Study 1), we coded whether gender was encoded grammatically. We used a coarse binary coding scheme, categorizing a language as encoding grammatical gender if it made any gender distinction on noun classes (male, female, common or neuter), and as not encoding gender grammatically otherwise. We coded this distinction on the basis of the WALS typological database where available (Feature 32a; Dryer & Haspelmath, 2013), and consulted additional resources as necessary. Our sample included 18 languages that encoded grammatical gender and 13 that did not.

Results

Languages that encode grammatical gender tended to have speakers with greater gender bias measured through the behavioral IAT (M = 1.07; SD = 0.08) compared to speakers of langauges that do not gramamtically encode gender (M = 1.02; SD = 0.07), though this difference was not reliable (d = 0.69 [-0.08, 1.45], t(27.53) = 1.89; p = 0.07; Fig. 4). In a post-hoc analysis, we excluded outliers located more than

[4]The language sample was determined by accessibility to native speakers, but included languages from a variety of language families.

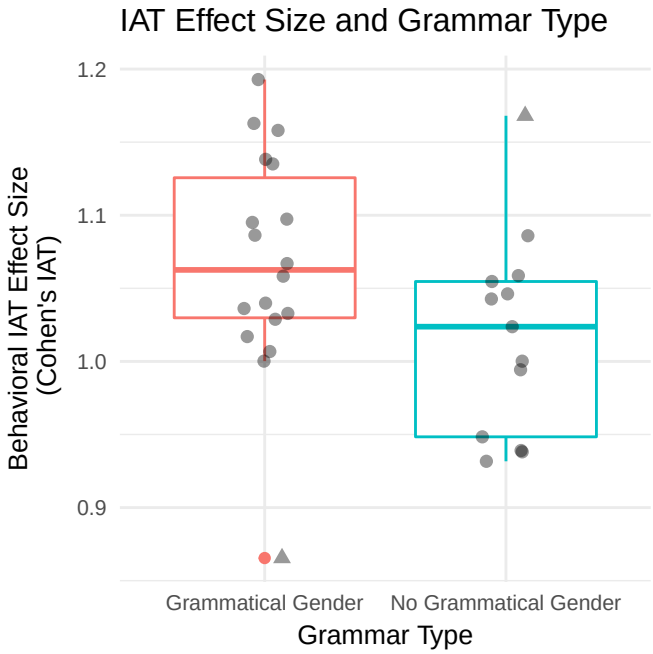

Figure 5: Behavioral IAT effect size as a function of whether participants' (assumed) native language encoded gender grammatically. Each point corresponds to a language (N = 31) with outliers shown as triangles (jittered along the x-axis for visibility).

two standard deviations from the group mean (Hungarian and Hindi). With these exclusions, we find a reliable difference between language types ($d = 1.3$ [0.46, 2.15], $t(25.12) = 3.46$; $p < .01$). In addition, we find the same pattern for language IAT (Study 2), with languages that encoded gender grammatical tending to have larger language IAT gender bias, compared to those who do not ($t(17.68) = 2.18$; $p = 0.04$).

Discussion

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.

Central Intelligence Agency. (2017). The World Factbook. Retrieved from `ttps://www.cia.gov/library/publications/the-world-factbook/index.html`

Chen, Dawn, Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv Preprint arXiv:1705.04416*.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/`

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.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of iat criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171.

Women's Peace and Security Index. (2017). Retrieved from `https://giwps.georgetown.edu/`