Language use shapes cultural norms: Large scale evidence from gender

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

Cultural norms vary dramatically across social groups. Here we use large scale data to examine the extent to which language plays a role in shaping one such norm—the gender norm to associate men with careers and women with family. We measure cross-cultural variability in this gender bias using previously-collected estimates from the Implicit Association Task (IAT; N = 663,709). We then try to predict bias variability by the way that gender is encoded in language semantics and grammar. We quantify gender bias in semantics using word-embedding models trained on different languages. Our data suggest that the linguistic encoding of gender predicts the degree of speakers' gender bias in the IAT, pointing to a causal role for language in shaping gender norms.

Keywords: cultural norms, IAT, gender, linguistic relativity

Introduction

The language we use to communicate a message shapes how our listener interprets that message (Loftus & Palmer, 1974; Tversky & Kahneman, 1981; Fausey & Boroditsky, 2010). A listener, for example, is more likely to infer that a person is at fault if the event is described actively (e.g., "she ignited the napkin"), as opposed to passively (e.g., "the napkin ignited"). The formative power of language is perhaps most potent in shaping meanings that necessarily must be learned from others: cultural norms. In the present paper, we consider one type of cultural norm—gender—and examine the extent to which differences in language use may lead to cross-cultural differences in understandings of gender.

Gender provides a useful case study of the relationship between language and thought for several reasons. First, more abstract domains like gender may be more subject to the influence of language relative to more perceptually grounded domains like natural kinds (Boroditsky, 2001). Second, many languages encode the gender of speakers and addressees explicitly in their grammar. Third, a large body of evidence suggests that language plays a key role in transmitting social knowledge to children (e.g., Master, Markman, & Dweck, 2012). And, fourth, gender norms are highly variable across cultures and have clear and important social implications.

For our purposes, we define the hypothesis space of possible relationships between language and gender norms with two broad extremes: (1) language reflects a pre-existing gender bias in its speakers (*language-as-reflection hypothesis*); (2) language causally influences gender biases (*language-as-causal hypothesis*). We assume that the language-as-reflection hypothesis is true to some extent: some of the ways we talk about gender reflect our knowledge and biases acquired independently of language. For example, we may observe that most nurses are women, and therefore be more likely to use a female pronoun to refer to a nurse of an unknown gender. Our goal here is to understand the extent to which language may also exert a causal influence on conceptualizations of gender.

In particular, we explore two possible mechanisms by which the way we speak may influence notions of gender[1]. The first is through the overt grammatical marking of gender, particularly on nouns, which is obligatory in roughly one quarter of languages (e.g., in Spanish, "niña" (girl) and "enfermera" (nurse) both take the gender marker *-a* to indicate grammatical femininity; Corbett, 1991). Because grammatical gender has a natural link to the real world, speakers may assume that grammatical markers are meaningful even when applied to inanimate objects that do not have a biological sex. In addition, the mere presence of obligatory marking of grammatical gender may promote bias by making the dimension of gender more salient to speakers.

A second route by which language may shape gender norms is via word co-occurrences. Words that tend to occur in similar contexts in language may lead speakers to assume—either implicitly or explicitly—that they have similar meanings. For example, statistically, the word "nurse" occurs in many of the same contexts as the pronoun "her," providing an implicit link between these two concepts that may lead to a bias to assume that nurses are female. This second route may be particularly influential because the bias is encoded in language in a way that is more implicit than grammatical markers of gender and thus more difficult to reject.

An existing body of experimental work points to a link between language and psychological gender bias in both adults (e.g., Phillips & Boroditsky, 2003) and children (e.g., Sera, Berge, & Castillo Pintado, 1994). For example, Phillips and Boroditsky (2003) asked Spanish-English and German-English adult bilinguals to make similarity judgements between pairs of pictures depicting an object with a natural gender (e.g., a bride) and one without (e.g., a toaster). They found that participants rated pairs as more similar when the pictures matched in grammatical gender in their native language. While these types of studies provide suggestive evidence for a causal link between language and psychological gender bias, they are limited by the fact that they typically only compare speakers of 2-3 different languages and measure bias in a way that is subject to demand characteristics.

In what follows, we ask whether the way gender is encoded linguistically across 31 different languages predicts cross-cultural variability in a particular manifestation of a gender

[1]These mechanisms are what Whorf (1945) refers to as phenotypes (overt) and cryptotypes (covert).

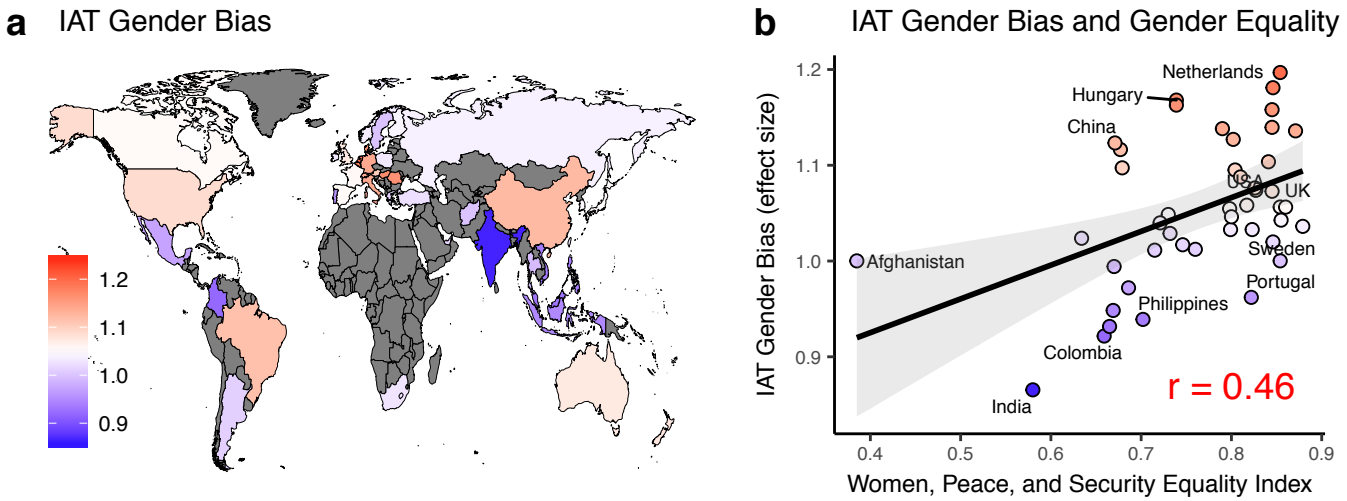

Figure 1: (a) Effect size (Cohen's d) for the bias to associate men with career and women with family, as measured by the IAT. All 48 countries with available data have a gender bias with red indicating above average and blue indicating below average bias. (b) Implicit gender bias predicted by an independent measure of gender equality, Women, Peace, and Security Index (WPS). Each point corresponds to a country with notable points labeled. Contra intuitions, we find that countries with greater gender equality have larger implicit gender bias.

bias—the bias to associate men with careers and women with family. We begin in Study 1 by describing cross-cultural variability in psychological gender bias using an implicit measure. In Study 2, we use semantic-embedding models to examine whether variability in lexical semantics predicts variability in psychological gender biases. In Study 3, we ask whether the presence of grammatical gender in a language is associated with greater implicit gender bias. Together, our data suggest that both language statistics and language structure likely play a causal role in shaping culturally-specific notions of gender.

Study 1: Gender bias across cultures

To quantify cross-cultural gender bias, we used data from a large-scale administration of an Implicit Association Task (IAT; Greenwald, McGhee, & Schwartz, 1998) by Project Implicit (Nosek, Banaji, & Greenwald, 2002). 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) accessed via words (e.g., "man," "business"). The underlying assumption of the IAT is that words denoting more similar meanings should be easier to pair together compared to more dissimilar pairs.

Meanings 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, meanings 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. 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).

In Study 1, we use the IAT to describe how the bias to associate women with family and men with careers varies across cultures. We find a gender bias in all countries and—unexpectedly—that the magnitude of this bias is positively correlated with objective gender equality.

Method

We analyzed an existing IAT dataset collected online by Project Implicit (`https://implicit.harvard.edu/implicit/`; Nosek et al., 2002). 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 error rates 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 the strength of the bias from participants' responses on congruent and incongruent blocks on the IAT. Here, we used the most robust measure, D-score, which measures the difference between critical blocks for each participant while controlling for individual differences in response time (Green-

wald, 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 career.

We compared implicit and explicit gender biases to a measure of objective gender equality, the Women, Peace, and Security Index (WPS, 2017). This metric describes the degree of inclusion, justice, and security of women by country, with larger values indicating higher gender equality.

Results

Our analyses confirm two key findings in the literature on the gender-career IAT (Nosek et al., 2002). First, participants showed an overall bias to associate men with career and females with family ($d = 1.08$). Figure 1a shows the 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 ($r = 0.31$; $p = 0.03$).

Our independent measure of gender equality—the Women, 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 high gender equality, have participants who show highest implicit gender bias according to the IAT ($r = 0.46$; $p < .01$; Fig. 1b).

We explored this surprising finding by testing two possible confounds with WPS. First, previous research has shown that women tend to have a larger implicit gender bias than men (Nosek et al., 2002). If countries with more gender parity have more female participants, this would explain the observed pattern. The data do not support this: In an additive linear model predicting IAT effect size with both WPS and the proportion female participants, proportion female participants was not a reliable predictor of IAT effect size ($\beta = 0.13$; $t = 0.8$; $p = 0.43$).

A second possibility is that participants in countries with greater gender parity have higher English proficiency and are thus relatively more influenced by the meaning of target words. We tested this possibility by including a measure of English proficiency as a covariate with WPS (EF English Proficiency Index, 2017). We find that English proficiency does not predict IAT effect size ($\beta < .001$; $t = 0.04$; $p = 0.97$), inconsistent with this explanation.

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: $d = .72$; explicit-implicit 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 (Forscher et al., 2016).

The novel finding from Study 1 is the direction of the correlation between the objective gender bias of a country (as measured by the WPS) and implicit gender bias—participants in countries with greater gender equality have *greater* implicit gender bias, even after controlling for possible confounds. In the General Discussion, we speculate about possible reasons for this positive correlation.

Study 2: Gender bias and semantics

In Study 2, we ask whether participants' implicit and explicit gender biases are correlated with the biases in the semantic structure of their native languages. For example, are the semantics of the words "woman" and "family" more similar in Hungarian than in English? Both the language-as-reflection and language-as-causal hypotheses predict a positive correlation between psychological and semantic gender biases. Importantly, we expect psychological and semantic gender biases to be correlated regardless of the direction of the relationship between psychological and objective gender bias (WPS) found in Study 1.

As a model of word meanings, we use large-scale distributional semantics models derived from auto-encoding neural networks trained on large corpora of text. 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 to "banana" because "dog" co-occurs with words more in common with "hound" than "banana."

Recent developments in machine learning allow the idea of distributional semantics to be implemented in a way that 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 correlated with human judgements of word similarity (e.g., Hill, Reichart, & Korhonen, 2015).

As it turns out, the biases previously reported using IAT tests can be predicted from distributional semantics models like word2vec using materials identical to those used in the IAT experiments. Caliskan, Bryson, and Narayanan (2017;

henceforth *CBN*) measured the distance in vector space between the words presented to participants in the IAT task. CBN found that these distance measures were highly correlated with reaction times in the behavioral IAT task. For example, CBN find a bias to associate males with career and females with family in the career-gender IAT, suggesting that the biases measured by the IAT are also found in the lexical semantics of natural language.

CBN only measured semantic biases in English, however. In Study 2, we use the method described by CBN to measure gender bias in the range of first languages spoken by participants in Study 1 by using models trained on those languages. To do this, we take advantage of a set of models pre-trained on corpora of Wikipedia text in different languages—a different corpus than that used by CBN (Bojanowski, Grave, Joulin, & Mikolov, 2016). In Study 2a, we replicate the original set of CBN findings using the model trained on English Wikipedia; In Study 2b, we apply this method to models trained on Wikipedia in other languages. We find 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 (a variant of word2vec; Bojanowski et al., 2016).[2] The model contains 2,519,370 words with each word represented by a 300 dimensional vector.

Using the Wikipedia-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 Word-Embedding Association Test (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 to the behavioral effect size measure in Study 1 where larger values indicate larger gender bias.

Results Figure 2 shows the effect size measures derived from the English 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 approximately 1.85.

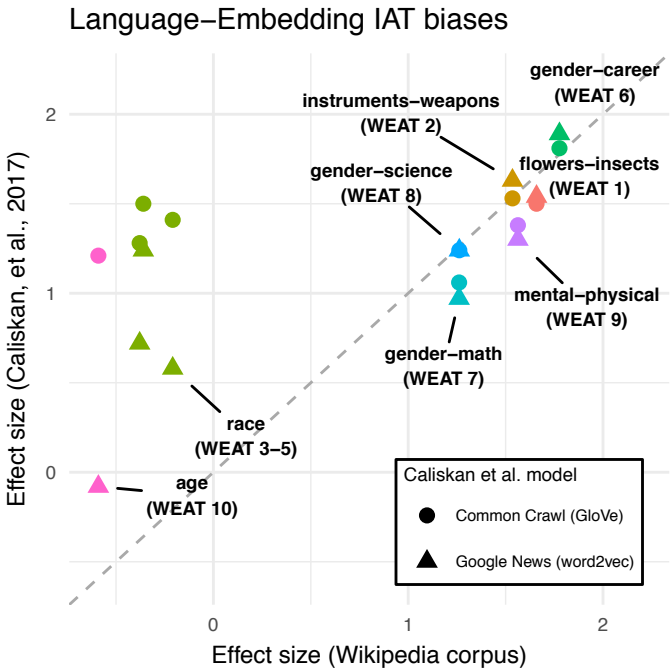

Figure 2: Study 2a: Effect sizes for the 10 IAT biases types (WEAT 1-10) reported in Caliskan et al. (2017; CBN). CBN effect sizes are plotted against effect sizes derived from the Wikipedia corpus. Point color corresponds to bias type, and point shape corresponds to the two CBN models trained on different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

Study 2b: Cross-linguistic gender semantics

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. 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 each country using the CIA factbook (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, we instead used a set of generic gendered words which had been previously used for a different version of the gender IAT (e.g., "man," "woman;" Nosek et al., 2002).

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

[3] CBN test three versions of this bias.

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

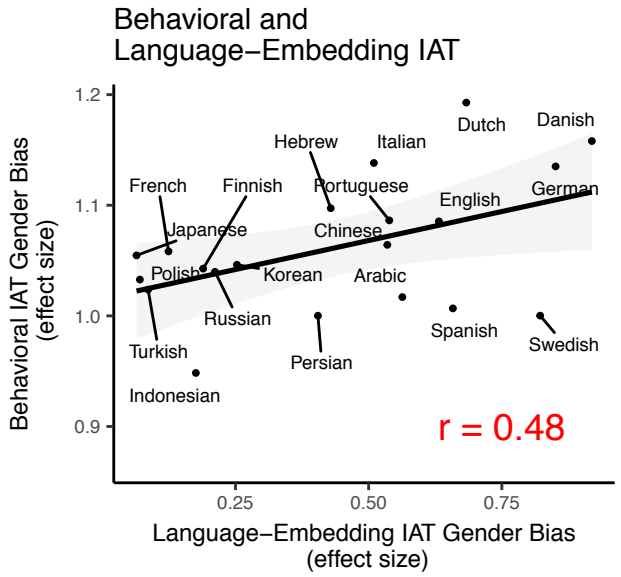

Figure 3: Study 2b: Behavioral IAT gender bias from Study 1 presented by language, versus language-embedding IAT gender bias. Language-embedding biases are estimated from models trained on each language.

	Behavioral IAT	Explicit Bias	WPS Index
Behavioral IAT			
Explicit Bias	.24		
WPS Index	.50*	-.09	
Language IAT	.48*	.23	.25

Table 1: Correlation (Pearson's r) for all measures in Studies 1 and 2. Asterisks indicate significance at the .05 level.

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 IAT gender bias from Study 1, averaging across countries that speak the same language and weighting by sample size.

Results and Discussion Table 1 presents the correlation between language IAT gender bias estimated from the native language embedding model and all other measures. Language IAT and Behavioral IAT effect sizes were positively correlated (r = 0.48; p = 0.03; Fig. 3). Explicit gender bias (r = 0.23; p = 0.33) and WPS index (r = .25; p = .29) were not reliably correlated with language gender bias.

Thus, as predicted, we find in Study 2 that countries with a larger gender bias in the semantics of their language tend to have speakers with greater implicit gender bias.

Study 3: Gender bias and grammar

The findings in Study 2 are consistent with both the language-as-causal and language-as-reflection hypotheses. In Study 3, we try to distinguish between the two hypotheses by asking whether there is a relationship between psychological gender bias and language along a linguistic dimension that is unlikely

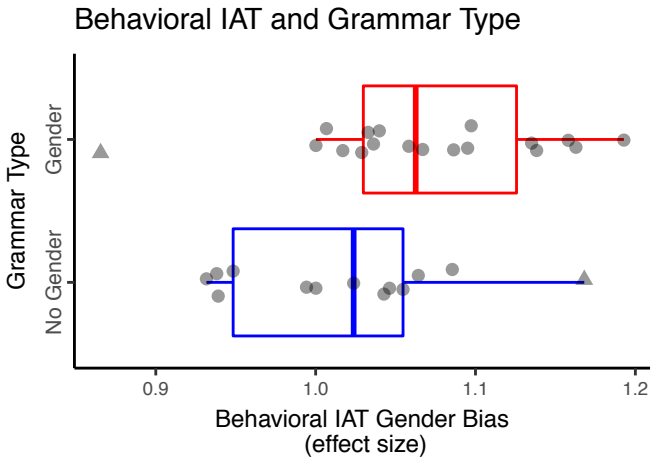

Figure 4: Study 3: Behavioral IAT gender bias from Study 1 as a function of whether participants' native language encodes gender grammatically. Each point corresponds to a language with outliers shown as triangles (jittered for visibility).

to be a subject of rapid change—namely, grammatical gender. While of course grammars do change over time, they are less malleable than the meanings of individual words, and thus less likely to be affected by psychological biases. We predict, therefore, that if language causally influences psychological gender biases, languages that encode gender grammatically will tend to have larger psychological gender biases.

Method

We coded each of the 31 languages in our sample (Study 1) for grammatical gender. We used a coarse binary coding scheme, categorizing a language as encoding grammatical gender if it made any gender distinction on noun classes (masculine, feminine, common or neuter), and as not encoding gender grammatically otherwise. We coded this distinction on the basis of the WALS typological database (Feature 32a; Dryer & Haspelmath, 2013) where available, and consulted additional resources as necessary. Our sample included 18 languages with and 13 without grammatical gender.

Results and Discussion

Languages that encode grammatical gender tended to have speakers with greater psychological gender bias (Study 1; M = 1.07; SD = 0.08) compared to speakers of languages that do not grammatically encode gender (M = 1.02; SD = 0.07), though this difference was not reliable (d = 0.68 [-0.08, 1.45], $t(27.48)$ = 1.87; p = 0.07; Fig. 4). In a post-hoc analysis, we excluded outliers located more than two standard deviations from the group mean (Hungarian and Hindi). With these exclusions, we find a reliable difference between language types (d = 1.29 [0.44, 2.14]; $t(25.02)$ = 3.43; $p < .01$).

In addition, we find the same pattern for language IAT (Study 2), with languages that encode gender grammatically tending to have larger language IAT gender biases, compared to those that do not ($t(17.68)$ = 2.18; p = 0.04).

In sum, Study 3 provides suggestive evidence that grammatical gender predicts implicit gender bias, as predicted by the language-as-causal hypothesis.

General Discussion and Conclusion

Across three studies, we explore the relationship between a culturally-constructed norm—gender—and the linguistic encoding of that norm. We find evidence for a close correspondence: Languages that have larger gender biases encoded in their lexical semantics (Study 2) and have grammatical gender markers (Study 3) tend to have speakers with larger implicit gender bias. Study 2 is consistent with both the language-as-reflection and the language-as-causal hypotheses, but Study 3 is most consistent with the language-as-causal hypothesis. Taken together, the most likely interpretation of our data is that both mechanisms are at play and act synergistically, such that the way we talk about gender shapes the way we think about it and the way we think about gender shapes the way we talk about it.

In addition, our work is the first to report the surprising positive correlation between implicit gender bias and objective gender equality. The source of this correlation is difficult to interpret in part because researchers do not agree on the nature of the construct that the IAT measures or its causal relationship to explicit bias and behavior (Forscher et al., 2016). One provocative implication, however, is that there is a causal relationship between individual and institutional sexism. We speculate that greater gender equality could emerge as a consequence of *increased* attention to gender inequality, leading to both objective equality but also increased implicit bias at the individual level. An important next step for understanding this relationship will be to examine whether this pattern holds for other types of biases, such as race.

More generally, the task for future work will be to specify the dynamics of the causal mechanisms between language and cultural norms with more precision. The ultimate goal is to describe the relative influence of different aspects of language—semantics and structure—on cultural norms and vice versa, particularly when children are first acquiring cultural knowledge in development. The data here provide an early step toward this goal.

All code and data for this project are available at `https://github.com/mllewis/IATLANG`

References

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

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

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 (CIA). (2017). The World Factbook. Retrieved from `https://www.cia.gov/library/publications/the-world-factbook/index.html`

Corbett, G. G. (1991). *Gender*. Cambridge: Cambridge University Press.

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

EF English Proficiency Index. (2017). Retrieved from `https://www.ef.edu/epi/`

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review*, *17*(5), 644–650.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, philological society. Oxford.

Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2016). A meta-analysis of change in implicit bias.

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: 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.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589.

Master, A., Markman, E., & Dweck, C. (2012). Thinking in categories or along a continuum: Consequences for children's social judgments. *Child Development*, *83*(4).

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

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.

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 928–933).

Sera, M. D., Berge, C. A., & Castillo Pintado, J. del. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development*, *9*(3), 261–292.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Whorf, B. (1945). Grammatical categories. *Language*, 1–11.

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