What are we learning from language? Associations between gender biases and distributional semantics in 25 languages

Molly Lewis[1,2] & Gary Lupyan[1]

[1] University of Wisconsin-Madison

[2] University of Chicago

Author Note

Portions of this manuscript appeared in Lewis & Lupyan, 2018, Cog. Sci. Proceedings.

Correspondence concerning this article should be addressed to Molly Lewis, . E-mail: mollyllewis@gmail.com

Abstract

Cultural stereotypes such as the idea that men are more suited for paid work while women for taking care of the home and family may contribute to gender imbalances in STEM fields (e.g., Leslie, Cimpian, Meyer, & Freeland, 2015) and other undesirable gender disparities. Here, we test the hypothesis that word co-occurrence statistics (e.g., the co-occurrence of "nurse" with "she") play a causal role in the formation of the men-career/women-family stereotype. We use word embedding models to measure bias in the distributional statistics of 25 languages and find that languages with larger biases tend to have speakers with larger implicit biases (N = 657,335). These biases are further related to the extent that languages mark gender in their lexical forms (e.g., "waiter"/"waitress") hinting that linguistic biases may be causally related to biases shown in people's implicit judgments.

Keywords: cultural stereotypes, implicit association task (IAT), gender

Word count: 1998 (excluding methods/results)

What are we learning from language? Associations between gender biases and distributional semantics in 25 languages

Introduction

By the time children are two years old, they have begun to acquire the gender stereotypes in their culture (Gelman et al., **???**). These stereotypes can have undesirable effects. For example, in one study, 6-year-old girls were less likely than boys to choose activities that were described as for children "who are very, very smart" and also less likely to think of themselves as "brilliant" (**???**). Such beliefs may, over time, translate to the observed lower rates of female participation in STEM fields (**???**; **???**; **???**; **???**). For this reason and others, it is important to better understand how cultural stereotypes are formed.

We can distinguish between two major sources of information on which gender stereotypes may be based. The first is direct experience. For example, one may observe that most nurses are women and most philosophers are men and conclude that women are better suited for nursing and men for philosophy. The second is language. Even without any direct experience with nurses or philosophers, one may learn about their stereotypical gender from language about nurses and philosophers. Languages encode gender in multiple ways. These include gender-specific titles ("Mr". vs. "Miss"), proper names ("Sam" vs. "Ashley"), pronouns ("he" vs. "she"), certain job titles ("waiter" vs. "waitress"), and higher-order linguistic associations (otherwise gender-neutral words can become gendered by being associated with explicitly gendered contexts). Another source of linguistic information comes from sex-based grammatical gender systems found in approximately 30% of languages (Dryer & Haspelmath, 2013). For example, in Spanish, the gender of a nurse must be specified grammatically ("enfermer*a*" vs. "enfermer*o*").

To the extent that language is a source of information for forming cultural stereotypes, two people with similar direct experiences, but different linguistic experiences, may develop

different stereotypes. Some past work hints at people's surprising sensitivity to stereotype-relevant information delivered through language. Young children perform worse in a game if they are told that someone of the opposite gender performed better than they did on a previous round (**???**), or merely told that the game is associated with a particular gender (**???**). In some cases, a subtle turn of phrase can influence children's gender-based generalization (**???**; **???**). For example, Cimpian and Markman found that children were more likely to infer that a novel skill is stereotypical of a gender if the skill is introduced with a generic as opposed a non-generic subject ("[Girls are/There is a girl who is] really good at a game called "gorp""). Such work shows that in certain experimental settings, language can influence stereotype formation. We were interested in whether it actually does, and by what means.

A widely used method for quantifying cultural stereotypes at an individual level is the *Implicit Association Test* (IAT; **???**). Here, we use previously administered IATs designed to measure a particular type of gender stereotype: A bias to associate men with careers and women with family (N = 657,335; Nosek, et al., 2002). These data span native speakers of 25 languages allowing us to assess how performance varies with properties of languages.

Discovering that linguistic bias predicts people's implicit biases can be interpreted in at least two ways. The first is that some cultures have stronger stereotypes and these are reflected in what people talk about. Language, on this view, simply *reflects* pre-existing biases. We refer to this as the *language as reflection* hypothesis. However, language may not simply reflect pre-existing biases, but may also provide a distinct source of information for learning about these stereotypes. We refer to this second possibility that language exerts a causal influence on people's biases as the *language as causal factor* hypothesis.

In Study 1, we examine whether language-derived gender biases predict responses on the gender-career IAT. Our analysis focuses on the *distributional* structure of language rather than the specifics of the communicated content. In Study 2, we examine how the

psychological biases measured by the IAT and the linguistic biases we measure relate to more structural aspects of language: sex-based grammatical gender and the prevalence of gender-specific occupation terms (e.g., "waiter"/"waitress" but "teacher"/"teacher"). The results of Study 2 suggest that language not only reflects existing gender biases, but may play a causal role in shaping them.

Description of Cross-Cultural Dataset of Psychological Gender Bias

Materials and Methods

To quantify cross-cultural gender bias, we used data from a large-scale administration of an Implicit Association Task (IAT; **???**) by Project Implicit (https://implicit.harvard.edu/implicit/; **???**). 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 two-alternative forced-choice 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 (see Study 1b Methods for list of target words). 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).

We analyzed gender-career IAT scores collected by Project Implicit between 2005 and

2016, restricting our sample based on participants' reaction times and error rates using the same criteria described in Nosek, Banjai, and Greenwald (2002, pg. 104). We only analyzed data for countries that had complete demographic information and complete data from the IAT for least 400 participants (2% of these respondents did not give responses to the explicit bias question). This cutoff was arbitrary, but the pattern of findings reported here holds for a range of minimum participant values (see SM[1]). Our final sample included 657,335 participants from 39 countries, with a median of 1,145 participants per country. Importantly, 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 a large fraction of the respondents from non-English speaking countries were native speakers of the dominant language of the country and second language speakers of English. The fact that the test was administered in English lowers the prior likelihood of finding language-specific predictors of the kind we report here.

To quantify participants' performance on the IAT we adopt the widely used *D-score*, which measures the difference between critical blocks for each participant while controlling for individual differences in response time (**???**). 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-career 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.

[1]SM available here: https://mollylewis.shinyapps.io/iatlang_SI/; All data and code available here: https://github.com/mllewis/IATLANG

Results

There was a reliable bias for participants to associate men with career and women with family (M = 0.38 [0.38, 0.38]; $t(657334) = 878.3$, $p < .0001$). At the participant level, implicit bias scores were positively correlated with participant age ($r(657333) = 0.06$ [0.06, 0.06], $p < .0001$). Male participants ($M = 0.32$, $SD = 0.37$) had a significantly smaller implicit gender bias than female participants ($M = 0.41$, $SD = 0.35$; $M = 0.09$ [0.09, 0.1]; $t(338217.04) = 96.82$, $p < .0001$; $d = 0.27$ [0.26, 0.27]), a pattern consistent with previous findings (**???**). Implicit bias scores were larger for participants that received the block of trials with bias-incongruent mappings first relative to the opposite order (M = -0.09 [-0.09, -0.09]; $t(652694.18)$ = -104.03, $p < .0001$; d = -0.26 [-0.26, -0.25]).

Because we did not have language information at the participant level, in the remaining analyses we examine gender bias and its predictors at the country level. To account for the above-mentioned influences on implicit bias, we calculated a residual implicit bias score for each participant, controlling for participant age, participant sex, and block order. We also calculated a residual explicit bias score controlling for the same set of variables. We then averaged across participants to estimate the country-level gender bias (Implicit: M = -0.01; $SD = 0.03$; Explicit: $M = 0.00$; $SD = 0.18$). Implicit gender biases were moderately correlated with explicit gender biases at the level of participants ($r(645072)$ = 0.16 [0.16, 0.16], $p < .0001$) but not countries ($r(37) = 0.26$ [-0.07, 0.53], $p = 0.116$)

Do the implicit and explicit biases measured by the Project Implicit dataset predict any real world outcomes? We compared our residual country-level implicit and explicit gender biases to a gender equality metric reported by the United Nations Educational, Scientific and Cultural Organization (UNESCO) for each country: the percentage of women among STEM graduates in tertiary education from 2012 to 2017 (Miller et al., 2015; Stoet & Geary, 2018; available here: http://data.uis.unesco.org/). These data were available for 33

out of 39 of the countries in our sample. Consistent with previous research (**???**), we found that implicit gender bias was negatively correlated with percentage of women in STEM fields: Countries with smaller gender biases tended to have more women in STEM fields ($r(31) =$ -0.54 [-0.75, -0.24], $p = 0.001$). In contrast, there was no relationship between the percentage of women in STEM fields and the explicit gender bias measure used by Project Implicit ($r(31) = 0.14$ [-0.21, 0.46], $p = 0.433$). In addition, we found a strong correlation between the median age of each country's population (as reported by the CIA factbook, 2017) and the residual implicit bias (in which participant age was held constant): Countries with older populations tended to have larger gender biases ($r(37) = 0.64$ [0.4, 0.79], $p < .0001$).

In sum, 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). The weak correlation between implicit and explicit measures is consistent with claims that these two measures tap into different cognitive constructs (**???**). In addition, we find that an objective measure of gender equality—female enrollment in STEM fields—is associated with implicit gender bias. The finding that older participants show stronger biases may stem from a cohort effect, but it is not obvious why there is a strong positive association between the median age of a country's population and a larger implicit bias when adjusting for the age of individual participants.

Study 1: Relating gender biases in distributional semantics and human behavior

Are participants' gender biases predictable from the language they speak? Both the language-as-reflection and language-as-causal-factor hypotheses predict a positive correlation between the two, but showing that such a relationship exists is the first step to investigating a possible causal link. We begin by validating word embedding measures of gender bias by comparing them to explicit human judgements of word genderness (Study 1a). We then

apply this method to models trained on text in other languages (Study 1b). We find that the implicit gender bias of participants in a country is correlated with the gender bias in the language spoken in that country.

Study 1a: Word embeddings as a measure of psychological gender bias

Methods. To model word meanings, we use semantic embeddings derived from a model that learns meanings by trying to predict a word from surrounding words, given a large corpus. The core assumption of these models is that the meaning of a word can be described by the words it tends to co-occur with—words occurring in similar contexts, tend to have similar meanings (**???**). A word like "dog," for example is represented as more similar to "cat" and "hound" than to "banana" because "dog" co-occurs with words more in common with "cat" and "hound" than with "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 language structure while remaining computationally tractable. The best known of these word embedding models is *word2vec* (**???**). By attempting to predict the words that surround another word, the model is able to learn a vector-based representation for each word that represents its similarity to other words, i.e., a semantic embedding. We can then compute the similarity between two words by taking the distance between their vectors (e.g., cosine of angle).

In order to validate word embeddings as a measure of psychological gender bias, we used an existing set of word norms in which participants were asked to rate "the gender associated with each word" on a Likert scale ranging from *very feminine* (1) to *very masculine* (7; **???**). We compared these norms to estimates of gender bias obtained from embedding models pre-trained on two different corpora of English text: Wikipedia (**???**) and subtitles from movies and TV shows (**???**; **???**). The Wikipedia corpus is a large, naturalistic corpus of written language; the Subtitle corpus is a smaller corpus of spoken

language. Both models were trained using the fastText algorithm (a variant of word2vec; **???**). There were 4,671 words in total that overlapped between the word-embedding models and human ratings.

Using the word embeddings, we calculated an estimate of gender bias for each word by measuring the average cosine distance to a standard set of male "anchor" words ("male," "man," "he," "boy," "his," "him," "son," and "brother"; Nosek, Banaji, & Greenwald, 2002) and the average cosine similarity to a set of female words ("female," "woman," "she," "girl," "hers," "her," "daughter," and "sister"). A gender score for each word was then obtained by taking the difference of the similarity estimates (mean male similarity - mean female similarity), such that larger values indicated a stronger association with males.

Results. Estimates of gender bias from the Subtitle corpus ($M = 0.01$; $SD = 0.03$) and the Wikipedia corpus ($M = 0$; $SD = 0.03$) were highly correlated with each other ($r(4669) = 0.71$ [0.7, 0.73], $p < .0001$). Critically, bias estimates from both word embedding models were also highly correlated with human judgements ($M = 4.10$; $SD = 0.92$; $r_{\text{Subtitle}} = r(4669) = 0.63$ [0.61, 0.65], $p < .0001$; $r_{\text{Wikipedia}} = r(4669) = 0.59$ [0.57, 0.6], $p < .0001$; Fig. 1). This suggests that the psychological gender bias of a word can be reasonably estimated from word embeddings.

Study 1b: Gender bias across languages

Having validated our method, we now use it to examine the relationship between psychological and linguistic gender biases. In Study 1b, we estimate the magnitude of the linguistic bias in the dominant language spoken in each country represented in the Project Implicit dataset, and compare this estimate to estimates of psychological gender bias from the Project Implicit participants.

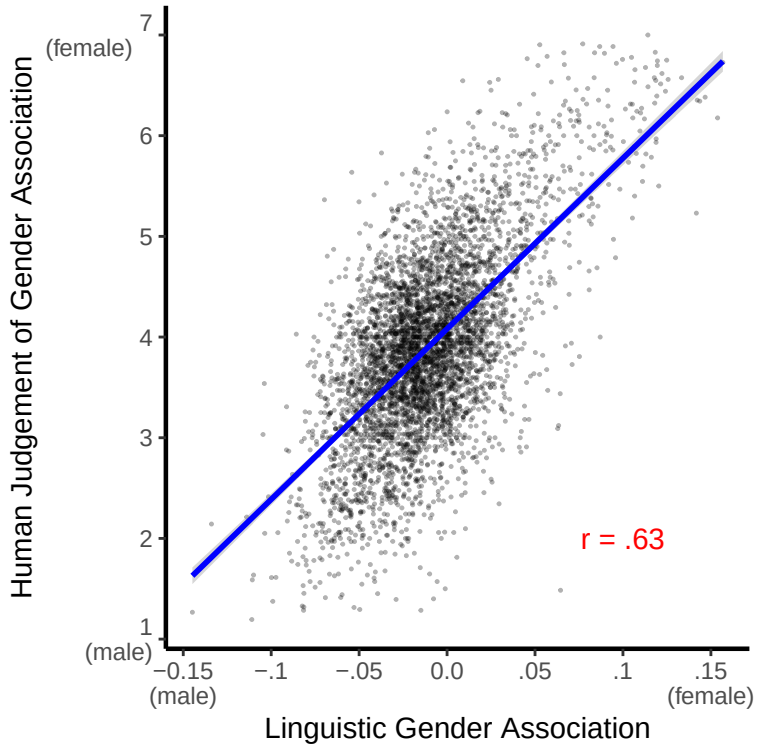

Figure 1. Human judgements of word gender bias as a function of gender bias from the Subtitle-trained embedding model (Study 1a). Each point corresponds to a word. Larger numbers indicate stronger association with females (note that this differs from the design of the rating task, but is changed here for consistency with other plots). Blue line shows linear fit and the error band indicates standard error (too small to be visible).

Methods. Previous work has shown biases studied using IATs can be predicted from the distributional statistics of language (word co-occurrences). Using these statistics, Caliskan, Bryson, and Narayanan (2017; henceforth, CBN) measured the distance between the words presented to participants in the IAT task. CBN found that these distances were highly correlated with the biases computed by a variety of IATs (e.g., valence and Caucasian vs. African-American names; gender and math vs. arts; permanence and mental vs. physical diseases). CBN only measured semantic biases in English. Here, we extend CBN's method to 25 languages examining whether languages with a stronger gender bias as expressed in distributional semantics predict stronger implicit and explicit gender biases on a large

dataset of previously administered gender-career IATs.

We identified the most frequently spoken language in each country in our analysis using Ethnologue (**???**). After exclusions (see below), our final sample included 25 languages.[2] For each language, we obtained translations from native speakers for the stimuli in the Project Implicit gender-career IAT behavioral task (**???**) with one slight modification. In the behavioral task, proper names were used to cue the male and female categories (e.g. "John," "Amy"), but 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;" **???**). Our linguistic stimuli were therefore a set of 8 female and 8 male Target Words (identical to Study 1a), and the set of 8 Attribute Words words used in the Project Implicit gender-career IAT: 8 related to careers ("career," "executive," "management," "professional," "corporation," "salary," "office," "business") and 8 related to families ("family," "home," "parents," "children," "cousins," "marriage," "wedding," "relatives"). For one language, Filipino, we were unable to obtain translations from a native speaker, and so Filipino translations were compiled from dictionaries.

We used these translations to calculate a gender bias effect size from word embedding models trained on text in each language. Our effect size measure is 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). Our effect size measure is identical to that used by CBN with an exception for grammatically gendered languages (see SM for replication of CBN on our corpora). Namely, for languages with grammatically gendered Attribute Words (e.g., niñas for female children in Spanish), we calculated the relationship between Target Words and Attribute Words of the same gender

[2]Note that while Hindi is identified as the most frequently spoken language in India, India is highly multilingual and so Hindi embeddings may be a poor representation of the linguistic statistics for speakers in India as a group.

(i.e. “hombre” (man) to “niños” and “mujer” (woman) to “niñas”). In cases where there were multiple translations for a word, we averaged across words such that each of our target words was associated with a single vector in each language. In cases where the translation contained multiple words, we used the entry for the multiword phrase in the model when present, and averaged across words otherwise. Like the psychological measures of bias from the Project Implicit data, larger values indicate larger gender bias.

We calculated gender bias estimates using the same word embedding models as in Study 1a (Subtitle and Wikipedia corpora). We excluded languages from the analysis for which 20% or more of the target words were missing from the model or the model did not exist. This led us to exclude one language (Zulu) from the analysis of the Wikipedia corpus and six languages from the analysis of the Subtitle corpus (Chinese, Croatian, Hindi, Japanese, Filipino, and Zulu). Our final sample included 25 languages in total ($N_{\text{Wikipedia}} = 25$; $N_{\text{Subtitle}} = 20$), representing 8 language families. Finally, we calculated language-level measures for four additional measures by averaging across countries whose participants speak the same language: implicit and explicit psychological gender bias (estimated from the Project Implicit dataset), percentage of women in STEM fields, and median country age.

Results. Despite the differences in the specific content conveyed by the Subtitles and Wikipedia corpus, the estimated gender bias for each language was similar across the two corpora ($M = 0$ [-0.17, 0.16]; $t(19) = -0.06$, $p = 0.95$; $d = -0.01$ [-0.65, 0.63]) . We next examined the relationship between these estimates of gender bias for each language and the mean IAT bias score for participants from countries where that language was dominant (and, we assume, was the native language of most of these individuals). Implicit gender bias was positively correlated with estimates of language bias from both the Subtitle ($r(18) = 0.5$ [0.08, 0.77], $p = 0.024$) and Wikipedia trained models ($r(23) = 0.48$ [0.11, 0.74], $p = 0.015$; Fig. 2; Table 1 shows the language-level correlations between all variables in Studies 1b and 2). The relationship between implicit gender bias and language bias remained reliable after

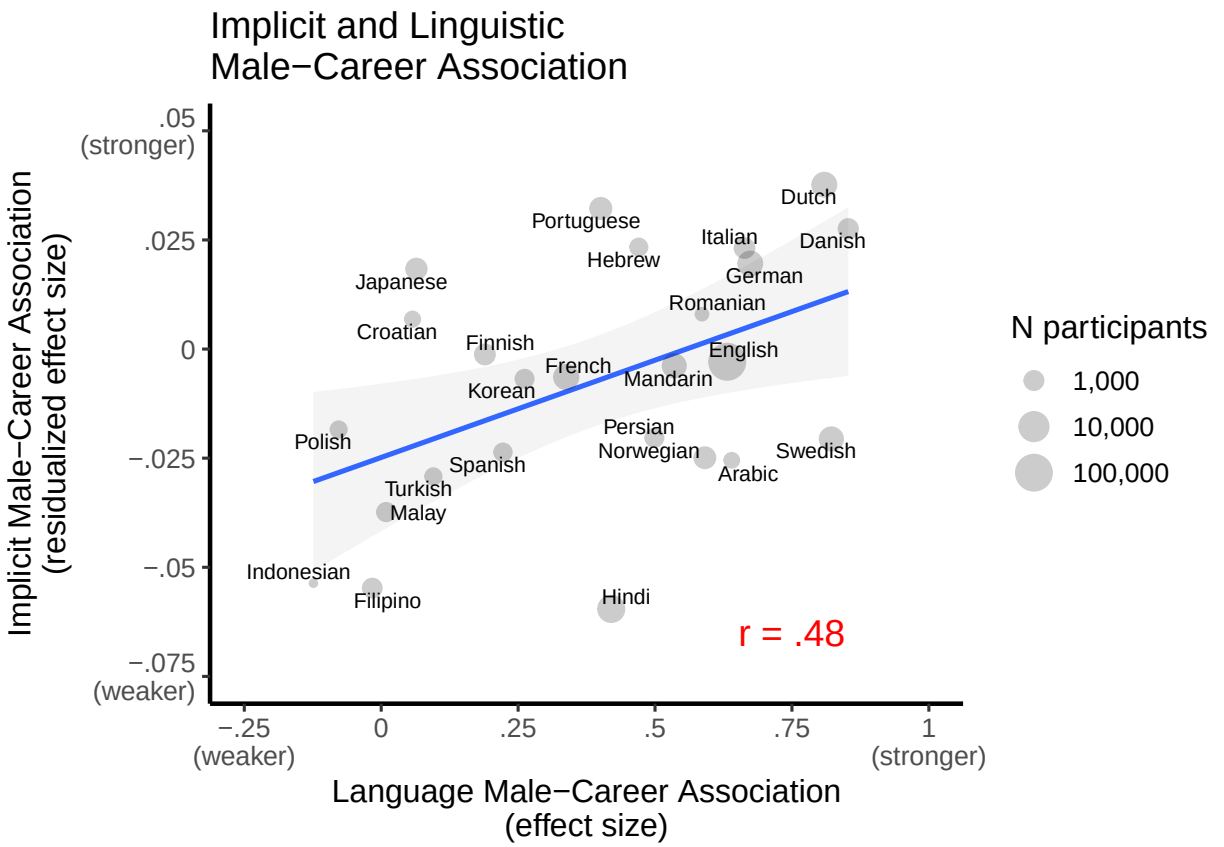

Figure 2. Implicit gender bias (adjusted for age, sex, and block order) as a function of the linguistic gender bias derived from word-embeddings (Study 1b). Each point corresponds to a language, with the size of the point corresponding to the number of participants speaking that langauge. Linguistic biases are estimated from models trained on text in each language from Subtitle (left) and Wikipedia (right) corpora. Larger values indicate a larger bias to associate men with the concept of career and women with the concept of family. Error bands indicate standard error of the linear model estimate.

partialling out the effect of median country age (Subtitle: $r = 0.42$, $p = 0.04$; Wikipedia: $r = 0.43$, $p = 0.04$). Linguistic gender bias was not correlated with explicit gender bias (Subtitle: $r(18) = -0.08$ [-0.5, 0.38], $p = 0.737$; Wikipedia: $r(23) = 0.34$ [-0.06, 0.65], $p = 0.093$). Estimates of language bias from the Subtitle corpus were correlated with the objective measure of gender equality, percentage of women in STEM fields ($r(16) = -0.55$ [-0.81, -0.11], $p = 0.018$); this relationship was not reliable for the Wikipedia corpus ($r(20) =$

-0.19 [-0.57, 0.25], $p = 0.401$).

Table 1

Correlation (Pearson's r) for all measures in Study 1 and 2 at the level of languages. Top panel shows simple correlations; bottom panel shows partial correlations controlling for median country age. Single asterisks indicate $p < .05$ and double asterisks indicate $p < .01$. The + symbol indicates a marginally significant p-value, $p < .1$.

	Residualized Explicit Bias	Residualized Implicit Bias (IAT)	Percent Women in STEM	Language IAT (Subtitle)	Language IAT (Wikipedia)	Prop. Gendered Occupation Labels	Occupation Bias (Subtitle)	Occupation Bias (Wikipedia)	Median Country Age
Simple Correlations									
Residualized Explicit Bias		.18	.18	-.08	.34+	.11	.16	.18	-.07
Residualized Implicit Bias (IAT)	.18		-.53*	.50*	.48*	.57**	.49*	.49*	.61**
Percent Women in STEM	.18	-.53*		-.55*	-.19	-.35	-.26	-.53*	-.42+
Language IAT (Subtitle)	-.08	.50*	-.55*		.51*	.28	.38	.41+	.31
Language IAT (Wikipedia)	.34+	.48*	-.19	.51*		.18	.51*	.53**	.25
Prop. Gendered Occupation Labels	.11	.57**	-.35	.28	.18		.60**	.77**	.35+
Occupation Bias (Subtitle)	.16	.49*	-.26	.38	.51*	.60**		.81**	.44+
Occupation Bias (Wikipedia)	.18	.49*	-.53*	.41+	.53**	.77**	.81**		.34+
Median Country Age	-.07	.61**	-.42+	.31	.25	.35+	.44+	.34+	
Partial Correlations									
Residualized Explicit Bias		.28	.16	-.06	.38+	.14	.21	.22	
Residualized Implicit Bias (IAT)	.28		-.38+	.42*	.43*	.48*	.31	.37+	
Percent Women in STEM	.16	-.38+		-.49*	-.09	-.23	-.10	-.46*	
Language IAT (Subtitle)	-.06	.42*	-.49*		.47*	.20	.28	.35+	
Language IAT (Wikipedia)	.38+	.43*	-.09	.47*		.11	.46*	.49*	
Prop. Gendered Occupation Labels	.14	.48*	-.23	.20	.11		.53**	.73**	
Occupation Bias (Subtitle)	.21	.31	-.10	.28	.46*	.53**		.79**	
Occupation Bias (Wikipedia)	.22	.37+	-.46*	.35+	.49*	.73**	.79**		

Study 1c: A pre-registered study of British versus American English biases

In Study 1c, we conducted a confirmatory, pre-registered analysis of our hypothesis that biases present in language statistics are reflected in the psychological biases of speakers

of those languages. To do this, we leveraged an exisiting dataset from the Attitiudes, Identities, and Individual Differences Study (AIID; **???**) containing measures of IAT performance from over 200,000 participants for a wide range of IAT types (e.g. career - family, team - individual, etc.). Because most participants in this sample were English speakers, we compared biases between participants who spoke two different dialects of English: British and American. For each of the 31 IAT types in the sample, we predicted that the degree to which that bias was present in a speaker's English dialect (British or American) would predict the magnitude of their psychological bias, as measured by the IAT.

Method

The AIID datset was partioned into two samples: exploratory (15%) and confirmatory (85%). Based on the exploratory sample, we pre-registered our analysis plan for the confirmatory sample (https://osf.io/3f9ed). We note where are analysis diverges from the preregistration plan below.

Of the 95 IAT types present in the dataset, we identified 31 types based on the following criteria: (1) stimuli were words rather than pictures, and (2) 75% of the target words for each IAT test were present in both the BNC and COCA models. To measure the bias in language, we trained word embedding models on comparably-sized corpora of British (British National Corpus; **???**) and American English (Corpus of Contemporary American English; see SM for details; **???**). We then calculated a language bias effect size for each IAT in each English dialect, using the same method as in Study 1b.

Within the confirmatory AIID dataset, there were 187,969 administrations of the IAT from participants in the United States (USA) or United Kingdom (UK). Our exclusion procedure for the behavioral data was similiar to Study 1a above (see SM for details). Our final sample included data from 135,240 administrations of the IAT across the 31 IAT types

(USA: N = 127,630; UK: N = 7,610). Each participant completed an average of 6.13 IAT types (SD = 3.99). For each administrations of the IAT, we calculated a residual IAT score, controlling for sex, age, education, task order (relative ordering of implicit versus explicit measures), and block order (relative ordering of congruent versus incongruent mappings on IAT task).

We fit a linear mixed effect model predicting the magnitude of the bias for each participant with language dialect (British or American English), magnitude of the bias difference between the two dialects, and an interaction betwen these two terms as fixed effect. We included participant and IAT type as random intercepts. This model differs from the pre-registered analysis, which is also consistent with results of the presented analysis, but does not account for participant-level variance (see SM for results of pre-registered model).

Results

As predicted, there was a reliable interaction between language dialect and the magnitude of the bias difference between the two dialects (β = .05, SE = .02, t = 2.88; see SM for full model results), suggesting that language bias was a reliable predictor of implicit bias. Figure 3 shows the difference in bias (British - American) in language and behavior for each of the 31 IATs in our dataset.

Discussion. In Study 1, we found that a previously-reported psychological gender bias – the bias to associate men with career and women with family – was correlated with the magnitude of that same bias as measured in the language statistics of 25 languages. Participants completing the IAT in countries where the dominant language had stronger associations between men and career words, and women and family words, showed stronger biases on the gender-career IAT. In a pre-registered, confirmatory analysis, we also find that this pattern extends to other biases: In a comparision of 31 different IAT types, the

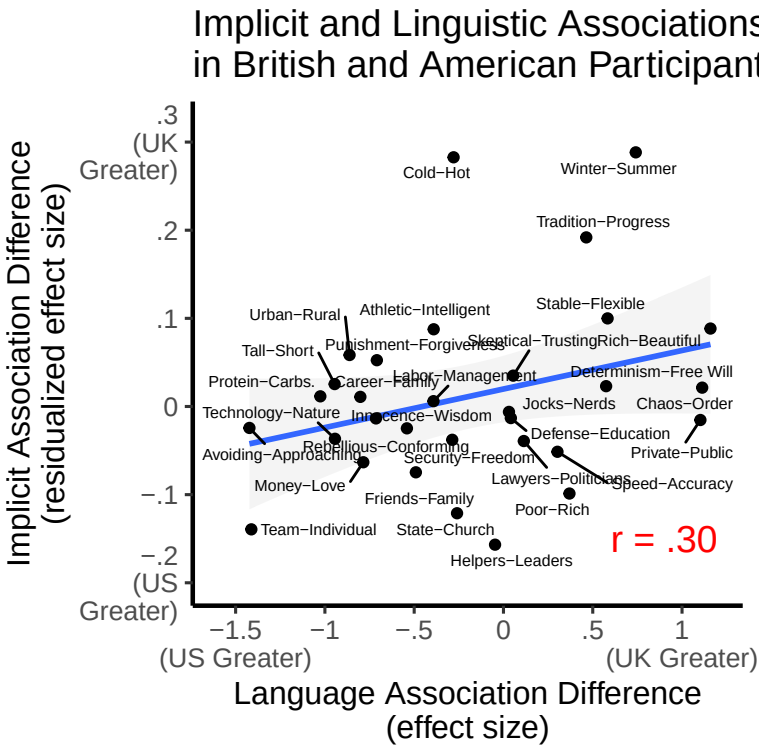

Figure 3. Difference (British - American) in implicit bias (y-axis) and linguistic bias (x-axis) for each of the 31 IAT types in our dataset.

magnitude of the bias in speaker's language predicted their behavioral bias, as measured by the IAT. These results are consistent with both the *language-as-reflection* and *language-as-causal-factor* hypotheses. In Study 2, we try to better distinguish between these hypotheses by investigating whether the gender-career bias is associated with two structural features of language: grammatical gender and the presence of gendered occupation terms (e.g., waiter/waitress).

Study 2: Gender bias and lexicalized gender

In Study 1 we examined cross-linguistic differences in gender bias without any reference to structural differences that exist in the languages in our sample. One such structural difference concerns the grammaticalization of gender. Some languages such as

Spanish mark gender distinctions in a grammatically obligatory way, e.g., "enfermero" (nurse-MASC) versus "enfermera" (nurse-FEM). Grammatical gender systems frequently demand gender-based agreement, e.g., "el enfermero alto" (the tall nurse-MASC) versus "la enfermera alta" (the tall nurse-FEM), which while informationally redundant, may act to amplify gender biases in the language. Another way languages convey gender is through gender-specific terms such as "waiter" vs. "waitress." Languages with grammatical gender do tend to use more such terms, but the two are distinct. French has grammatical gender, but many occupation terms are gender-neutral (e.g., auteur, athlète, juge).

In Study 2, we examined whether grammatical gender and use of gender-specific occupation terms are associated with a greater psychological gender bias and whether this relationship is further mediated by language statistics. Finding such associations would lend support to the language-as-causal-factor hypothesis because grammatical gender and (to a lesser degree) lexical gender encoding are relatively stable features of language. Although both can change over time, these changes are somewhat independent of the propositional content conveyed by language. For example, a Finnish document about nursing being unsuitable for men would still use a gender-neutral form of "nurse" while a Spanish document promoting nursing careers to men would be committed to using gender-marked forms.

Methods. We identified 20 occupation names that were likely to have corresponding terms in all 25 of our languages, and that were balanced in terms of their perceived gender bias in the workforce (**???**). We then translated these words into each of the 25 languages in our sample, distinguishing between male and female variants (e.g., "waiter" vs. "waitress") where present. The words were translated by consulting native speakers and dictionaries, as necessary.

We coded each language for the presence or absence of a sex-based grammatical gender system using WALS (**???**) and other sources, as necessary. To estimate the extent to which a language lexically encoded gender, we calculated the proportion of occupations within each

language for which the male and female forms differed. Larger values indicate a preponderance for more gender-specific forms in a language. Finally, we also estimated the extent to which each occupation term was gender biased in its language statistics using word embedding models trained in each language on the Subtitle and Wikipedia corpora. For each occupation form, we estimated its bias in language statistics using the same pairwise similarity metric as in Study 1a, and then averaged across occupations within a language to get a language-level estimate of gender bias. Larger values indicate greater gender bias in language statistics. We then compared each of the three language measures (grammatical gender, proportion specific gender forms, and bias in language statistics for occupation words) to the psychological gender measures described in Study 1b (implicit and explicit bias, adjusted for age, sex and block order).

Results. Speakers of languages with a grammatical gender system ($N = 12$) did not differ from those without ($N = 13$) in terms of implicit ($M = 0.01$ [-0.01, 0.03]; $t(22.99) = 0.74$, $p = 0.47$; $d = 0.29$ [-0.54, 1.13]) or explicit bias ($M = 0.08$ [-0.07, 0.23]; $t(17.67) = 1.17$, $p = 0.26$; $d = 0.48$ [-0.36, 1.32]). Languages with grammatical gender systems were more likely to have gender-specific terms for occupations ($M = 0.51$ [0.28, 0.73]; $t(14.89) = 4.85$, $p < .001$; $d = 2$ [0.98, 3.01]). Implicit gender bias was reliably correlated with degree of gender-specific marking on occupation words: Languages with more gender-specific forms tended to have speakers with greater psychological gender bias ($r(23) = 0.57$ [0.22, 0.79], $p = 0.003$), even after partialling out the effect of median country age ($r = 0.48$, $p = 0.02$; Table 1). There was no relationship between explicit psychological gender bias and lexical marking of occupation words after partialling out the effect of median country age ($r(23) = 0.11$ [-0.3, 0.48], $p = 0.609$).

We next examined whether having gender-specific forms for a particular occupation was associated with greater gender bias in the language statistics for that form. We fit a mixed effect model predicting degree of gender bias in language statistics (estimated from

word embedding models) as a function of degree of distinctiveness between male and female forms for that word, with random intercepts and slopes by language. Degree of form distinctiveness was a strong predictor of language statistics for models trained on both the Subtitle corpus ($\beta = 0.46$; $SE = 0.08$; $t = 6.08$) and Wikipedia corpus ($\beta = 0.89$; $SE = 0.1$; $t = 8.93$), with words with shared male and female forms tending to have less gender bias. This relationship also held at the level of languages: Languages with more distinct forms had a greater gender-career bias in language statistics (Subtitle: $r(17) = 0.6$ [0.2, 0.83], $p = 0.006$; Wikipedia: $r(23) = 0.77$ [0.53, 0.89], $p < .0001$; Fig. 4).

Finally, we examined the relationship between gender bias in language statistics and psychological gender biases at the level of languages. Unlike in Study 1, all the target words in the present study referred to people (occupations) and thus potentially could be marked for the gender of the referenced person. Consequently, if explicit gender marking drives language statistics, we should expect to see a strong positive relationship at the level of languages between bias in language statistics *for occupation words* and psychological biases for speakers of that language. Consistent with this prediction, gender bias in language statistics for occupation words was positively correlated with implicit gender bias (Subtitle: $r(17) = 0.49$ [0.04, 0.77], $p = 0.034$; Wikipedia: $r(23) = 0.49$ [0.11, 0.74], $p = 0.014$; Fig. 5). There was no relationship between language statistics for occupation words and explicit psychological gender bias (Subtitle: $r(17) = 0.16$ [-0.32, 0.57], $p = 0.526$; Wikipedia: $r(23) = 0.18$ [-0.23, 0.54], $p = 0.389$).

To understand the relative predictive power of language statistics and distinct forms, we fit an additive linear model predicting implicit bias from language statistics and proportion distinct forms, controlling for median country age. Because language statistics for occupation terms and proportion distinct forms were highly colinear (Wikipedia: $r(23) = 0.77$ [0.53, 0.89], $p < .0001$; $r(17) = 0.6$ [0.2, 0.83], $p = 0.006$), we used the estimate of bias in language statistics for each language based on the set IAT words described in Study 1b.

Both gender bias in language statistics (based on IAT words) and the proportion of gender-specific occupation titles were independent predictors of implicit bias. The two predictors accounted for 49% of variance in implicit bias when using the Subtitle corpus and 60% of variance for the Wikipedia corpus. Full model results are reported in the SM.

The high degree of collinearity between language statistics for occupation terms and proportion gender-specific occupations forms is consistent with a causal model in which language statistics mediate the effect of gender-specific forms on implicit bias: The presence of distinct forms referring to people of different genders *leads to* biased language statistics, which in turn leads to gender bias in behavior. Consistent with this model, a bootstrap test of mediation revealed a marginal effect for the Subtitle model (path-ab = 0.13, $p = 0.40$; **???**), and significant mediation effect for the Wikipedia model (path-ab = 0.15, $p = 0.72$).[3]

Discussion

In Study 2, we asked whether structural features of language – the presence of a grammatical gender systems and the propensity to lexicalize gender distinctions – correlated with implicit bias. Grammatical gender was not reliably correlated with implicit bias. Languages that use more gender-specific occupation terms, however did predict a greater implicit bias. There is some evidence that the effect of lexical gender distinctions on implicit bias may be mediated by the influence this terminology introduces on the ways that gender is statistically encoded in different language. What does this finding mean for our two hypotheses? The fact that, e.g., German explicitly marks the gender of professors while English does not, has cognitive consequences for German speakers; it is not simply a matter of current cultural differences being reflected in language. Language does not merely reflect our biases, it seems to contribute to them.

[3]Though our power to detect this effect is relatively low (approximately, .4; Schoemann, Boulton, & Short, 2017).

General Discussion

Where do we get our gender stereotypes? Non-linguistic experiences surely play a role, but might we also be learning our biases from the statistics of language to which we are exposed? We used a large-scale dataset of Implicit Association Tests (IATs) measuring the bias to associate men with career and women with family and related people's measured implicit bias to the statistics of the dominant language spoken in the country of the participants. In Study 1, we found that languages with a greater gender bias in their distributional structure, tend to have speakers that have stronger implicit biases. In Study 2, we found a positive relationship between a structural language feature – the prevalence of gender-marked occupation terms – and implicit bias. There is suggestive evidence that this greater implicit bias is mediated by the greater gender bias encoded in the distributional patterns of gender-marked terms.

Our work is the first to characterize the relationship between broad structural patterns in language and cultural stereotypes. The positive correlation that we find between gender bias in language and gender bias in speakers is consistent both with language playing a causal role in the emergence of cultural stereotypes and the idea that language merely reflects existing stereotypes of its speakers. The positive correlation we find in Study 2 between prevalence of gender-specific terms and implicit bias is most parsimoniously explained by the language-as-causal-factor hypothesis because it is unlikely that language forms change on a timescale that could directly reflect behavior. The two causal forces are not mutually exclusive, and in fact may amplify the effects of each other. Future work could use experimental methods to manipulate language statistics in order to more directly examine these causal influences.

A central contribution of the current work is that it sheds light onto the potential origins of psychological biases that exist at the level of the individual. Given observed

large-scale stuctural patterns of gender inequality, such as differences in STEM participation, researchers from a range of fields have sought to understand the individual-level causal processes that led to the emergence of structural inequality. But, critically, these previous efforts have taken properties of the indidividual – such as feelings of self-efficacy in science (**???**) and general preferences (**???**) – as largely exogeneous. Here, we provide a potential explanation of the origins of these psychological biases by arguing that exposure to biased language statistics could play a causal role in the emergence of these biases at the level of the invididual. Consistent with this account, biases in language statistics are correlated with previous individual-level predictors of STEM inequality, such as self-efficacy in science and general preferences (see SM for details).

One limitation of our work is the reliance on the IAT, which has been criticized for both its low reliability (**???**) and limited external validity (**???**). Issues of reliability are less relevant here because we use the IAT to measure group-level differences rather than as an individual-difference measure. However, concerns about validity are important particularly because we find that language measures and explicit psychological measures of gender bias are uncorrelated, though explicit bias was measured in a fairly coarse way. Understanding the full import of linguistic biases on cultural stereotypes would therefore require obtaining measures more closely related to real-world behavior.

Cultural stereotypes are acquired through experience. Here, we show that group-level differences in implicit bias are strongly correlated with the strength of gender bias encoded in the statistics of different languages. This pattern suggests that the statistics of language use are an important source of cultural experience: The mere process of listening to and producing language exposes one to statistics that may lead to the formation of cultural stereotypes. Many cultural associations present in the statistics of language may be innocuous – indeed, these statistics may be an important mechanism through which cultural information is transmitted (**???**). But, in other cases, like the kind of gender stereotypes

investigated here, language may play a powerful role in their formation, and ultimately contribute to undesirable real-world consequences like gender inequality in STEM. Understanding the causal role that language plays in the formation of these stereotypes is therefore an important first step to changing these consequences.

References