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

Molly Lewis[a,b,1] **and Gary Lupyan**[b,1,2]

[a]University of Chicago; [b]University of Wisconsin-Madison

This manuscript was compiled on May 28, 2019

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 science, technology, engineering and mathematics (STEM) fields 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.

distributional semantics | cultural stereotypes | IAT | gender

By the time they are two, children have begun to acquire the gender stereotypes in their culture (1). 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" (2). Such beliefs may, over time, translate to the observed lower rates of female participation in science, technology, engineering and mathematics (STEM) fields (3–6) and are reflected in large differences in perceived self-efficacy; boys reported having greater ability to understand and explain various scientific findings (independent of actual ability, 6). Here we attempt to understand where such beliefs may come from.

We can distinguish between two major sources of information that contribute to gender stereotypes. 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 (7). 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 (8), or merely told that the game is associated with a particular gender (9). In some cases, a subtle turn of phrase can influence children's gender-based generalization (10, 11). 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 to 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, 12). 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 (13, N = 657,335). These data span native speakers of 25 languages allowing us to assess how implicit gender biases vary as a function of language to which participants are exposed.

To measure cultural stereotypes in language, we use semantic embeddings derived from a distributional semantics model that tries to predict a word from surrounding words

Significance Statement

Where do stereotypes come from? We asked whether language provide implicit information about gender stereotypes through the statistics of word occurrences (e.g., how often "nurse" and "she" co-occur). These linguistic biases may in turn shape speakers' psychological biases and ultimately contribute to structural gender inequalities. We compared speakers of 25 languages and measured both their psychological gender biases and the gender bias in the language they spoke. We found that speakers of more gender-biased languages tended to have greater psychological gender bias suggesting that language statistics could lead to the emergence of individuals' gender stereotypes.

Both authors designed research and wrote the paper. M.L. performed research and analyzed the data.

The authors declare no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: mollyllewis@gmail.com

in a large corpus. The core assumption of these models is that the meaning of a word can be described by the words it co-occurs with—words occurring in similar contexts tend to have similar meanings (14). A word like "dog," for example is represented as more similar to "cat" than to "banana" because contexts containing "dog" are more similar to contexts containing "cat" than to contexts containing "banana" (15–17). Gender stereotypes can become encoded in the distributional semantics of language because a word like "woman" may occur in more similar contexts to words like "home" and "family" while a word like "man" in contexts more similar to "job" and "money." Previous work has shown biases studied using IATs can be predicted from the distributional statistics of language (word co-occurrences, 18). This previous work only measured semantic biases in English. Here, we examine gender biases in the distributional semantics of 25 languages and ask whether languages with a stronger gender bias predict stronger implicit and explicit gender biases in speakers of those languages.

Discovering that gender biases in language are correlated with people's implicit and explicit gender 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 gender biases derived from the distributional structure of different languages predict responses on the IAT. 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.

A cross cultural dataset assessing gender biases

To quantify gender biases, we used data from a large-scale administration of an Implicit Association Task (IAT, 12) by Project Implicit (13, 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 are easier to pair together compared to words denoting 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, 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). Our final sample included 657,335 participants from 39 countries, with a median of 1,145 participants per country.

To quantify the strength of participants' implicit bias as assessed by 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 (12). 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). An explicit gender-career bias score was defined as their Career response minus their Family response such that greater values indicate a greater bias to associate males with career.

At the participant level, implicit bias scores were positively correlated with participant age ($r = 0.06$, $p < .0001$). Male participants ($M = 0.32$, $SD = 0.37$) had a significantly smaller implicit gender bias than female participants ($M = 0.41$, $SD = 0.35$; $t = 96.82$, $p < .0001$), a pattern consistent with previous findings (13). Implicit bias scores were larger for participants that received the block of trials with bias-incongruent mappings first relative to the opposite order (t = -104.03, $p < .0001$).

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 ($r = 0.16$, $p < .0001$) but not countries ($r = 0.26$, $p = 0.12$).

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 (5, 6). Consistent with previous research (Miller et al., 2015), 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 = -0.54, $p < .001$). In contrast, there was no relationship between the percentage of women in STEM fields and the explicit gender bias measure used by Project Implicit ($r = 0.14$, $p = 0.43$). In addition, we found a strong correlation between the median age of each country's population (19) and the residual implicit bias (in which participant age was held constant): Countries with older populations tended to have larger gender biases ($r = 0.64$, $p < .0001$).

In sum, we replicate previously-reported patterns of gender bias in the gender-career IAT literature, with roughly comparable effect sizes (c.f. 13). The weak correlation between implicit and explicit measures is consistent with claims that these two measures tap into different cognitive constructs (20). In addition, we find that implicit gender bias predicts an objective measure of gender equality—female enrollment in STEM fields. 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 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 judgments of word genderness (Study 1a). We then apply this method to models trained on text in other languages (Study 1b and 1c). 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.

In Studies 1a-1c we estimate linguistic gender biases using distributional semantics. By attempting to predict the words that surround another word in large corpora, these models (e.g., 21) are 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). We estimated a gender score 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"; 13) 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 female similarity - mean male similarity), such that larger values indicated a stronger association with females. We estimated gender scores for each word from models pre-trained on two different corpora of English text: subtitles from movies and TV shows (22, 23) and Wikipedia (24).

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 = 0.71; $p <$ .0001). Critically, bias estimates from both word embedding models were also highly correlated with human judgements (M = 4.10; SD = 0.92; r_{Subtitle} = 0.63; $p <$ .0001; $r_{\text{Wikipedia}}$ = 0.59; $p <$ .0001; Fig. 1). This suggests that the psychological gender bias of a word can be reasonably estimated from word embeddings.

Having validated our basic method, we now use it to examine the relationship between psychological and linguistic gender biases. In Study 1b, we estimated 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.

Despite the differences in the specific content conveyed by the Wikipedia and the Subtitles corpus, the estimated gender bias for each language was similar across the two corpora

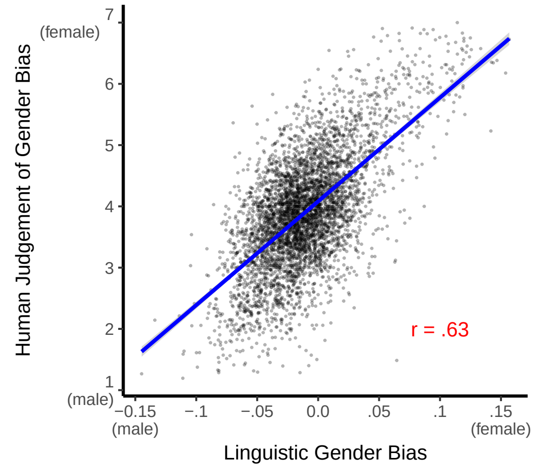

Fig. 1. Human judgments of word gender bias as a function of gender bias from the Subtitle-trained embedding model (Study 1a). Each point corresponds to a word. Larger numbers indicate stronger association with females (note that this differs from the design of the rating task, but is changed here for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

($t(19)$ = -0.06, p = 0.95). 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 = 0.5, p = 0.02) and Wikipedia trained models (r = 0.48, p = 0.01; Fig. 2a; 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 partialling out the effect of median country age (r_{Subtitle} = 0.42, p = 0.04; $r_{\text{Wikipedia}}$ = 0.43, p = 0.04). Linguistic gender bias was not correlated with explicit gender bias (r_{Subtitle} = -0.08, p = 0.74; $r_{\text{Wikipedia}}$ = 0.34, p = 0.09). Estimates of language bias from the Subtitle corpus were correlated with the objective measure of gender equality, percentage of women in STEM fields (r = -0.55, p = 0.02); this relationship was not reliable for the Wikipedia corpus (r = -0.19, p = 0.4).

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. We leveraged the Attitudes, Identities, and Individual Differences Study dataset (AIID, 25) containing measures of IAT performance from over 200,000 participants for a wide range of IATs (e.g. career - family, team - individual, etc.). All the tests were conducted using English words and most participants were English speakers. The dataset allowed us to compare biases between participants who spoke two different dialects of English: British and American. For each of the 31 IATs 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.

Figure 2b visualizes the critical interaction term. Behavioral performance on the different IATs was correlated with language statistics. When language statistics predicted that US-English had a greater bias, American participants showed a greater bias. When language statistics predicted that UK-

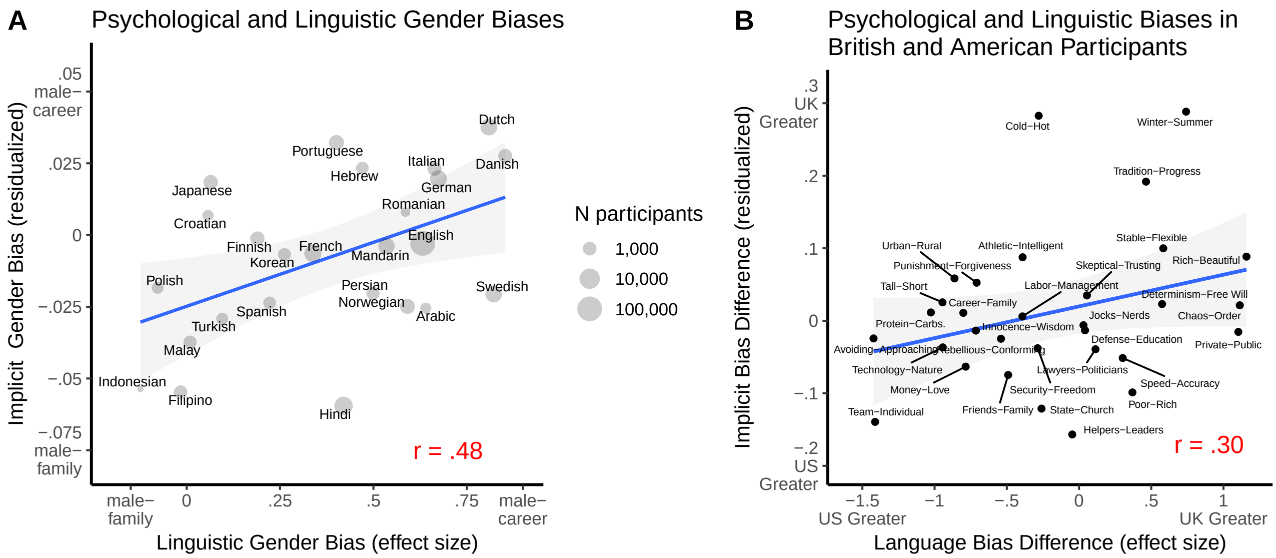

Fig. 2. (A) Implicit gender bias (adjusted for participants' age, gender, and congruent/incongruent block order) as a function of the linguistic gender bias derived from word-embeddings (Study 1b). Each point corresponds to a language. The size of the point is proportional to the number of participants who come from the country in which the language is dominant. Linguistic biases are estimated from models trained on text in each language from the Wikipedia corpus. 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. (B) Difference (UK minus US) in implicit bias versus linguistic bias for 31 IAT types (Study 1c). Error bands indicate standard error of the linear model estimate.

English had a greater bias, British participants showed a greater bias (β = -.05, SE = .02, t = -2.88; see SI Sec. 3.4 for full model results).

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 biases beyond associating males with career and women with family: In a comparison of 31 different IATs, the magnitude of the bias in speaker's dialect of English (US vs. UK) 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).

	Residualized Explicit Bias	Residualized Implicit Bias (IAT)	Percent Women in STEM	Language IAT (Subtitle)	Language IAT (Wikipedia)	Gendered Occupation Labels	Occupation Bias (Subtitle)	Occupation Bias (Wikipedia)
Residualized Explicit Bias		.28	.16	-.06	.38+	.14	.33	.34
Residualized Implicit Bias (IAT)	.18		-.38+	.42*	.43*	.48*	.57**	.52**
Percent Women in STEM	.18	-.53*		-.49*	-.09	-.23	-.28	-.20
Language IAT (Subtitle)	-.08	.50*	-.55*		.47*	.20	.35+	.33
Language IAT (Wikipedia)	.34+	.48*	-.19	.51*		.11	.21	.39+
Gendered Occupation Labels	.11	.57**	-.35	.28	.18		.71**	.66**
Occupation Bias (Subtitle)	.28	.64**	-.39	.42+	.28	.75**		.77**
Occupation Bias (Wikipedia)	.29	.59**	-.32	.40+	.44*	.70**	.80**	
Median Country Age	-.07	.61**	-.42+	.31	.25	.35+	.36	.34+

Table 1. Correlation (Pearson's *r*) for all measures in Study 1b and 2 at the level of languages. Bottom triangle shows simple correlations; top triangle shows partial correlations controlling for median country age. * denotes $p < .05$; ** denotes $p < .01$; + denotes $p < .1$.

Study 2: Gender bias and lexicalized gender

The association between language bias and implicit bias is predicted by both the language-as-reflection and language-as-causal-factor hypotheses, but for different reasons. If language is causally related to implicit biases, then differences in the structural aspects of language that act to exaggerate linguistic gender bias should predict greater implicit bias. This relationship is not predicted by the language-as-reflection hypothesis.

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 structural difference is the existence of gender-specific terms such as "waiter" vs. "waitress," which are more frequent in some languages than others. 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,

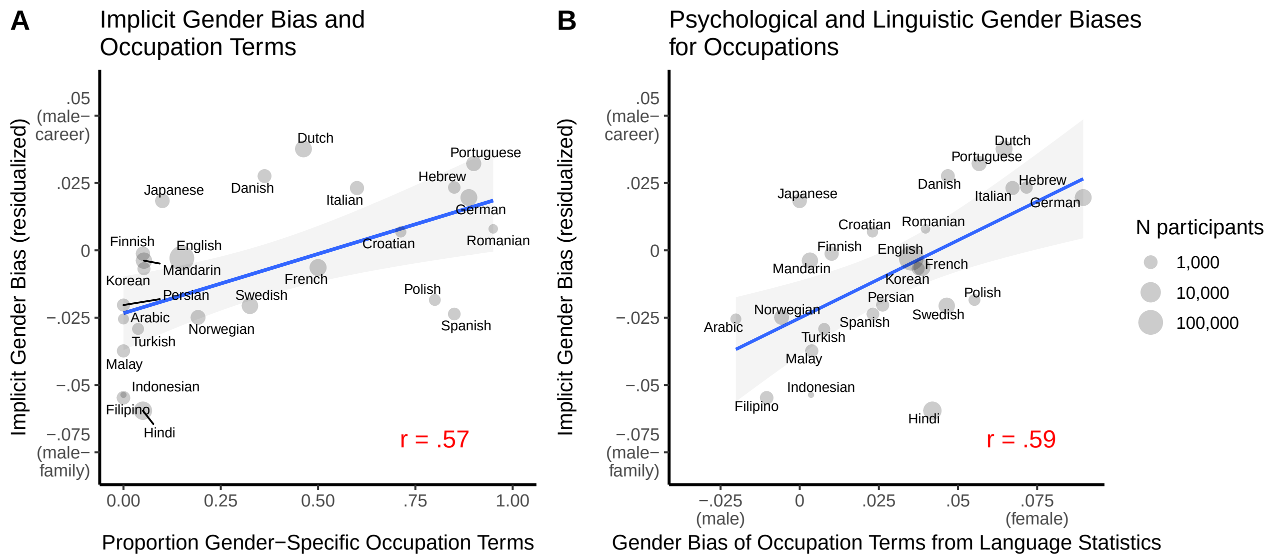

Fig. 3. Implicit gender bias (adjusted for participant age, gender, and block order) as a function of the proportion of gender-specific labels for set of words referring to occupations (A, left) and mean gender bias of words referring to occupations (B, right). Each point corresponds to a language, with the size of the point corresponds the number of participants speaking that language. Occupation gender bias is estimated from language statistics (B) from word embedding models trained on the Wikipedia corpus. Error bands indicate standard error of the linear model estimate.

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 somewhat lesser degree) lexical gender encoding are relatively stable features of language. Although both can change over time, these changes are largely 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.

In additive linear models controlling for median country age, there was no difference in implicit or explicit psychological gender bias for speakers of languages with a grammatical gender system (N = 12), compared to those without (N = 13; Implicit: $\beta = 0$; $SE = 0.01$; t = -0.43; Explicit: β = -0.09; SE = 0.07; t = -1.23). 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 implicit gender bias ($r = 0.57$, $p <$ .01; Fig. 3a). This relationship remained 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 ($r = 0.14$, p = 0.51).

We next examined whether the existence of gender-specific occupation terms was associated with a greater encoding of gender bias in the distributional statistics of the language. We fit a mixed effects model predicting degree of gender bias in language statistics (estimated from word embedding models) from distinctiveness between male and female forms for that word, with random intercepts and slopes by language. Having more distinct occupation terms was associated with greater linguistic gender bias for those occupations. This was true for models trained on both the Subtitle corpus ($\beta = 0.59$; SE = 0.07; $t = 8.72$) and Wikipedia corpus ($\beta = 0.81$; $SE = 0.09$; t = 9.48). For example, "secretary" had greater gender bias in Italian, which has distinct male and female terms, compared to English, which has only one term.

This relationship also held at the level of languages: Languages with more distinct forms had a greater bias in language statistics ($r_{\text{Subtitle}} = 0.75$, $p < .01$; $r_{\text{Wikipedia}} = 0.7$, $p <$ .01; i.e., Italian has greater overall gender bias compared to English).

Finally, we examined the relationship between gender bias in language statistics for occupation words 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 ($r_{\text{Subtitle}} = 0.64$, $p < .01$; $r_{\text{Wikipedia}} = 0.59$, $p < .01$), and remained reliable after partialling out the effect of median country age ($r_{\text{Subtitle}} = 0.57$, $p < .01$; $r_{\text{Wikipedia}} = 0.52$, p = 0.01; Fig. 3b). In contrast, explicit psychological gender bias was not predicted by language statistics, even after partialling out the effect of median country age ($r_{\text{Subtitle}} = 0.33$, p = 0.12; $r_{\text{Wikipedia}} = 0.34$, $p = 0.11$).

To understand the relative predictive power of language statistics and distinct occupation terms, 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 ($r_{\text{Subtitle}} = 0.75$, $p < .001$; $r_{\text{Wikipedia}} = 0.70$, $p < .001$), 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 SI (Sec. 4.4).

The strong 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 mediation effect for the Wikipedia model (path-ab = 0.35, $p = 0.04$), and a marginal effect for the Subtitle model (path-ab = 0.28, $p = 0.10$).*

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.

Discussion

Our work is the first to characterize the relationship between cultural stereotypes and cross-linguistic differences in language statistics. The positive correlation 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. However, the positive association between prevalence of gender-specific terms and implicit bias (Study 2) is most parsimoniously explained by language statistics (partially) causing the observed differences in implicit bias. It is implausible that nonlinguistically acquired gender biases could have changed the lexical inventory of the language rapidly enough to explain the differences in IAT performance that we observed. Future work could use experimental methods to manipulate language statistics as a way to more directly examine these causal influences.

Where do we get our gender stereotypes? Non-linguistic experiences surely play a role, but might we also be learning our biases from the language to which we are exposed? We used a large-scale dataset of Implicit Association Tests (IATs) that measure people's implicit bias to associate men with career and women with family. We related these biases to the linguistic gender biases computed from patterns of word co-occurrences in the dominant language spoken in the country of the participants. In Study 1, we found that languages with a greater gender biases embedded 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 findings join several recent attempt to understand large-scale correlates of gender stereotypes (6) and differences in gender preferences more broadly (27). These earlier reports have argued that increases in institutional gender equality (which are strongly associated with increases in national GDP) allow greater personal freedom, unmasking inherent gender differences and explaining why greater institutional equality is associated with a *lower* female STEM participation (6) and larger stereotypical gender differences (e.g., women being more risk averse and less patient than men; 27). Although our results do not contradict this possibility, they suggest that biases learned from language may be a part of the fuller picture. The encoding of gender stereotypes in different languages is itself correlated with GDP (larger GDP correlates with a stronger linguistic bias, $r = .43$; $p < .01$) and also with previously reported individual-level predictors of STEM inequality such as self-efficacy in science ($r = .48$; $p = .01$) and general gender preferences ($r = .59$; $p < .001$; see SI Sec. 5). Determining the causal pathways requires additional work.

One limitation of our work is its reliance on the IAT, which has been criticized for both its low reliability (28) and limited external validity (29). 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. The strong negative correlation we find between the proportion women in STEM and language bias (r = -.55) provides compelling evidence that language biases are related to real-world consequences. However, understanding the full import of linguistic biases on cultural stereotypes will 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 (30). 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 structural inequality. Understanding the causal role that language plays in the

*Though our power to detect this effect is relatively low, approximately, .4 (ref. 26).

formation of these stereotypes is therefore an important first step to changing these consequences.

Materials and Methods

All data and code are available online (https://github.com/mllewis/IATLANG). Supplementary Information available at: https://mollylewis.shinyapps.io/iatlang_SI/.

Description of IAT dataset. 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 (13, 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 SI Sec. 1.1). 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 make our analyses conservative, lowering the likelihood of finding language-specific predictors of the kind we report here.

Country-level estimates of female STEM participation was calculated from 2012 to 2017 data; these data were available for 33 out of 39 of the countries in our sample.

Study 1a. 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, 31). Both models were trained using the fastText algorithm (32, a variant of word2vec). There were 4,671 words in total that overlapped between the word-embedding models and human ratings.

Study 1b. We identified the most frequently spoken language in each country in our analysis using Ethnologue (33). After exclusions (see below), our final sample included 25 languages (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). For each language, we obtained translations from native speakers for the stimuli in the Project Implicit gender-career IAT behavioral task (13) 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;" 13). 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 SI Sec. 2.1 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 (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.

Study 1c. The AIID datset was partitioned 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) and were given access to the confirmatory dataset only after our pre-registration was approved.

Of the 95 IATs present in the dataset, we selected 31 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 our US and UK English corpora. To measure linguistic bias, we trained word embedding models on equally-sized subsets of British National Corpus (BNC; 34) and Corpus of Contemporary American English (COCA; 35). The model was trained using the fastText alogrithm (32), with a vector size of 400 and window size of 10. 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. After data exclusion (using criteria similiar to Study 1a; see SI Sec. 3.2 for details), our final sample included data from 135,240 administrations of the IAT across the 31 IATs (USA: N = 127,630; UK: N = 7,610). Each participant completed an average of 6.13 different IATs (SD = 3.99). For each administrations of an IAT, we calculated a residual D-score which controlled for participant gender, age, education, task order (whether implicit or explicit measures were completed first), and block order (whether congruent or incongruent mappings occurred first).

We fit a linear mixed effect model predicting the magnitude of the IAT bias for each participant from their location (US vs. UK), the linguistic bias from US-English and UK-English trained models, and the interaction of the two factors. We included participant and IAT test as random intercepts. We fit this and subsequent mixed effect models with the *lme4* R package (36). 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 SI Sec. 3.3 for results of the exact pre-registered model).

Study 2. We identified 20 occupation terms that can be translated into all 25 of our languages, and that were balanced in terms of their perceived gender bias in the workforce (37). 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.

We coded each language for the presence or absence of a sex-based grammatical gender system using WALS (7) and other sources, as necessary. We quantified lexical encoding of gender as the proportion of the 20 occupations within each language for which the male and female forms differed. Larger values indicate a preponderance for more gender-specific forms. Languages with grammatical gender systems were more likely to have gender-specific terms for occupations ($t(14.89) = 4.85$, $p < .001$). We then 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 term, 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, gender and block order). The reported mediation analysis was conducted using the *robmed* R package (38).

1. Gelman SA, Taylor MG, Nguyen SP, Leaper C, Bigler RS (2004) Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development* pp. i–142.
2. Bian L, Leslie SJ, Cimpian A (2017) Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* 355(6323):389–391.
3. Ceci SJ, Williams WM (2011) Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences* pp. 3157–3162.
4. Leslie SJ, Cimpian A, Meyer M, Freeland E (2015) Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347(6219):262–265.
5. Miller DI, Eagly AH, Linn MC (2015) Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology* 107(3):631.
6. Stoet G, Geary DC (2018) The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science* 29(4):581–593.
7. Dryer MS, Haspelmath M, eds. (2013) *WALS Online*. (Max Planck Institute for Evolutionary Anthropology, Leipzig).
8. Rhodes M, Brickman D (2008) Preschoolers' responses to social comparisons involving relative failure. *Psychological Science* 19(10):968–972.
9. Cimpian A, Mu Y, Erickson LC (2012) Who is good at this game? linking an activity to a social category undermines children's achievement. *Psychological Science* 23(5):533–541.
10. Cimpian A, Markman EM (2011) The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development* 82(2):471–492.
11. Rhodes M, Leslie SJ, Yee KM, Saunders K (2019) Subtle linguistic cues increase girls' engagement in science. *Psychological Science*.
12. Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the Implicit Association Test. *Journal of Personality and Social Psychology* 74(6):1464.
13. Nosek BA, Banaji MR, Greenwald AG (2002) Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6(1):101.
14. Firth J (1957) A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, Philological Society.
15. Landauer TK, Dumais ST (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211.
16. Lund K, Burgess C (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2):203–208.
17. Lenci A (2008) Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics* 20(1):1–31.
18. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.
19. CIA (2017) The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/index.html.
20. Forscher PS, et al. (2016) A meta-analysis of change in implicit bias.
21. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space.
22. Van Paridon J, Thompson B (in prep.) Sub2vec: Word embeddings from OpenSubtitles in 62 languages.
23. Lison P, Tiedemann J (2016) Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles in *Proceedings of the 10th International Conference on Language Resources and Evaluation*.
24. Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information.
25. Hussey I, et al. (2019) Attitudes, Identities, and Individual differences (AIID) Study. https://doi.org/https://osf.io/pcjwf/.
26. Schoemann AM, Boulton AJ, Short SD (2017) Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science* 8(4):379–386.
27. Falk A, Hermle J (2018) Relationship of gender differences in preferences to economic development and gender equality. *Science* 362(6412):eaas9899.
28. Lane KA, Banaji MR, Nosek BA, Greenwald AG (2007) Understanding and using the Implicit Association Test: Iv. *Implicit measures of attitudes* pp. 59–102.
29. Fazio RH, Olson MA (2003) Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology* 54(1):297–327.
30. Lupyan G, Lewis M (2017) From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience* pp. 1–19.
31. Scott GG, Keitel A, Becirspahic M, Yao B, Sereno SC (2018) The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods* pp. 1–13.
32. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
33. Simons GF, Charles DF, eds. (2018) Ethnologue: Languages of the world.
34. Burnard L (1995) *Users reference guide for the British National Corpus*. (Oxford University Computing Services).
35. Davies M (2008) The corpus of contemporary American English (COCA). https://corpus.byu.edu/coca/.
36. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1–48.
37. Misersky J, et al. (2014) Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods* 46(3):841–871.
38. Alfons A (2018) *robmed: (Robust) Mediation Analysis*. R package version 0.3.0.