nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0918-6

Check for updates

Gender stereotypes are reflected in the distributional structure of 25 languages

Molly Lewis [1,2] ✉ and Gary Lupyan [3]

Cultural stereotypes such as the idea that men are more suited for paid work and women are more suited for taking care of the home and family, may contribute to gender imbalances in science, technology, engineering and mathematics (STEM) fields, among other undesirable gender disparities. Might these stereotypes be learned from language? Here we examine whether gender stereotypes are reflected in the large-scale distributional structure of natural language semantics. We measure gender associations embedded in the statistics of 25 languages and relate these to data on an international dataset of psychological gender associations ($N = 656{,}636$). People's implicit gender associations are strongly predicted by gender associations encoded in the statistics of the language they speak. These associations are further related to the extent that languages mark gender in occupation terms (for example, 'waiter'/'waitress'). Our pattern of findings is consistent with the possibility that linguistic associations shape people's implicit judgements.

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, six-year-old girls were less likely than boys to choose activities that were described as being 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 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 are better suited 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.' versus 'Miss.'), proper names ('Sam' versus 'Ashley'), pronouns ('he' versus 'she'), certain job titles ('waiter' versus '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 ('enfermera' versus 'enfermero').

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 if they are 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[10] ("[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. In this study, we investigate whether a similar correspondence between language associations and stereotypes exists in a large corpus of naturalistic text and among an international sample of participants.

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 tendency to associate men with careers and women with family ($N = 657{,}335$)[13]. These data span 39 countries, allowing us to assess how group-level implicit gender associations[14–16] 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 is trained by predicting words from surrounding words as they occur 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[17]. The word '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'[18–20]. 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' may occur in contexts more similar to 'job' and 'money.' Previous work has shown stereotypes such as those studied using IATs can be predicted from the distributional statistics of language (co-occurrences)[21–24]. This previous work only measured semantic associations in English. Here, we examine gender associations in the distributional semantics of 25 languages and ask whether languages

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Psychology Department, University of Wisconsin-Madison, Madison, WI, USA. ✉e-mail: mollylewis@cmu.edu

with stronger career–gender associations predict stronger implicit and explicit gender associations in speakers of those languages.

Discovering that gender associations in language are correlated with people's implicit and explicit gender associations can be interpreted in several ways[22,25]. One possibility is that some cultures have stronger gender stereotypes and these are reflected in what people talk about. Language, in this view, simply reflects pre-existing associations. However, language may not only reflect pre-existing stereotypes, but may also provide a distinct source of information for learning about them, thereby constituting a causal influence on the associations that people learn[26]. Another possibility is that a third variable influences both language and psychological associations. The correlational approach of the present work does not allow us to distinguish between these possibilities; our goal is to establish whether there is in fact a correspondence between psychological and linguistic gender associations. Establishing whether such a correspondence exists is a prerequisite to understanding the underlying causal model.

In study 1, we examine whether gender associations derived from the distributional structure of different languages predict responses on the IAT. In study 2, we examine how the psychological associations measured by the IAT and the linguistic associations we measure relate to more structural aspects of language: sex-based grammatical gender and the prevalence of gender-specific occupation terms (for example, 'waiter'/'waitress', but 'teacher'/'teacher'). Our results suggest that languages that encode gender stereotypes more strongly—through either distributional semantics or structural features—tend to have speakers with stronger stereotypical gender associations.

Results

Study 1: relating associations in distributional semantics and human behaviour. To quantify gender associations, we used data from a large-scale administration of an IAT[12] by Project Implicit[13]. The IAT measures the strength of respondents' implicit associations between two pairs of concepts (for example, male–career/female–family versus male–family/female–career) accessed via words (for example, 'man' and '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 stereotype-congruent (that is, Key A = male/career; Key B = female/family) or stereotype-incongruent (that is, 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 Methods, 'Study 1b' for a list of target words). Slower reaction times in the stereotype-incongruent blocks relative to the stereotype-congruent blocks are interpreted as indicating an implicit association between the corresponding concepts (that is, a tendency 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 (Extended Data Fig. 1).

To quantify the strength of participants' implicit association 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[27]. 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 association score was defined as their Career response minus their Family response such that greater values indicate a greater tendency to associate males with career.

Replicating previous analyses[13], participants tended to implicitly associate men with career and women with family (*D*-score mean $(M) = 0.38$ [0.38, 0.38]; $t(657{,}334) = 878.3$, $P < 0.001$). Older participants showed greater implicit associations between women–family and men–career ($r(657{,}333) = 0.06$ [0.06, 0.06], $P < 0.001$). The measured associations were stronger for female participants ($M = 0.41$, s.d. = 0.35) than male participants ($M = 0.32$, s.d. = 0.37; $t(338{,}217.04) = 96.82$, $P < 0.001$; $d = 0.27$ [0.26, 0.27]) and were larger for participants that received the block of trials with stereotype-incongruent mappings first than those who received the stereotype-incongruent mappings second ($M = -0.09$ [−0.09, −0.09]; $t(652{,}694.18) = -104.03$, $P < 0.001$; $d = -0.26$ [−0.26, −0.25]; Extended Data Fig. 2).

Because we did not have language information at the participant level, in the remaining analyses we examine the career–gender association and its predictors at the country level. To account for the influences on implicit associations mentioned above, we calculated a residual implicit-association score for each participant, controlling for participant age, participant gender and block order. We also calculated a residual explicit association score controlling for the same set of variables. We then averaged across participants to estimate the country-level gender association (implicit: $M = -0.01$; s.d. = 0.03; explicit: $M = 0.00$; s.d. = 0.18; Extended Data Fig. 3). Implicit gender associations were correlated with explicit gender associations at the level of participants ($r(645{,}072) = 0.16$ [0.16, 0.16], $P < 0.001$); at the level of countries, this relationship was stronger, but not statistically reliable ($r(37) = 0.26$ [−0.07, 0.53], $P = 0.12$). The weak correlation between implicit and explicit measures is consistent with claims that these two measures tap into different cognitive constructs[28].

Do the implicit and explicit associations measured by the Project Implicit dataset predict any real-world outcomes? We compared our residual country-level implicit and explicit gender associations 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, we found that implicit gender association was negatively correlated with percentage of women in STEM fields: countries with weaker associations between men and career tended to have more women in STEM fields ($r(31) = -0.54$ [−0.75, −0.24], $P = 0.001$). By contrast, there was no relationship between the percentage of women in STEM fields and the explicit gender-association measure used by Project Implicit ($r(31) = 0.14$ [−0.21, 0.46], $P = 0.43$). In addition, we found a strong correlation between the median age of each country's population[29] and the residual implicit association (in which participant age was held constant): Countries with older populations tended to have larger gender associations ($r(37) = 0.64$ [0.4, 0.79], $P < 0.001$).

In sum, we replicate previously reported patterns of gender association in the gender–career IAT literature, with roughly comparable effect sizes[13]. We also find that implicit gender associations predict an objective measure of gender equality—female enrollment in STEM fields. In the Discussion, we comment further on our findings that older participants and participants from countries with older populations show stronger implicit gender associations.

Are participants' gender associations predictable from the language they speak? Showing that such a relationship exists is the first step towards investigating the underlying causal relationships. In study 1, we estimate linguistic gender associations using distributional semantics. By attempting to predict the words that surround another word in large corpora, word-embedding models are able to learn a vector-based representation for each word that represents its similarity to other words; that is, a semantic embedding (for example, ref. [30]). We can then compute the similarity between two words by taking the distance between their vectors (for example, cosine of angle). We begin by validating word-embedding measures of gender association by comparing them with explicit human judgements

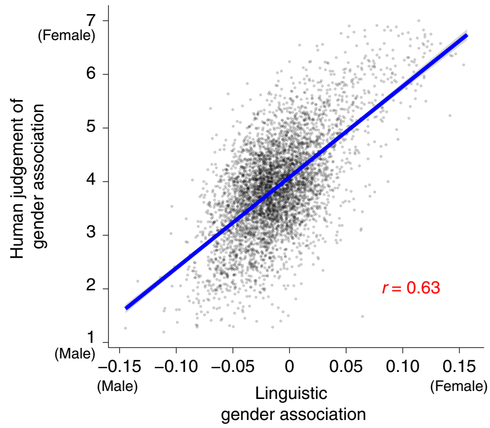


Fig. 1 | Human judgements of word gender association as a function of gender association from the subtitle-trained embedding model. Each point corresponds to a word. Larger numbers indicate stronger association with females (note that this differs from the design of the rating task, but is changed here for consistency with other plots). Blue line shows linear fit and the error band indicates standard error of the linear fit. Study 1a: $r(4{,}669) = 0.63$ [0.61, 0.65]; $P < 0.001$; $n = 4{,}671$.

of word genderness (study 1a). We then apply this method to models trained on text in other languages (studies 1b and 1c). We find that the implicit gender association of participants in a country is correlated with the gender associations embedded in the statistics of the dominant language spoken in that country.

In study 1a, 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'). We then obtained a gender score for each word 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[31,32], and Wikipedia[33].

Estimates of gender association from the subtitle corpus ($M = 0.01$; s.d. $= 0.03$) and the Wikipedia corpus ($M = 0$; s.d. $= 0.03$) were highly correlated with each other ($r(4{,}669) = 0.71$ [0.70, 0.73], $P < 0.001$). Critically, association estimates from both word-embedding models were also highly correlated with human judgements of word gender (the degree to which a word is associated with females versus males; $M = 4.10$; s.d. $= 0.92$; subtitle: $r(4{,}669) = 0.63$ [0.61, 0.65], $P < 0.001$; Wikipedia: $r(4{,}669) = 0.59$ [0.57, 0.60], $P < 0.001$; Fig. 1). This suggests that the psychological gender association 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 associations of men with career and women with family. In study 1b, we estimated the magnitude of these associations in the dominant language spoken in each country represented in the Project Implicit dataset, and compare this estimate with estimates of psychological career–gender associations from the Project Implicit participants.

Despite the differences in the specific content conveyed by the Wikipedia and the subtitle corpora, the estimated career–gender association for each language was similar across the two corpora (mean difference (M_{diff}) $= 0$ [−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 for each language and the mean career–gender association score for participants from countries where that language was dominant (and, we assume, was the native language of most of these individuals). Implicit career–gender association was positively correlated with estimates of career–gender association in language from both the subtitle- ($r(18) = 0.5$ [0.08, 0.77], $P = 0.02$) and Wikipedia-trained models ($r(23) = 0.48$ [0.11, 0.74], $P = 0.01$; Fig. 2a; Table 1 shows the language-level correlations between all variables in studies 1b and 2; Extended Data Figs. 4–6). Linguistic career–gender association was not correlated with explicit career–gender association (subtitle: $r(18) = -0.08$ [−0.5, 0.38], $P = 0.74$; Wikipedia: $r(23) = 0.34$ [−0.06, 0.65], $P = 0.09$). Estimates of the career–gender association from the subtitle corpus were correlated with the objective measure of gender equality and percentage of women in STEM fields ($r(16) = -0.55$ [−0.81, −0.11], $P = 0.02$). This relationship was not reliable for the Wikipedia corpus ($r(20) = -0.19$ [−0.57, 0.25], $p = 0.4$).

In study 1c, we conducted a confirmatory, pre-registered analysis of our hypothesis that associations present in language statistics are reflected in the psychological associations of speakers of those languages. We leveraged the Attitudes, Identities and Individual Differences Study dataset (AIID)[34] containing measures of IAT performance from more than 200,000 participants for a wide range of IATs (including, for example, career–family and team–individual). All the tests were conducted using English words and most participants were English speakers. The dataset allowed us to compare associations between participants who spoke two different dialects of English: British and American English. For each of the 31 IATs in the sample, we predicted that the degree to which those associations were present in a speaker's English dialect (British or American) would predict the magnitude of their psychological association, as measured by the IAT.

Figure 2b visualizes the critical interaction term. Behavioural performance on the different IATs was correlated with language statistics. When language statistics predicted that American English had a greater association, US participants showed a stronger association in the IAT. When language statistics predicted that British English had a stronger association, British participants showed a stronger association in the IAT ($\beta = -0.05$, s.e. $= 0.02$, $t = -2.88$; see Extended Data Fig. 7 for full model results).

In study 1, we found that a previously reported psychological gender association—the tendency to associate men with career and women with family—was correlated with the magnitude of that same association 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 between women and family words, showed stronger associations on the gender–career IAT. In a pre-registered, confirmatory analysis, we also find that this pattern extends to associations beyond career and gender. In a comparison of 31 different IATs, the magnitude of the association in speaker's dialect of English (American versus British) predicted their behavioural association, as measured by the IAT. These results suggest a close correspondence between psychological and linguistic gender associations. In study 2, we try to better understand the source of the gender–career association in language by investigating whether it is related to two structural features of language: grammatical gender and the presence of gendered occupation terms (for example, waiter/waitress).

Study 2: gender association and lexicalized gender. The similarity between language associations and implicit associations found in study 1 is consistent with multiple causal pathways. If language is causally related to implicit associations, then differences in the structural aspects of language that act to exaggerate linguistic gender association should predict greater implicit association. This relationship is difficult to explain if language merely reflects cultural stereotypes, since structural aspects of language are relatively fixed.

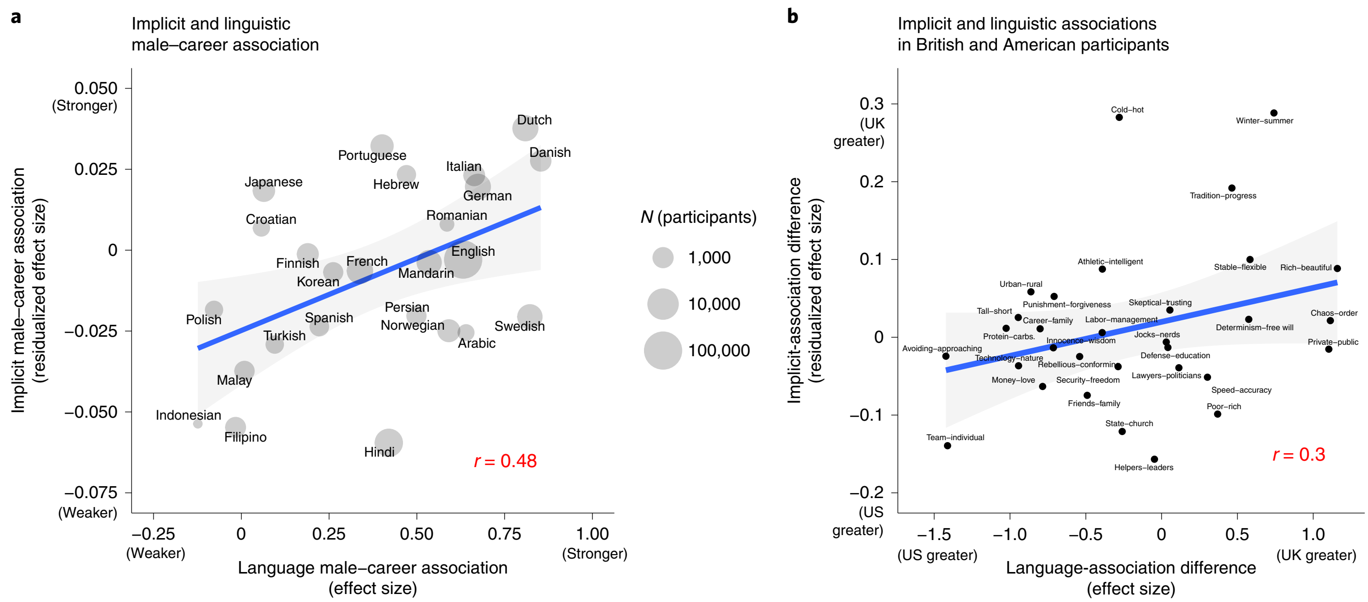


Fig. 2 | Implicit versus linguistic associations. a, Implicit male–career association (adjusted for participant age, gender and congruent or incongruent block order) as a function of the linguistic male–career association derived from word embeddings ($r(23) = 0.48$ [0.11, 0.74]; $P = 0.01$; $n = 25$; study 1b). Each point corresponds to a language. The size of the point is proportional to the number of participants from the country in which the language is dominant (total $N = 656{,}636$ participants). Linguistic associations are estimated from models trained on text in each language from the Wikipedia corpus. Larger values indicate a greater tendency to associate men with the concept of career and women with the concept of family. **b**, Difference (UK minus US) in implicit association versus linguistic association for 31 IAT types (study 1c, $N = 27{,}045$ participants). Error bands indicate standard error of the linear model estimate.

One such structural difference concerns the grammaticalization of gender. Some languages, such as Spanish, mark gender distinctions in a grammatically obligatory way, for example, 'enfermero' (nurse, masculine) versus 'enfermera' (nurse, feminine). Grammatical gender systems frequently demand gender-based agreement, for example, 'el enfermero alto' (the tall nurse, masculine) versus 'la enfermera alta' (the tall nurse, feminine), which may act to amplify gender associations in the language. Another structural difference is the existence of gender-specific terms such as 'waiter' versus '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 (for example, 'auteur', 'athlète' and 'juge').

In study 2, we examined whether grammatical gender and use of gender-specific occupation terms are associated with a greater psychological gender association and whether this relationship is further mediated by language statistics. Finding such associations would lend support to the hypothesis that language has a causal role in shaping gender associations 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 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', whereas a Spanish document promoting nursing careers to men would be committed to using gender-marked forms.

Speakers of languages with a grammatical gender system ($N = 12$ languages) did not differ from those without ($N = 13$ languages) in terms of implicit ($M_{\text{diff}} = 0.01$ [−0.01, 0.03]; $t(22.99) = 0.74$, $P = 0.47$; $d = 0.29$ [−0.54, 1.13]) or explicit career–gender associations ($M_{\text{diff}} = 0.08$ [−0.07, 0.23]; $t(17.67) = 1.17$, $P = 0.26$; $d = 0.48$ [−0.36, 1.32]). However, the strength of the women–family and men–career associations, as measured by the IAT, were reliably correlated with degree of gender-specific marking on occupation words: languages with more gender-specific forms tended to have speakers with greater implicit career–gender association ($r(23) = 0.57$ [0.22, 0.79], $P = 0.003$; Fig. 3a and Table 1). There was no relationship between explicit psychological career–gender association and lexical marking of occupation words ($r(23) = 0.11$ [−0.3, 0.48], $P = 0.61$).

We next examined whether the existence of gender-specific occupation terms was predicted by a greater encoding of gender associations (male versus female) in the distributional statistics of the language. We fit a mixed-effects model predicting degree of gender association 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 association for those occupations. This was true both for models trained on the subtitle corpus ($\beta = 0.46$; s.e. $= 0.08$; $t = 6.08$) and those trained on the Wikipedia corpus ($\beta = 0.89$; s.e. $= 0.1$; $t = 8.93$). For example, 'secretary' has a greater gender association in Italian, which has distinct male and female terms, compared with English, which has only a gender-neutral form.

This relationship also held at the level of languages: languages with more gendered occupation terms had stronger career–gender associations in their language statistics (subtitle: $r(17) = 0.6$ [0.2, 0.83], $P = 0.006$; Wikipedia: $r(23) = 0.77$ [0.53, 0.89], $P < 0.001$).

Finally, we examined the relationship between gender association in language statistics for occupation words and psychological career–gender associations. Unlike in study 1, all the target words in the present study referred to people (occupations) and thus could potentially 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 association in language statistics for occupation words and psychological associations for speakers of that language. Consistent with this prediction, gender association in language

Table 1 | Correlation (Pearson's *r*) for all measures in studies 1b and 2 using language as the unit of analysis

	Explicit male–career assoc.	Implicit male–career assoc. (IAT)	Percent women in STEM	Lang. male–career assoc. (subt.)	Lang. male–career assoc. (Wiki.)	Prop. gendered occupation terms	Lang. occup. genderness (subt.)	Lang. occup. genderness (Wiki.)
Implicit male–career assoc. (IAT)	0.18 [−0.23, 0.54], 0.39							
Percent women in STEM	0.18 [−0.26, 0.56], 0.43	−0.53 [−0.78, −0.14], 0.01						
Lang. male–career assoc. (subt.)	−0.08 [−0.5, 0.38], 0.74	0.5 [0.08, 0.77], 0.02	−0.55 [−0.81, −0.11], 0.02					
Lang. male–career assoc. (Wiki.)	0.34 [−0.06, 0.65], 0.09	0.48 [0.11, 0.74], 0.01	−0.19 [−0.57, 0.25], 0.4	0.51 [0.09, 0.78], 0.02				
Prop. gendered occupation terms	0.11 [−0.3, 0.48], 0.61	0.57 [0.22, 0.79], 0.002	−0.35 [−0.67, 0.09], 0.12	0.28 [−0.18, 0.64], 0.23	0.18 [−0.23, 0.54], 0.38			
Lang. occup. genderness (subt.)	0.16 [−0.32, 0.57], 0.53	0.49 [0.04, 0.77], 0.03	−0.26 [−0.66, 0.25], 0.31	0.38 [−0.09 0.71], 0.11	0.51 [0.07, 0.78], 0.03	0.6 [0.2, 0.83], 0.01		
Lang. occup. genderness (Wiki.)	0.18 [−0.23, 0.54], 0.39	0.49 [0.11, 0.74], 0.01	−0.53 [−0.78, −0.14], 0.01	0.41 [−0.03, 0.72], 0.07	0.53 [0.18, 0.77], 0.01	0.77 [0.53, 0.89], <0.001	0.81 [0.57, 0.93], <0.001	
Median country age	−0.07 [−0.45, 0.33], 0.73	0.61 [0.28, 0.81], 0.001	−0.42 [−0.72, 0], 0.05	0.31 [−0.15, 0.66], 0.18	0.25 [−0.16, 0.59], 0.22	0.35 [−0.05, 0.65], 0.09	0.44 [−0.02, 0.74], 0.06	0.34 [−0.07, 0.65], 0.1

Implicit and explicit male–career association measures are residualized for participant age, gender and task order; 95% confidence intervals are given in brackets, followed by the corresponding *P*-value. Assoc., association; lang., language; subt., subtitle corpus; Wiki., Wikipedia corpus; prop. gendered occup. terms., proportion of occupation terms that are gendered; occup. genderness, degree to which occupation terms in a language tend to be associated with a particular gender in the language statistics.

statistics for occupation words was positively correlated with implicit career–gender association (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. 3b). By contrast, explicit psychological career–gender association was not predicted by gender association in language statistics (subtitle: $r(17)=0.16$ [−0.32, 0.57], $P=0.57$; Wikipedia: $r(23)=0.18$ [−0.23, 0.54], $P=0.39$).

To understand the relative predictive power of language statistics and distinct occupation terms, we fit an additive linear model predicting implicit association from language statistics and proportion distinct forms. Because language statistics for occupation terms and the proportion of gendered forms in each language were highly correlated (subtitle: $r(18)=0.75$ [0.46, 0.9], $P<0.001$; Wikipedia: $r(23)=0.70$ [0.42, 0.86], $P<0.001$), we used a measure of language statistics that was more weakly correlated with proportion of gendered forms, namely, the degree of gender association in language statistics based on the set of IAT words described in study 1b. Both gender association in language statistics (based on IAT words) and the proportion of gender–specific occupation words were independent predictors of implicit associations as measured by the IAT. The two predictors accounted for 41% of variance when using the subtitle corpus and 45% of variance for the Wikipedia corpus (Extended Data Fig. 8).

In Study 2, we investigated whether structural features of language—the presence of a grammatical gender systems and the propensity to lexicalize gender distinctions—correlated with implicit association. Grammatical gender was not reliably correlated with implicit association. Languages that use more gender-specific occupation terms, however, did predict a greater implicit association. This finding suggests that one driver of the relationship between language and psychological career–gender associations observed in study 1 may be the presence of gender–specific occupation terms.

Discussion

Where do we get our gender stereotypes? Non-linguistic experiences surely play a role, but might we also be learning our associations from the language to which we are exposed? We used a large-scale dataset of IATs that measure people's implicit associations of men with career and women with family. We related these associations to the linguistic gender associations computed from patterns of word co-occurrences in the dominant language spoken in the country of each participant. In study 1, we found that languages with stronger gender associations embedded in their distributional structure tend to have speakers that have stronger implicit associations. In study 2, we found a positive relationship between a structural language feature—the prevalence of gender-marked occupation terms—and the strength of people's implicit associations.

Our work characterizes the relationship between cultural stereotypes and cross-linguistic differences in language statistics. Establishing that this relationship exists is the first step to understanding the underlying causal pathways. The positive correlation between the strength of the career–gender association in language and speakers' IAT results is consistent both with language playing a causal role in the emergence of cultural stereotypes and with language merely reflecting existing stereotypes of its speakers[22,25]. The correlational approach of our studies does not allow us to fully distinguish between these possibilities. Among the findings that could help confirm or disconfirm the hypothesis that language plays a causal role in shaping psychological associations are (1) longitudinal analyses testing whether changes in language statistics predict or follow changes in measured implicit associations[25];

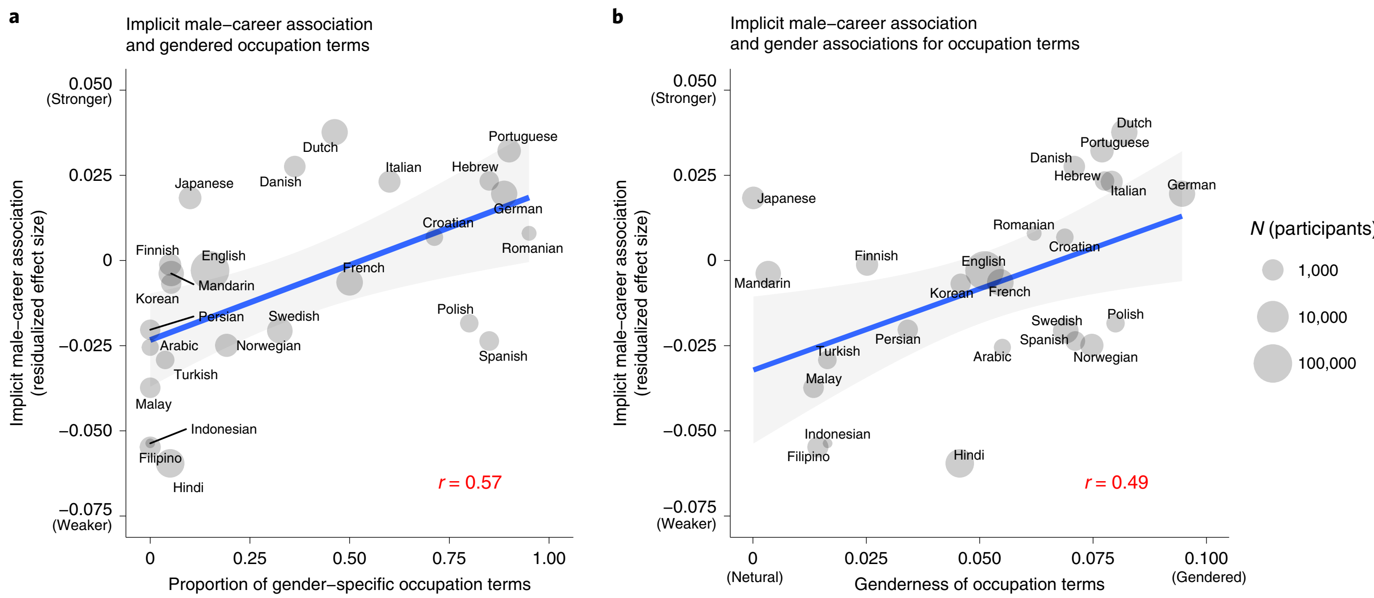


Fig. 3 | Implicit male–career association and mean gender association. a, Implicit male–career association (adjusted for participant age, gender and congruent or incongruent block order) as a function of the proportion of gender-specific labels for the set of words referring to occupations ($r(23) = 0.57$ [0.22, 0.79]; $P = 0.003$; $n = 25$). **b**, Mean gender association of words referring to occupations from word embeddings trained on the Wikipedia corpus ($r(23) = 0.49$ [0.11, 0.74]; $P = 0.01$). Each point corresponds to a language, with the size of the point corresponding to the number of participants speaking that language (total $N = 656{,}636$ participants). Error bands indicate standard error of the linear model estimate.

(2) quasi-experimental tests that involve, for example, measuring implicit associations in bilingual subjects using stimuli in languages that embed different linguistic associations; and (3) experimental designs that measure the effect of manipulating language statistics on people's implicit associations.

Our results speak to several recent attempts to understand large-scale correlates of gender stereotypes[6] and differences in gender preferences more broadly[35]. These studies have argued that increases in institutional gender equality (which are strongly associated with increases in national gross domestic product (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 gender differences in preferences (for example, women being more risk averse and less patient than men)[35]. Although our results do not contradict this possibility, they suggest that associations learned from language may be a part of the fuller picture. The encoding of gender stereotypes in different languages is itself correlated with GDP (higher GDP correlates with stronger career–gender linguistic associations, $r(31) = 0.58$ [0.29, 0.77], $P < 0.001$) and also with previously reported individual-level predictors of STEM inequality, such as self-efficacy in science (ref. [6]; $r(28) = 0.59$ [0.3, 0.79], $P < 0.001$) and general gender preferences (ref. [35]; $r(25) = 0.48$ [0.12, 0.73], $P = 0.01$; Extended Data Fig. 9).

One unexpected finding is the substantial relationship between median country age (for example, 29.9 in Israel versus 47.1 in Germany) and the gender–career IAT: countries with older populations have stronger career–gender associations ($r = 0.61$; see Table 1). The direction of this relationship is consistent with the by-participant analyses (older participants have stronger career–gender associations, $r = 0.06$), and is consistent with older populations having more traditional gender norms. Importantly, the two effects are distinct: participants from countries with an older population show stronger career–gender associations after adjusting for their own age. This effect holds even after controlling for the percentage of women in STEM fields within a country (Extended Data Fig. 2b).

One (admittedly speculative) possibility is that younger participants from countries with older populations are more likely to be exposed to stronger career–gender associations from language produced by older individuals.

One limitation of our work is its reliance on the IAT, which has been criticized for both its low reliability[36] and limited external validity[37]. Issues of reliability are less relevant here because we use the IAT to measure group-level differences rather than individual-difference measures, and group-level estimates have been shown to be stable[38]. However, concerns about validity are important, particularly because we find that language measures and explicit psychological measures of gender associations are uncorrelated, although this lack of a relationship may be due to the explicit association measure being too coarse. Nevertheless, the strong negative correlation we find between the proportion women in STEM and gender–career associations in language statistics ($r = -0.55$) provides compelling evidence that language associations are related to real-world consequences. Understanding the full import of linguistic associations on cultural stereotypes will require obtaining measures more closely related to real-world behaviour. Two additional questions for further research is how much exposure to the relevant language statistics is sufficient to produce differences in beliefs, and how resilient the learned associations are to other sources of information. For example, if a bilingual individual is exposed to conflicting gender associations in two languages, is the net effect a combination of the two sources of information, or does it vary dynamically with the linguistic context of a given interaction (for example, refs. [39,40]).

Cultural stereotypes are acquired through experience. Here we show that group-level differences in implicit associations are strongly correlated with the strength of gender associations encoded in the statistics of different languages. This pattern suggests that the statistics of language use could be 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[26]. In other cases, such as the gender stereotypes investigated here, language may play a powerful role in their formation and ultimately contribute to undesirable structural inequality. Understanding the extent to which language has a causal role in the formation of these stereotypes is therefore an important first step towards changing these consequences.

Methods

All reported correlations are Pearson's r values. Two-sample t-test are calculated using Welch's test. Effect size measures are classic Cohen's d. Brackets indicate 95% confidence intervals. All statistical tests are two-sided analyses. Data distributions were assumed to be normal but this was not formally tested. Data analysis was not performed blind to the identify of the variables.

Description of the IAT dataset. We analysed 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 on page 104 of ref. [13]. We only analysed 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 association question). This cut-off was arbitrary, but the pattern of findings reported here holds for a range of minimum participant values (Supplementary Fig. 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 were 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 associations 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)[41]. Both models were trained using the fastText algorithm (a variant of word2vec)[42]. 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[43]. 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 behavioural task[13] with one slight modification. In the behavioural task, proper names were used to cue the male and female categories (for example 'John' and 'Amy'), but because there are not direct translation equivalents of proper names, we instead used a set of generic gendered words that had been previously used for a different version of the gender IAT (for example, 'man' and 'woman')[13]. Our linguistic stimuli were therefore a set of eight female and eight male 'target words' (identical to study 1a, described in the main text), and the set of eight 'attribute words' used in the Project Implicit gender–career IAT: eight related to careers ('career', 'executive', 'management', 'professional', 'corporation', 'salary', 'office' and 'business') and eight related to families ('family', 'home', 'parents', 'children', 'cousins', 'marriage', 'wedding' and '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-association 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 (that is, relative similarity of male to career versus relative similarity of female to career). Our effect size measure is identical to that used in ref. [21] with an exception for grammatically gendered languages. Namely, for languages with grammatically gendered attribute words (for example, niñas for female children in Spanish), we calculated the relationship between target words and attribute words of the same gender (that is 'hombre' (man) to 'niños' (male children) 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 gender association from the Project Implicit data, larger values indicate larger association between males and career and between females and family.

We calculated gender–career association 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 (Wikipedia, $N=25$; subtitle, $N=20$), representing 8 language families, and corresponded to 656,636 participants in the Project Implicit dataset.

Study 1c. The AIID dataset was partitioned into two samples: exploratory (15%) and confirmatory (85%). On the basis of the exploratory sample, we pre-registered our analysis plan for the confirmatory sample (https://osf.io/3f9ed, 8 February 2019) 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 American and British English corpora. To measure the associations in language, we trained word-embedding models on equally sized subsets of British National Corpus (BNC)[44] and Corpus of Contemporary American English (COCA)[45]. The model was trained using the fastText algorithm[42], with a vector size of 400 and window size of 10. We then calculated an association effect size for each IAT in each English dialect, using the same method as in study 1b.

After data exclusion (using criteria similar to study 1a; see Supplementary Methods), our final sample in the confirmatory AIID dataset included data from 27,045 administrations of the IAT across the 31 IATs (American: $N=25{,}523$; British: $N=1{,}522$). Each participant completed an average of 1.23 different IATs (s.d. $=0.80$). For each administration 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 implicit association for each participant from their location (US versus UK), the linguistic association from American English- and British English-trained models, and the interaction of the two factors. We included participant and IAT test type as random intercepts. We fit this and subsequent mixed-effect models with the lme4 R package[46]. 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 Extended Data Fig. 7 for results of this model and exact pre-registered model).

Study 2. We identified 20 occupation terms that could be translated into all 25 of our languages, and that were balanced in terms of their perceived gender associations in the workforce[47]. We then translated these words into each of the 25 languages in our sample, distinguishing between male and female variants (for example, 'waiter' versus '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[48] 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 ($M=0.51$ [0.28, 0.73]; $t(14.89)=4.85$, $P<0.001$; $d=2$ [0.98, 3.01]).

We then estimated the extent to which each occupation term was associated with a specific gender ('genderness') 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 linguistic gender association to males and females using the same pairwise similarity metric as in study 1a. A genderness score was calculated for each word as the absolute value of the difference in association between males and females. Larger values indicate greater association with females relative to males or males relative to females. We averaged across occupations within a language to get a language-level estimate of occupation genderness. One language was excluded from the subtitle analysis (German) because more than 50% of the words were missing from the model, but the results remain the same when this language is included. We then compared each of the three language measures (grammatical gender, proportion specific gender forms, and genderness in language statistics for occupation words) to the psychological male–career measures described in study 1b (implicit and explicit associations, adjusted for age, gender and block order).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available at https://github.com/mllewis/IATLANG. Source data are provided with this paper.

Code availability
All code that supports the findings of this study is available at https://github.com/mllewis/IATLANG.

Received: 24 June 2019; Accepted: 26 June 2020;
Published online: 03 August 2020

References

1. Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C. & Bigler, R. S. Mother–child conversations about gender: understanding the acquisition of essentialist beliefs. *Monogr. Soc. Res. Child Dev.* **69**, 1–142 (2004).
2. Bian, L., Leslie, S. J. & Cimpian, A. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* **355**, 389–391 (2017).
3. Ceci, S. J. & Williams, W. M. Understanding current causes of women's underrepresentation in science. *Proc. Natl Acad. Sci. USA* **108**, 3157–3162 (2011).
4. Leslie, S. J., Cimpian, A., Meyer, M. & Freeland, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262–265 (2015).
5. Miller, D. I., Eagly, A. H. & Linn, M. C. Women's representation in science predicts national gender–science stereotypes: Evidence from 66 nations. *J. Educ. Psychol.* **107**, 631 (2015).
6. Stoet, G. & Geary, D. C. The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychol. Sci.* **29**, 581–593 (2018).
7. Dryer, M. S. & Haspelmath M. eds. *WALS Online* (Max Planck Institute for Evolutionary Anthropology, 2013); https://wals.info/
8. Rhodes, M. & Brickman, D. Preschoolers' responses to social comparisons involving relative failure. *Psychol. Sci.* **19**, 968–972 (2008).
9. Cimpian, A., Mu, Y. & Erickson, L. C. Who is good at this game? Linking an activity to a social category undermines children's achievement. *Psychol. Sci.* **23**, 533–541 (2012).
10. Cimpian, A. & Markman, E. M. The generic/nongeneric distinction influences how children interpret new information about social others. *Child Dev.* **82**, 471–492 (2011).
11. Rhodes, M., Leslie, S. J., Yee, K. M. & Saunders, K. Subtle linguistic cues increase girls' engagement in science. *Psychol. Sci.* **30**, 455–466 (2019).
12. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
13. Nosek, B. A., Banaji, M. R. & Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn. Theory Res. Pract.* **6**, 101–115 (2002).
14. Payne, B. K., Vuletich, H. A. & Brown-Iannuzzi, J. L. Historical roots of implicit bias in slavery. *Proc. Natl Acad. Sci. USA* **116**, 11693–11698 (2019).
15. Hehman, E., Calanchini, J., Flake, J. K. & Leitner, J. B. Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *J. Exp. Psychol. Gen.* **148**, 1022–1040 (2019).
16. Charlesworth, T. E. & Banaji, M. R. Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).
17. Firth, J. R. *Studies in Linguistic Analysis* (Philological Society, 1957).
18. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
19. Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**, 203–208 (1996).
20. Lenci, A. Distributional semantics in linguistic and cognitive research. *Ital. J. Linguist.* **20**, 1–31 (2008).
21. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
22. Bhatia, S. The semantic representation of prejudice and stereotypes. *Cognition* **164**, 46–60 (2017).
23. von der Malsburg, T., Poppels, T. & Levy, R. Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 US and 2017 UK election. *Psychol. Sci.* **31**, 115–128 (2020).
24. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).
25. Greenwald, A. G. An AI stereotype catcher. *Science* **356**, 133–134 (2017).
26. Lupyan, G. & Lewis, M. From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Lang. Cogn. Neurosci.* **34**, 1319–1337 (2017).
27. Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the implicit association test: I. an improved scoring algorithm. *J. Pers. Soc. Psychol.* **85**, 197 (2003).
28. Forscher, P. S. et al. A meta-analysis of procedures to change implicit measures. *J. Pers. Soc. Psychol.* **117**, 522–559 (2019).
29. CIA. *The CIA World Factbook 2017* (2016); https://www.cia.gov/library/publications/the-world-factbook/index.html
30. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* https://arxiv.org/abs/1301.3781 (2013).
31. van Paridon, J. & Thompson, B. subs2vec: word embeddings from subtitles in 55 languages. Preprint at *OSF* https://doi.org/10.31234/osf.io/fcrmy (2019).
32. Lison, P. & Tiedemann, J. OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In *Proc. 10th International Conference on Language Resources and Evaluation* (ELRA, 2016).
33. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Ling.* **5**, 135–146 (2017).
34. Hussey, I., et al. *The Attitudes, Identities, and Individual differences (AIID) Study and Dataset* https://osf.io/pcjwf/ (2019).
35. Falk, A. & Hermle, J. Relationship of gender differences in preferences to economic development and gender equality. *Science* **362**, eaas9899 (2018).
36. Lane, K. A., Banaji, M. R., Nosek, B. A. & Greenwald, A. G. in *Implicit Measures of Attitudes* (eds. Wittenbrink, B. & Schwarz, N.) 59–102 (2007).
37. Fazio, R. H. & Olson, M. A. Implicit measures in social cognition research: their meaning and use. *Annu. Rev. Psychol.* **54**, 297–327 (2003).
38. Payne, B. K., Vuletich, H. A. & Lundberg, K. B. The bias of crowds: how implicit bias bridges personal and systemic prejudice. *Psychol. Inq.* **28**, 233–248 (2017).
39. Marian, V. & Kaushanskaya, M. Language context guides memory content. *Psychon. Bull. Rev.* **14**, 925–933 (2007).
40. Athanasopoulos, P. Cognitive representation of colour in bilinguals: the case of Greek blues. *Biling. Lang. Cogn.* **12**, 83–95 (2009).
41. Scott, G. G., Keitel, A., Becirspahic, M., Yao, B. & Sereno, S. C. The Glasgow norms: ratings of 5,500 words on nine scales. *Behav. Res. Methods* **51**, 1258–1270 (2019).
42. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of tricks for efficient text classification. Preprint at *arXiv* https://arxiv.org/abs/1607.01759 (2016).
43. Simons, G. F. & Charles, D. F. (eds) *Ethnologue: Languages of the World* (SIL International, 2018).
44. Burnard, L. *Users Reference Guide for the British National Corpus* (Oxford Univ. Computing Services, 1995).
45. Davies M. *The Corpus of Contemporary American English* (2008); https://corpus.byu.edu/coca/
46. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
47. Misersky, J. et al. Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behav. Res. Methods* **46**, 841–871 (2014).
48. Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. (eds) *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, 2008); http://wals.info

Acknowledgements
The authors acknowledge National Science Foundation Perception, Action and Cognition 1734260 for funding support for this work. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
M.L. and G.L. designed the research and wrote the manuscript. M.L. conducted the data analysis.

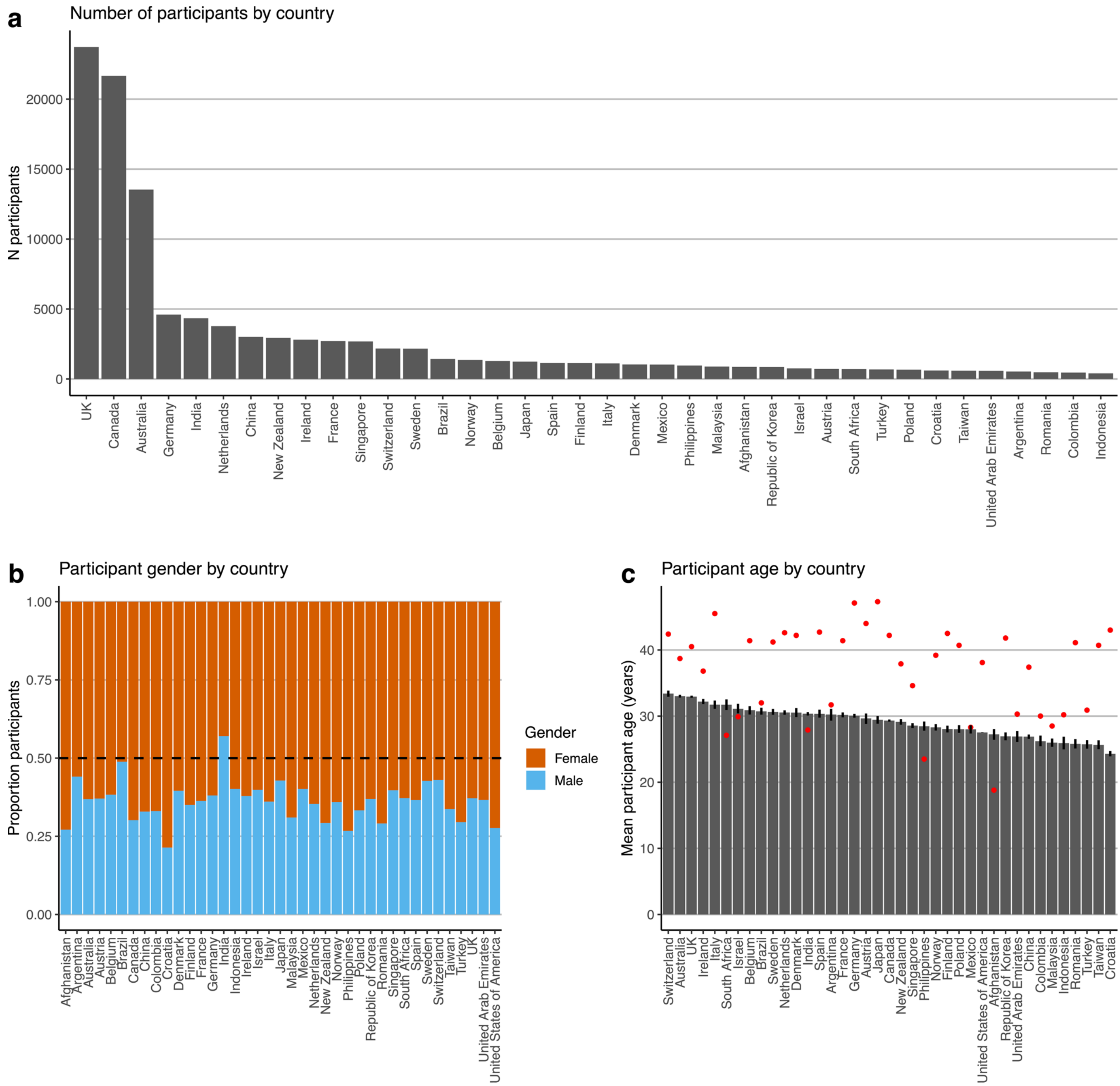
Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41562-020-0918-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0918-6.

Correspondence and requests for materials should be addressed to M.L.

Peer review information Primary handling editor: Aisha Bradshaw



Extended Data Fig. 1 | Sample size and demographic characteristics of Project Implicit data. a, Number of participants by country after exclusions (note that US participants are excluded from the visualization because of the large sample size; $N = 545,673$). Our final sample included 657,335 participants from 39 countries (see Supplementary Information for exclusion criteria). **b,** Gender distribution of participants by country after exclusions. Across countries, there tended to be more female participants relative to male participants ($M = 0.64$ proportion females; $SD = 0.06$). **c,** Age distribution of participants by country after exclusions. Ranges correspond to 95% CIs. Red points show median age by country.

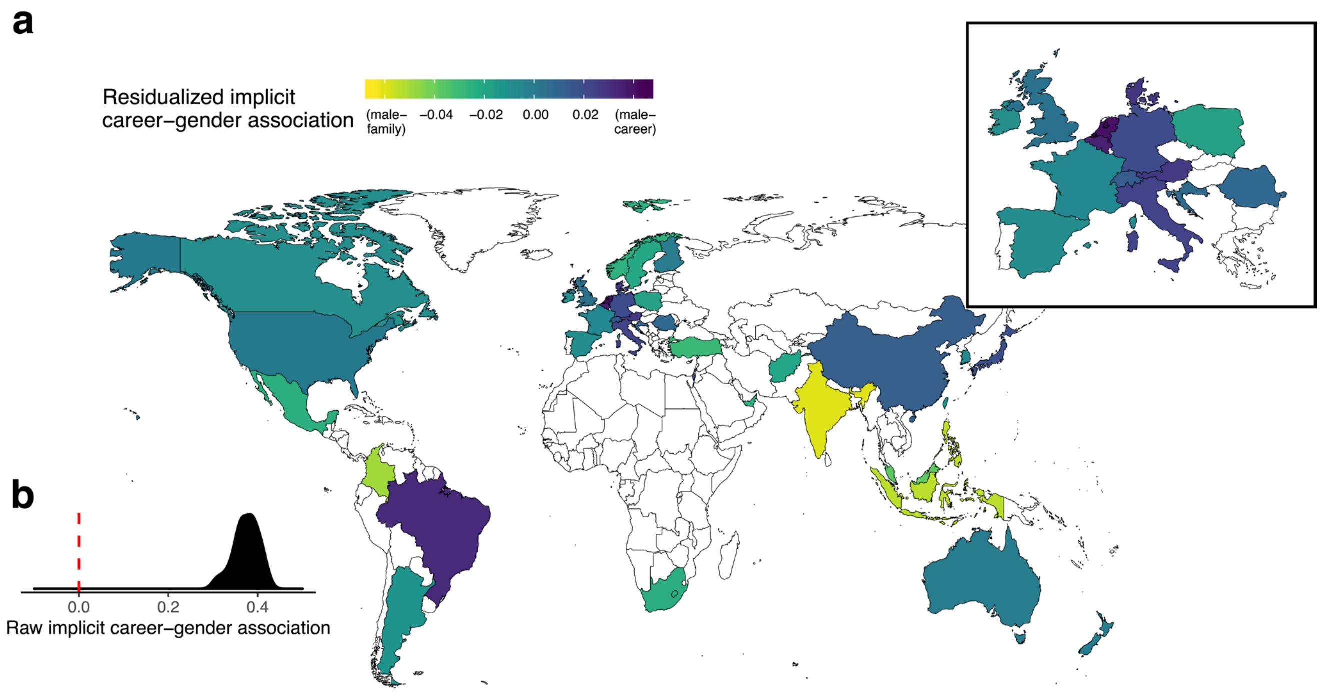
a

	Overall IAT D-score		
Predictors	Estimate	SE	Statistic
(intercept)	-0.03	0.02	-1.53
median country age	< 0.01	< 0.01	4.96
log age	0.06	< 0.01	55.05
sex (M)	-0.10	< 0.01	-103.14
task order	0.09	< 0.01	105.67
Random Effects			
s2	0.12		
T00 country code	< 0.01		
N country code	39		
Observations	657335		

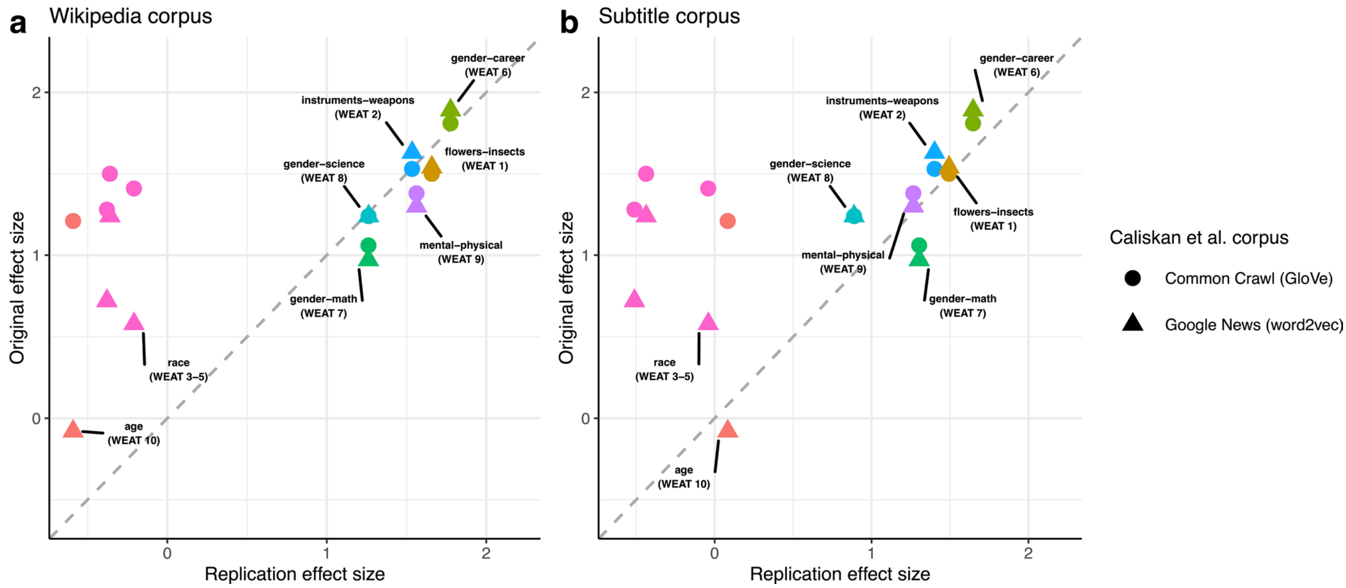
b

	Overall IAT D-score		
Predictors	Estimate	SE	Statistic
(intercept)	< -0.01	0.03	-0.05
median country age	< 0.01	< 0.01	4.11
log age	0.06	< 0.01	54.56
sex (M)	-0.10	< 0.01	-102.78
task order	0.09	< 0.01	104.76
% women in STEM	< -0.01	< 0.01	-2.73
Random Effects			
s2	0.12		
T00 country code	< 0.01		
N country code	33		
Observations	646271		

Extended Data Fig. 2 | Models predicting IAT effect size at the participant level. Median country age predicts IAT effect size over and above participant age at the participant level: Countries with older populations tend to have individuals with stronger implicit career-gender associations, even after controlling for participant age. The table presents an additive mixed-effect regression predicting IAT D-score at the participant level with participant age and median country age, controlling for participant sex and trial order. The model includes by-country random intercepts. **b,** The relationship between median country age and IAT effect size holds, even after controlling for the percentage women in STEM. The table presents an additive mixed effect model predicting IAT D-score at the participant level with participant age, median country age and percentage women in STEM in country, controlling for participant sex and trial order. The model includes by-country random intercepts.



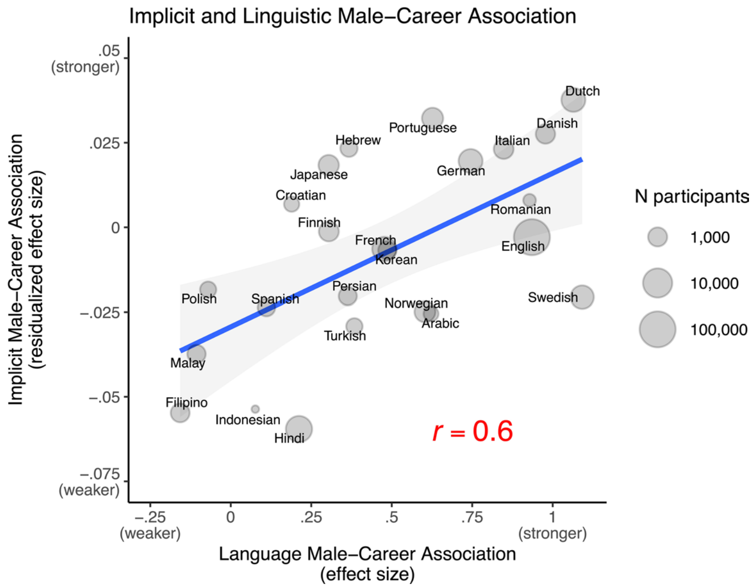
Extended Data Fig. 3 | Geographic distribution of IAT scores. a, Residualized implicit career-gender association (IAT score) shown by country. IAT scores are residualized for participant age, gender, and task order (N = 657,335). Larger values (blue) indicate a larger bias to associate men with the concept of career and women with the concept of family. Countries in white correspond to countries for which there was insufficient data to estimate the country-level career-gender association. Inset shows IAT scores for European countries only. 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. **b,** Distribution of raw (unresidualized) implicit career-gender association (IAT D-score) across countries. All countries in our sample showed a tendency to associate men with career and women with family.



Extended Data Fig. 4 | Replication of Caliskan et al. (2017) with our corpora. We replicate the original set of Caliskan, Bryson, and Narayanan (2017; *CBN*)[21] findings using the English-trained versions of the models used in our main analyses (models trained on the Wikipedia and Subtitles corpora). For each model, we calculate an effect size for each of the 10 IAT types reported in CBN: flowers/insects-pleasant/unpleasant, instruments/weapons-pleasant/unpleasant, European-American/Afro-American-pleasant/unpleasant, males/females-career/family, math/arts-male/female, science/arts-male/female, mental-disease/physical-disease-permanent/temporary, and young/old-pleasant/unpleasant (labelled as Word-Embedding Association Test (WEAT) 1-10 in CBN). We calculate the bias using the same effect size metric described in CBN, a standardized difference score of the relative similarity of the target words to the target attributes (that is relative similarity of male to career vs. relative similarity of female to career). This measure is analogous to the behavioural effect size measure where larger values indicate larger bias. The figure shows the effect size measure derived from the English Wikipedia corpus (**a**) and the English Subtitle corpus (**b**) plotted against effect size estimates reported by CBN from two different models (trained on Common Crawl and Google News corpora). Point color corresponds to bias type, and point shape corresponds to the two CBN models. With the exception of biases related to race and age, effect sizes from our corpora are comparable to those reported by CBN. In particular, for the gender-career IAT-the bias relevant to our current purposes-we estimate the effect size to be 1.78 (Wikipedia)/1.65 (Subtitle), while CBN estimates it to be 1.81 (Common Crawl)/1.89 (Google News).

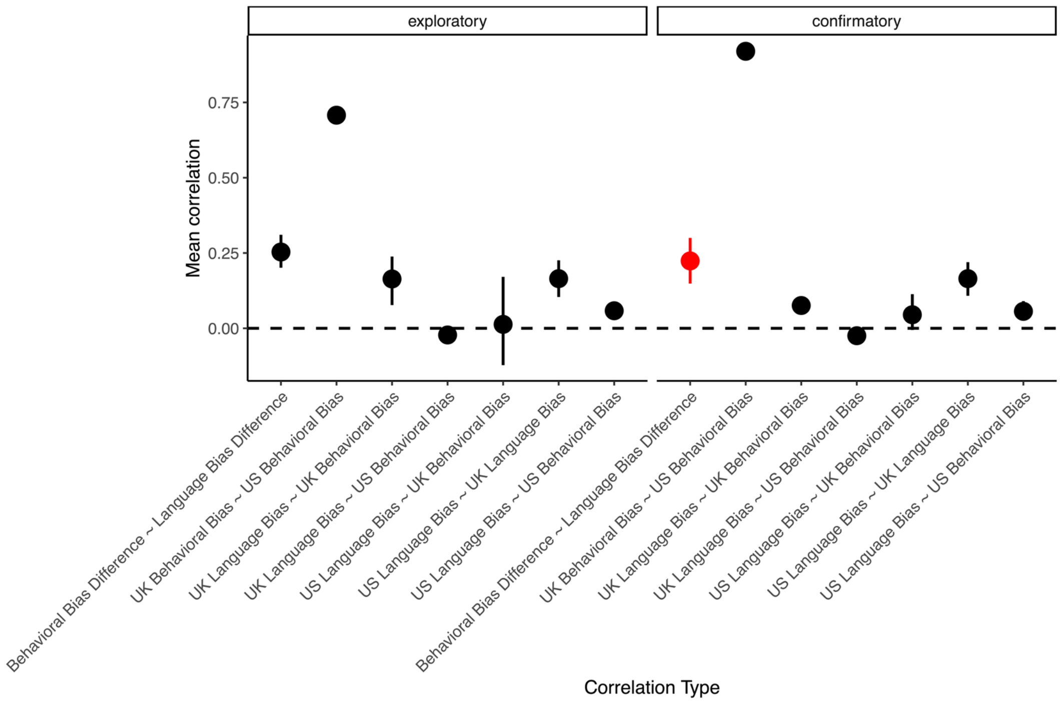
	Explicit Male-Career Assoc.	Implicit Male-Career Assoc. (IAT)	Percent Women in STEM	Male-Career Assoc. (Subt.)	Male-Career Assoc. (Wiki.)	Prop. Gendered Occup. Terms	Lang. Occup. Genderness (Subt.)	Lang. Occup. Genderness (Wiki.)
Explicit Male-Career Assoc.		.28 [−.14, .62], 0.18	.16 [−.26, .53], 0.45	−.06 [−.45, .35], 0.78	.38 [−.03, .68], 0.07	.14 [−.28, .52], 0.51	.21 [−.21, .56], 0.33	.22 [−.20, .57], 0.31
Implicit Male-Career Assoc. (IAT)	.28 [−.14, .62], 0.18		−.38 [−.68, .03], 0.07	.42 [.02, .70], 0.04	.43 [.03, .71], 0.04	.48 [.09, .74], 0.02	.31 [−.11, .63], 0.14	.37 [−.03, .68], 0.07
Percent Women in STEM	.16 [−.26, .53], 0.45	−.38 [−.68, .03], 0.07		−.49 [−.75, −.11], 0.02	−.09 [−.48, .32], 0.67	−.23 [−.58, .19], 0.27	−.10 [−.48, .32], 0.65	−.46 [−.73, −.07], 0.02
Male-Career Assoc. (Subt.)	−.06 [−.45, .35], 0.78	.42 [.02, .70], 0.04	−.49 [−.75, −.11], 0.02		.47 [.08, .73], 0.02	.20 [−.23, .56], 0.36	.28 [−.14, .61], 0.18	.35 [−.07, .66], 0.1
Male-Career Assoc. (Wiki.)	.38 [−.03, .68], 0.07	.43 [.03, .71], 0.04	−.09 [−.48, .32], 0.67	.47 [.08, .73], 0.02		.11 [−.31, .49], 0.62	.46 [.06, .73], 0.03	.49 [.11, .75], 0.01
Prop. Gendered Occup. Terms	.14 [−.28, .52], 0.51	.48 [.09, .74], 0.02	−.23 [−.58, .19], 0.27	.20 [−.23, .56], 0.36	.11 [−.31, .49], 0.62		.53 [.17, .77], 0.01	.73 [.47, .88], <.001
Lang. Occup. Genderness (Subt.)	.21 [−.21, .56], 0.33	.31 [−.11, .63], 0.14	−.10 [−.48, .32], 0.65	.28 [−.14, .61], 0.18	.46 [.06, .73], 0.03	.53 [.17, .77], 0.01		.79 [.56, .90], <.001
Lang. Occup. Genderness (Wiki.)	.22 [−.20, .57], 0.31	.37 [−.03, .68], 0.07	−.46 [−.73, −.07], 0.02	.35 [−.07, .66], 0.1	.49 [.11, .75], 0.01	.73 [.47, .88], <.001	.79 [.56, .90], <.001	

Extended Data Fig. 5 | Pairwise Correlations partialing out the effect of median country age. Partial correlations (Pearson's r) for all measures in Study 1b and 2 using language as the unit of analysis, controlling for median country age. 95% CIs are given in brackets followed by the corresponding p-value. Implicit and explicit male-career association measures are residualized for participant age, gender, and task order. 'Assoc.' = association; 'Lang.'= language; 'Subt.'/ 'Wiki.' = Subtitle/Wikipedia corpora; 'Prop. Gendered Occup. Terms.' = proportion of occupation terms that are gendered. 'Occup. Genderness' = degree to which occupation terms in a language tend to be associated with a particular gender in the language statistics.



Extended Data Fig. 6 | Replication of Study 1b on Wikipedia corpus excluding translations. Both the Subtitle and Wikipedia corpora likely contain some documents that are translated from other languages (for example, the Wikipedia article on 'Paris' is written in French and then translated into English). The parallel content across languages allows us to estimate the gender bias in language statistics, while holding content constant across languages. Nevertheless, content may itself be a driver of gender bias (for example one language may have more articles about male politicians relative to another). To understand the contribution of language-specific content on gender bias, we constructed a corpus of Wikipedia articles in each language that were originally written in the target language (that is, untranslated), and trained word embedding models on the corpus in each language (see Supplemental Methods for details). We then used these models to calculate by-language male-career association scores using the same procedure as in Study 1b. Using models trained on the untranslated corpora, we replicate the key finding from Study 1b showing a positive correlation between the bias measured behaviorally with the IAT and measured in language ($r = .60$; $p = .002$; N participants = 656,636). Notably, the effect size is somewhat larger relative to the other two corpora types, presumably because additional bias is introduced by allowing the corpus content to vary across languages.

a



b

	Behavioral Effect Resid		
Predictors	Estimate	SE	Statistic
(intercept)	-0.00	0.05	-0.09
country (uk)	0.02	0.01	1.74
language bias difference (uk - us)	-0.03	0.07	-0.42
country:language bias difference	0.05	0.02	2.88
Random Effects			
s2	0.20		
T00 user id	0.01		
T00 domain	0.07		
N user id	22059		
N domain	31		
Observations	27045		

Extended Data Fig. 7 | Models examining UK-US bias difference in AIID dataset (Study 1c). a, The exact pre-registered analysis of Study 1c is presented. Pairwise correlations between all variables (language bias, behavioral bias, and UK-US difference measures) are shown, averaging across estimates of language bias from the 5 model runs (N participants = 27,045). Error bars are 95% CIs. As stated in the pre-registration, the key test of our hypothesis is that the correlation between the UK - US linguistic difference ('Language Bias Difference') and the UK - US behavioral difference ('Behavioral Bias Difference') is greater than 0 (shown in red). That data are consistent with this prediction. The confirmatory dataset is shown on the right, along with the smaller exploratory dataset on the left for reference. **b,** The full results of the mixed-effect model described in the Main Text are presented.

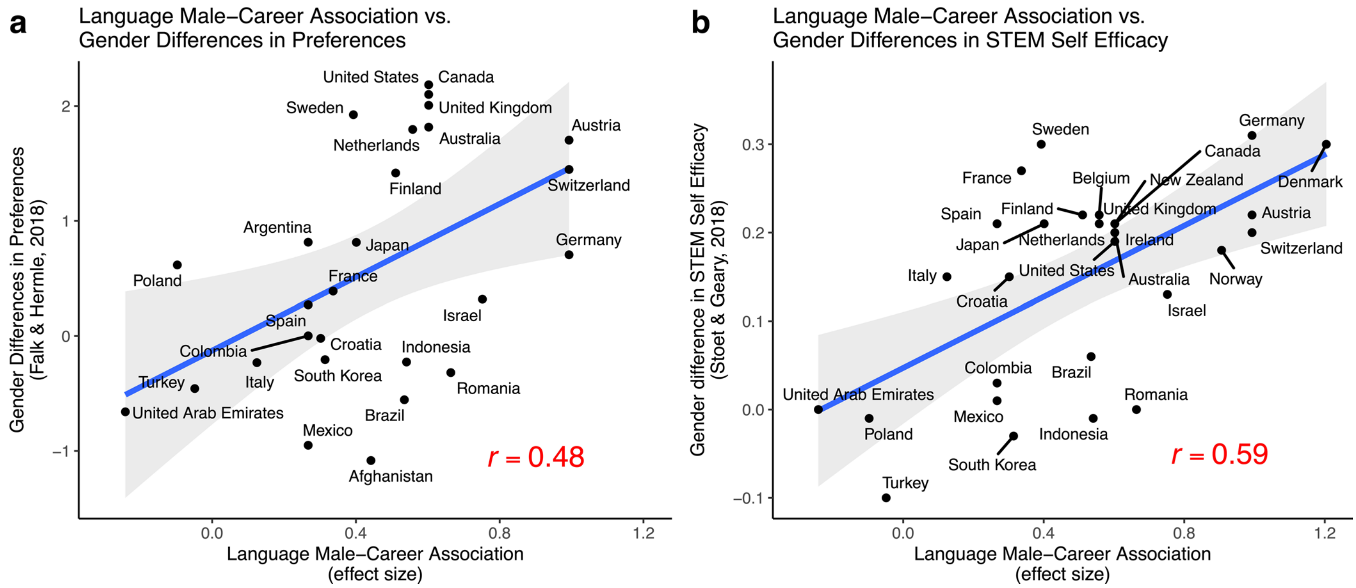
a

Predictors	Estimate	SE	Statistic	p-value
(Intercept)	0.11	0.16	0.69	0.50
Prop. Gendered Occup. Terms	0.39	0.16	2.36	0.03
Male-Career Assoc. (Subt.)	0.33	0.17	2.01	0.06

b

Predictors	Estimate	SE	Statistic	p-value
(Intercept)	0.00	0.15	0.00	$>.99$
Prop. Gendered Occup. Terms	0.52	0.16	3.27	$<.01$
Male-Career Assoc. (Wiki.)	0.36	0.16	2.28	0.03

Extended Data Fig. 8 | Models predicting implicit male-career association with proportion gender distinct labels and language career-gender association (Study 2). We predict the magnitude of implicit male-career association by language with an additive linear model. Predictors are proportion of occupation terms that are gendered ('Prop. Gendered Occup. Terms') and language male-career association as measured by word embeddings of the IAT words ('Male-Career Assoc.'). Model coefficients are shown for two models using estimates of language career-gender association from embedding models trained on Subtitle (**a**) and Wikipedia (**b**) corpora. The linear models account for 40.63% (Subtitle) and 45.32% (Wikipedia) of the variance in implicit male-career association. 'Subt.'/ 'Wiki.' = Subtitle/Wikipedia corpora.



Extended Data Fig. 9 | Gender associations in language and other psychological measures. Several recent studies[6,35] have presented novel theories to account for cases of structural inequality related to gender. Both of these studies argue that psychological differences play a causal role in the emergence of structural inequality. Here, we show that degree of gender bias in language is correlated with these psychological differences at the country level, consistent with the idea that language experience could be playing a causal role in the emergence of psychological differences. **a,** Gender differences in preferences[35] (composite score of 'six fundamental preferences with regard to social and nonsocial domains: willingness to take risks; patience, which captures preferences over the intertemporal timing of rewards; altruism; trust; and positive and negative reciprocity, which capture the costly willingness to reward kind actions or to punish unkind actions, respectively.') as a function of language male-career association measured in the Subtitle corpus. These two measures are correlated ($r(25) = 0.48$ [0.12, 0.73],$p = 0.01$): Countries with greater differences in gender preferences also have greater gender bias present in their languages. We also find that per capita GDP[49] is correlated with language gender male-career association measured in both corpora (Wikipedia: $r(35) = 0.64$ [0.4, 0.8],$p < .0001$; Subtitle: $r(31) = 0.58$ [0.29, 0.77],$p < .001$). However, the magnitude of the male-career association in the language spoken in a country predicts the magnitude of the male-career association measured via the behavioral IAT, controlling for both national GDP and median country age, in an additive mixed-effect model. **b,** Gender difference in STEM Self Efficacy[6] ('The sex difference in self efficacy (boys - girls)') as a function of male-career association measured in the Subtitle corpus. These two measures are correlated ($r(28) = 0.59$ [0.3, 0.79], $p < .001$): Countries with greater gender differences in self-efficacy also have greater gender bias present in their languages. Further, self-efficacy mediated the effect of language statistics on percentage of women in stem (path-ab = -0.33, $p = 0.01$), suggesting that language statistics could be critical causal factor underlying gender differences in STEM participation.

49. The World Bank. *World Development Indicators* (2017); http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

nature research

Corresponding author(s): Molly Lewis

Last updated by author(s): 4/21/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data data were collected.
Data analysis	Data were analyzed with R Version 3.5.1. The following packages were used in statistical reporting: lme4 (Bates, Maechler, Bolker, & Walker, 2015) , version 1.1-21 effsize (Torchiano, 2018), version 0.7.4 robmed (Alfons, 2018), version 0.3.0 The full analysis pipeline is available at https://github.com/mllewis/IATLANG.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data and code that support the findings of this study are available in an online repository (https://github.com/mllewis/IATLANG).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is a quantitative analysis of existing corpora and behavioral data.
Research sample	N/A
Sampling strategy	N/A
Data collection	N/A
Timing	N/A
Data exclusions	In the primary analysis of the Project Implicit data, 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, Banaji, 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 Supplemental Methods and Supplemental Figure 1).
Non-participation	N/A
Randomization	N/A.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☐	Antibodies
☐	☐	Eukaryotic cell lines
☐	☐	Palaeontology
☐	☐	Animals and other organisms
☐	☐	Human research participants
☐	☐	Clinical data

Methods

n/a	Involved in the study	
☐	☐	ChIP-seq
☐	☐	Flow cytometry
☐	☐	MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

- [ ] Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- [ ] Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

- [ ] The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- [ ] The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- [ ] All plots are contour plots with outliers or pseudocolor plots.
- [ ] A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

- [ ] Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*

Design specifications | *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures | *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) | *Specify: functional, structural, diffusion, perfusion.*

Field strength | *Specify in Tesla*

Sequence & imaging parameters | *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition | *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software | *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization | *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template | *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal | *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring | *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings | *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested | *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction | *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity | *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis | *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis | *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*