Language use shapes cultural stereotypes: Large scale evidence from gender

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

[1] University of Wisconsin-Madison

[2] University of Chicago

Author Note

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

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

Abstract

this is an abstract

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

Word count: X

Language use shapes cultural stereotypes: Large scale evidence from gender

Introduction

By the time they are two years old, children have already begun to acquire the gender stereotypes in their culture (Gelman, Taylor, Nguyen, Leaper, & Bigler, 2004). This finding is concerning given evidence that children's conceptualizations of gender can have undesirable real-world consequences. For example, in one study, girls, compared to boys, were less likely to think that girls are "brilliant" and also less likely to choose activities that were described as for "children who are very, very smart" (Bian, Leslie, & Cimpian, 2017). In the aggregate, these behavioral choices could lead to the observed lower rates of female participation in science, technology, engineering, and mathematics (STEM) fields (Ceci & Williams, 2011; Leslie, Cimpian, Meyer, & Freeland, 2015; Miller, Eagly, & Linn, 2015; Stoet & Geary, 2018). Given the potential downstream consequences of cultural stereotypes, it is therefore important to understand their origins.

While there are likely many factors that shape cultural stereotypes, a large body of evidence suggests that language may play an important causal role. For instances, young children perform worse in a game if they are told that an anonymous member of the opposite gender performed better than they did on a previous round (Rhodes & Brickman, 2008), or merely told that the game is *associated* with a particular gender (Cimpian, Mu, & Erickson, 2012). Further, there is evidence that descriptions as minimal as a single word can influence children's stereotypes: In one study, children were more likely to infer that a novel skill is stereotypical of a gender if the skill was introduced to children with a generic as opposed a non-generic subject ("[Girls are/There is a girl who is] really good at a game called 'gorp"'; Cimpian & Markman, 2011).

In our work here, we ask whether language influences psychological gender biases in ways that are more subtle than linguistic descriptions or word choice. In particular, we test

two possible mechanisms through which language might shape gender biases. The first is through word meaning derived from a word's co-occurrence with other words, an approach to word meaning known as *distributional semantics*. Under this approach, words that tend to occur in similar contexts in language may lead speakers to assume—either implicitly or explicitly—that they have similar meanings. For example, statistically, the word "nurse" occurs in many of the same contexts as the pronoun "her," providing an implicit link between these two concepts that may lead to a bias to assume that nurses are female. This route may be particularly influential because the bias is encoded in language in a way that is implicit and thus may be more difficult to discount. To our knowledge, no work to date has tested the link between distributional semantics and social stereotypes.

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

Past experimental work suggests a causal link between grammatical gender and psychological gender bias in both adults (e.g., Phillips & Boroditsky, 2003) and children (e.g., Sera, Berge, & Castillo Pintado, 1994). For example, Phillips and Boroditsky (2003) asked Spanish-English and German-English adult bilinguals to make similarity judgments between pairs of pictures depicting an object with a natural gender (e.g., a bride) and one without (e.g., a toaster). They found that participants rated pairs as more similar when the pictures matched in grammatical gender in their native language. While these types of studies provide suggestive evidence for a causal link between language and psychological gender bias,

they are limited by the fact that they typically only compare speakers of 2-3 different languages and measure bias in a way that is subject to demand characteristics.

Of course, evidence for a close correspondence between language and psychological gender biases is also consistent with the simpler explanation that language reflects a pre-existing gender bias in its speakers (*language-as-reflection hypothesis*). We assume that the language-as-reflection hypothesis is true to some extent: some of the ways we talk about gender reflect our knowledge and biases acquired independently of language. For example, we may observe that most nurses are women, and therefore be more likely to use a female pronoun to refer to a nurse of an unknown gender. Our present goal is to understand the extent to which language may also exert a causal influence on conceptualizations of gender (*language-as-causal hypothesis*).

In what follows, we ask whether the way gender is linguistically encoded across 26 different languages predicts cross-cultural variability in a specific manifestation of a gender bias—the bias to associate men with careers and women with family. We begin by describing our cross-cultural dataset of psychological gender bias. We then examine whether variability in language, as captured by distributional semantics, predicts differences in the gender-career implicit bias (Study 1). We next ask whether the presence of explicit marking of gender on nouns referring to people in a language is associated with greater implicit gender bias (Study 2). Together, our data suggest that language not only reflects existing biases, but plays a causal role in shaping culturally-specific notions of gender.

Description of Cross-Cultural Dataset of Psychological Gender Bias

To quantify cross-cultural gender bias, we used data from a large-scale administration of an Implicit Association Task (IAT; Greenwald, McGhee, & Schwartz, 1998) by Project Implicit (https://implicit.harvard.edu/implicit/; Nosek, Banaji, & Greenwald, 2002). The

IAT measures the strength of respondents' implicit associations between two pairs of concepts (e.g., male-career/female-family vs. male-family/female-career) accessed via words (e.g., "man," "business"). The underlying assumption of the IAT is that words denoting more similar meanings should be easier to pair together compared to more dissimilar pairs.

Meanings are paired in the task by assigning them to the same response keys in a two-alternative forced-choice categorization task. In the critical blocks of the task, meanings are assigned to keys in a way that is either bias-congruent (i.e. Key A = male/career; Key B = female/family) or bias-incongruent (i.e. Key A = male/family; Key B = female/career). Participants are then presented with a word related to one of the four concepts and asked to classify it as quickly as possible. Slower reaction times in the bias-incongruent blocks relative to the bias-congruent blocks are interpreted as indicating an implicit association between the corresponding concepts (i.e. a bias to associate male with career and female with family).

In the present study, we analyzed a dataset of gender-career IAT scores collected by Project Implicit between 2005 and 2016. We restricted our sample based on participants' reaction times and error rates using the same criteria described in Nosek, Banjai, and Greenwald (2002, pg. 104). We only analyzed data for countries that had complete demographic and IAT data for least 400 participants (2% of respondents did not give responses to the explicit bias question). This cutoff was arbitrary but the pattern of findings reported here holds for a range of minimum participant values (see SM). Our final sample included 764,520 participants from 39 countries, with a median of 1,311 participants per country. Note that although the respondents were from largely non-English speaking countries, the IAT was conducted in English. We do not have language background data from the participants, but we assume that most respondents from non-English speaking countries were native speakers of the dominant language of the country and L2 speakers of English.

Several measures have been used in the literature to quantify the strength of the bias from participants' responses on congruent and incongruent blocks on the IAT. Here, we used

the most robust measure, D-score, which measures the difference between critical blocks for each participant while controlling for individual differences in response time (Greenwald, Nosek, & Banaji, 2003). In addition to the implicit measure, we also analyzed an explicit measure of gender bias. After completing the IAT, participants were asked, "How strongly do you associate the following with males and females?" for both the words "career" and "family." Participants indicated their response on a Likert scale ranging from *female* (1) to *male* (7). We calculated an explicit gender bias score for each participant as the Career response minus the Family response, such that greater values indicate a greater bias to associate males with career.

At the participant level, implicit bias scores were correlated with participant age such that older participants tended to have a larger gender bias than younger participants (r = 0.06, $p < .0001$). Male participants ($M = 0.32$, $SD = 0.39$) had a significantly smaller implicit gender bias than female participants ($M = 0.42$, $SD = 0.36$; t = 105.60, $p < .0001$), a pattern consistent with previous findings (Nosek et al., 2002). Finally, implicit bias scores varied as a function of block order on the IAT task (t = -114.08, $p < .0001$).

For the present purposes, our goal was to estimate gender bias at the country level. To account for covariates of gender 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: $M = 0.00$; $SD = 0.17$). Implicit gender biases were moderately correlated with explicit gender biases at the level of participants ($r = 0.16$, $p < .0001$) but not countries ($r = 0.26$, $p = 0.10$).

We compared our residual country-level implicit and explicit gender biases to a gender equality metric reported by the United Nations Educational, Scientific and Cultural Organization (UNESCO) for each country: the percentage of women among STEM

graduates in tertiary education from 2012 to 2017 (Miller et al., 2015, available here: http://data.uis.unesco.org/; Stoet & Geary, 2018). These data were available for 33 out of 39 of the countries in our sample. 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 a smaller gender bias tended to have more women in STEM fields (r = -0.59, $p < .001$). 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 = 0.08, $p = 0.65$). In addition, we found a strong correlation between the median age of each country's population (as reported by the CIA factbook, 2017) and the residual implicit bias (for which participant age was controlled for): Countries with older populations tended to have larger gender biases ($r = 0.63$, $p < .0001$).

In sum, we replicate previously-reported patterns of gender bias in the gender-career IAT literature, with roughly comparable effect sizes (c.f. Nosek, et al., 2002). The weak correlation between explicit and implicit measures is consistent with claims that these two measures tap into different cognitive constructs (Forscher et al., 2016). In addition, we find that an objective measure of gender equality—female enrollment in STEM fields—is associated with implicit gender bias.

Study 1: Gender bias and semantics

Are participants' implicit and explicit gender biases predictable from biases found in the semantic structure of their native languages? For example, are the semantics of the words "woman" and "family" more similar in Spanish than in English? Both the language-as-reflection and language-as-causal hypotheses predict a positive correlation between the measured biases and biases present in language.

As a model of word meanings, we use large-scale distributional semantics models

derived from auto-encoding neural networks trained on large text corpora. The underlying assumption of these models is that the meaning of a word can be described by the words it tends to co-occur with—words occurring in similar contexts, tend to have similar meanings (Firth, 1957). The word like "dog", for example is represented as more similar to "hound" than to "banana" because "dog" co-occurs with words more in common with "hound" than "banana."Recent developments in machine learning allow the idea of distributional semantics to be implemented in a way that takes into account many features of language structure while remaining computationally tractable. The best known of these word embedding models is *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013). The model takes as input a corpus of text and outputs a vector for each word corresponding to its semantics. From these vectors, we can derive a measure of the semantic similarity between two words by taking the distance between their vectors (e.g., cosine distance).

As it turns out, many of the biases previously reported using implicit association tests can be predicted from distributional semantics models like word2vec. Caliskan, Bryson, and Narayanan (2017; henceforth *CBN*) measured the distance in vector space between the words presented to participants in the IAT task. CBN found that these distance measures were highly correlated with reaction times in the behavioral IAT task. For example, CBN find a bias to associate males with career and females with family in the career-gender IAT, suggesting that the biases measured by the IAT are also found in the lexical semantics of natural language.

CBN only measured semantic biases in English, however. In Study 1, we use the method described by CBN to examine whether the gender bias of the participants in the Project Implicit dataset is correlated with the gender bias measured in the dominant languages spoken in the countries of these participants. We begin by validating word embedding measures of gender bias by comparing them to explicit human judgements of word genderness (Study 1a). We then apply this method to models trained on text in other

languages (Study 1b). To foreshadow the results, we find that the implicit gender biases reported in Study 1 for individual countries are correlated with the biases found in the distributional semantics of the language spoken in the countries of the participants.

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

To validate word embeddings as a measure of psychological gender bias, we asked whether words that were closely associated with males in the word embedding models tended to be rated by human participants as being more male biased. We found human and word-embedding estimates of gender bias to be highly correlated.

Methods. 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; Scott, Keitel, Becirspahic, Yao, & Sereno, 2018). We compared these gender norms to estimates of gender bias obtained from embedding models pre-trained on two different corpora of English text: Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2016) and subtitles from movies and TV shows (Lison & Tiedemann, 2016; Van Paridon & Thompson, in prep.). The Wikipedia corpus is a large, naturalistic corpus of written language trained using the fastText algorithm (a variant of word2vec; Bojanowski et al., 2016; Joulin, Grave, Bojanowski, & Mikolov, 2016); The subtitle corpus is a smaller corpus of spoken language, trained using the XX algorithm.

To calculate a gender score from the word embeddings, for each word we calculated the average cosine distance to a standard set of male "anchor" words: ("male", "man", "he", "boy", "his", "him", "son", "brother") and the average cosine similarity to a set of female words ("female", "woman", "she", "girl", "hers", "her", "daughter", "sister"). A gender score for each word was then obtained by taking the difference of the similarity estimates (mean male similarity - mean female similarity), such that larger values indicated a stronger

association with males. There were 4,671 words in total that overlapped between the word-embedding models and human ratings.

Results and Discussion. 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; $r_{\text{subtitles}}$ = 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.

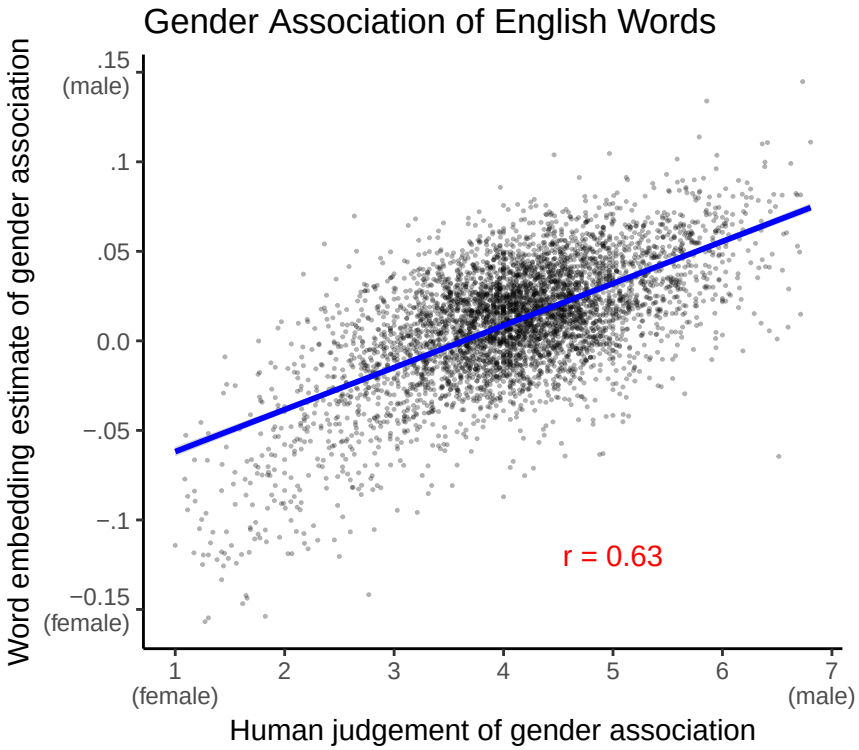

Figure 1. Word estimates of gender bias from the Subtitle-trained embedding model as a function of human judgments of gender bias (Study 1a). Each point corresponds to a word. Larger numbers indicate stronger association with males. Blue line shows linear fit and the error band corresponds to a standard error (too small to be visible).

Study 1b: Gender bias across languages

Having validated our method, we next turn toward examining the relationship between psychological and linguistic gender biases. In Study 1b, we estimate the magnitude of the gender-career bias in the dominant language spoken in the countries of the Project Implicit participants and compare it with estimates of behavioral gender bias from the Project Implicit data set.

Methods. For each country represented in our analysis of the Project Implicit, we identified the most frequently spoken language in each country using Ethnologue (Simons & (eds.)., 2018). This included a total of 26 unique languages. For each language, we then obtained translations from native speakers for the stimuli in the Project Implicit gender-career IAT behavioral task (Nosek et al., 2002) 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;" Nosek et al., 2002). Our linguistic stimuli were therefore a set of 8 female and 8 male Target Words (identical to Study 1a), a set of 8 Attribute Words associated with the concept "career" ("career," "executive," "management," "professional," "corporation," "salary," "office," "business") and 8 Attribute Words associated with the concept "family" ("family," "home," "parents," "children," "cousins," "marriage," "wedding," "relatives"). For one language, Tagalog, we were unable to obtain translations from a native speaker, and so translations were gathered from several translations sources. All analyses remain the same when this language is excluded [check this].

We used these translations to calculate a gender bias effect size from word embedding models trained on text in each language. Our effect size measure is a standardized difference score of the relative similarity of the target words to the target attributes (i.e. relative

similarity of male to career vs. relative similarity of female to career). Our effect size measure is identical to that used by CBN with an exception for grammatically gendered languages (see SM for replication of CBN on our corpora). Namely, for languages with grammatically gendered Attribute Words (e.g., niñas for female children in Spanish), we calculated the relationship between target words and attribute words of the same gender (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 behavioral effect size from the Project Implicit data, larger values indicate larger gender bias.

We calculated gender bias estimates from word-embedding models that had been trained on texts Wikipedia (Bojanowski et al., 2016) and subtitles from movies and TV shows (Lison & Tiedemann, 2016, as in Study 1a; Van Paridon & Thompson, in prep.) in each of our target languages. 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 (see SM for more details). This lead 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, Tagalog, and Zulu). Our final sample included 25 languages in total ($N_{\text{Wikipedia}} = 25$; $N_{\text{Subtitle}} = 20$), representing 9 different language families. We then compared estimates of linguistic gender bias for each language from each model to the behavioral IAT gender bias estimated from Project Implicit, averaging across countries whose participants speak the same language.

As before, we included two country level variables in our analysis, percentage of women in STEM fields and median country age. To obtain language-level estimates of these variables, we took the mean across countries whose participants speak the same primary

language.

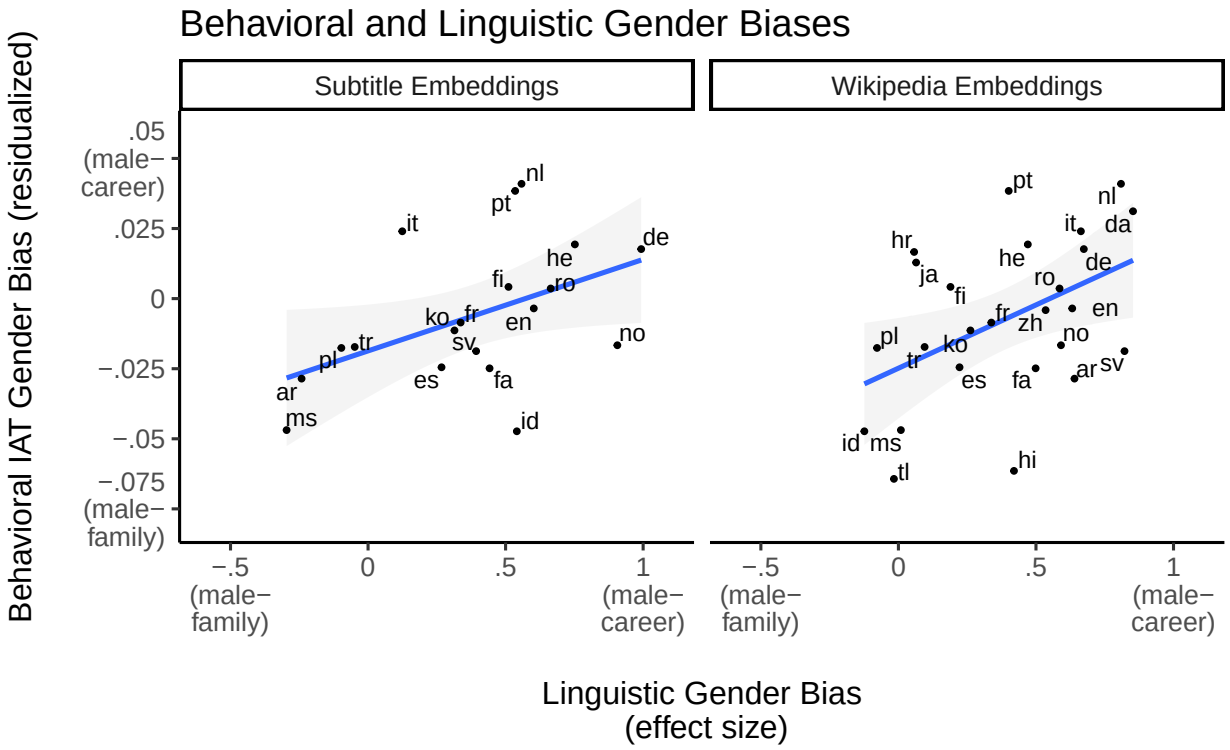

Figure 2. Residualized behavioral IAT gender bias as a function of linguistic gender bias, with each point corresponding to a language (Study 1b). Linguistic biases are estimated from models trained on text in each language from a subtitle corpus (left) and a Wikipedia corpus (right). Larger values indicate a larger bias to associate men with the concept of career and men with the concept of family. Error bands indicate standard error of the model estimate.

Results. There was no overall difference in the estimates of gender bias between models trained on the subtitle corpora versus the Wikipedia corpora ($t(19)$ = -0.06, p = 0.95). We next asked about the relationship between estimates of gender bias for each language and implicit gender bias of participants from countries where that language was dominant (and, we assume, was the native language of most of these individuals). Implicit gender bias were positively correlated with estimates of language bias from both Subtitles (r = 0.54, p = 0.01) and Wikipedia (r = 0.47, p = 0.02; Fig. 2, Table 1 shows the language-level correlations between all variables). The relationship between behavioral bias

and language bias remained reliable after partialling out the effect of median country age (Subtitles: $r = 0.46$, $p = 0.02$; Wikipedia: $r = 0.41$, $p = 0.04$). Linguistic gender bias was not correlated with explicit gender bias (Subtitles: $r = -0.06$, $p = 0.8$; Wikipedia: $r = 0.31$, $p = 0.13$). Finally, estimates of language bias from the Subtitles corpus were correlated with the objective measure of gender equality, percentage of women in STEM fields ($r = -0.55$, $p = 0.02$), though this relationship was not reliable for the Wikipedia corpus ($r = -0.19$, $p = 0.4$).

Discussion. In Study 1, we demonstrate that a well-studied psychological gender bias – the bias to associate men with career and women with family – is correlated with the magnitude of that same bias as measured in language statistics across a sample of 25 languages. Our pattern of findings is consistent with the possibility language statistics play a causal role in shaping psychological biases in indiviual speakers (language-as-causal). However, this pattern of findings is also consistent with the possibility that people's implicit biases are in part caused by input from language and that the linguistic biases simply reflect existing biases (language-as-reflection). In Study 2, we more directly investigate the causal mechanism linking language statistics to psychological biases.

Study 2: Gender bias and lexicalized gender

If language statistics play a causal role in shaping psychological gender biases, we predicted that those statistics would be influenced by the presence of explictly-marked gender distinctions referring to people. Languages make gender distinctions on words referring to people in order to indicate the biological sex of the referent, as in "waiter" versus "waitress" in English. We hypothesized that the presence of these kinds of distinctions might *lead to* more gender biased language statistics for those words. This prediction is a stronger test of the language-as-causal hypothesis because these gender markings are fossilized as part of the lexicon, and thus less likely to be caused by people's gender biases.

Languages can mark gender distinctions on words referring to people either as part of a grammatical gender system, as in Spanish (nurse: "enfermero" versus "enfermera"), or through idiosyncratic marking on particular nouns (e.g., "waiter" versus "waitress" in English). Languages that have grammatical gender systems make gender distinctions more frequently, since it is an obligatory part of the grammar. Further, languages that mark gender on the noun tend to mark gender on other arguments in the sentence as part of an agreement system ("el enfermo alto" (the tall $\text{nurse}_{\text{male}}$) versus "la enferma alta" (the tall $\text{nurse}_{\text{female}}$)), potentially making the reference to biological sex even more salient to speakers.

Thus, in Study 2, we asked whether languages that make gender distinctions more often on words referring to people tend to have more biased language statistics for those words. In this cases, gender marking highlights the gender of the person and thereby might exaggerate the gender biases present in the language. To test this possibility, we used word embedding models to examine the semantic associates that emerge from language statistics for a set words referring to occupations (e.g., "nurse" and "waiter"). We hypothesized that words referring to occupations that contained lexicalized gender distinctions would be more likely to have biased language statistics. To the extent that language statistics are causally related to people's implicit biases, we also hypothesized that speakers of languages with more biased language statistics for these words would also have larger overall psychological gender biases, as measured by the IAT.

Method. We identified 20 words referring to occupations that were relatively high-frequency and balanced for their perceptions of gender bias in the workforce, on the basis of previously-collected gender perception norms (Misersky et al., 2014). We then translated these words into each of the 26 languages in our sample, distinguishing between male and female forms where present. Forms were translated by consulting native speakers or English translation dictionaries.

To estimate the extent to which a language lexically encoded gender, we calculated the

proportion of male and female forms that were identical for each item, and then averaged across items within each language. This measure reflects the degree to which a language marks gender lexically, with larger values indicating more gender-neutral forms. We also estimated the extent to which each occupation word was gender biased in its language statistics using word embedding models trained. This measure was obtained using the same procedure as in Study 1a and 1b based on models trained on both subtitle and Wikipedia corora in each language. Larger values indicate greater gender bias (larger difference between associations to females vs. males). [Note here the measure is female - male, where as male - female in 1a above -> should maybe change this so consistent]. We compared these measures to the psychological gender measures described in Study 1b (residualized implicit association bias effect size and explicit bias score).

Results. Languages with higher degree of overlap in forms for male and female referents tended to have speakers with lower psychological gender bias (M = 0.66, SD = 0.36; r = -0.56, $p < .01$), even after partialling out the effect of median country age (r = -0.45, p = 0.02; Table 1; Figure 3).

Next, we asked whether a shared form for male and female referents for a particular occupation was associated with less gender bias in the statistics of use for that word. We fit a mixed effect model predicting degree of gender bias in language statistics (estimated from word embedding models) as a function of degree of overlap between male and female forms for that word, with random intercepts and slopes by language. Degree of form overlap was a strong predictor of language statistic for models trained on both the subtitle corpus (β = -0.59; SE = 0.07; t = -8.72) and Wikipedia corpus (β = -0.81; SE = 0.09; t = -9.48), with words with shared male and female forms tending to have less gender bias. This relationship also held at the level of languages, with languages with higher degrees of form overlap tending to have less gender bias in language statistics (Subtitle corpus: r = -0.75, $p < .01$; Wikipedia corpus: r = -0.66, $p < .01$).

Finally, we examined the relationship between gender bias in language statistics and psychological gender bias at the level of languages. Gender bias in language statistics were a strong predictor of behavioral gender bias for both language models (Subtitle corpus: $r = 0.62$, $p < .01$; Wikipedia corpus: $r = 0.54$, $p = 0.01$), and remained reliable after partialling out the effect of median country age (Subtitle corpus: $r = 0.54$, $p = 0.01$; Wikipedia corpus: $r = 0.46$, $p = 0.02$; Figure 4). To examine the relative predictive power of gender bias in language statistics and proportion form overlap, we fit a linear regression predicting behavioral gender bias with both measures, controlling for median country age. [NEITHER IS PREDICTIVE - try mediation model?]. Bias in language statistics did not reliably predict explicit gender bias.

Table 1

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

	Language IAT (Subtitles)	Language IAT (Wikipedia)	Residualized Behavioral IAT	Residualized Explicit Bias	Percent Women in STEM	Median Country Age	Occupation Bias (Subtitles)	Occupation Bias (Wikipedia)	Prop. Gender-Neutral Labels
Residualized Behavioral IAT	.51* (.45*)	.35+ (.22)		.1 (.23)	.59** (.5*)	.62**	.62** (.54**)	.54** (.46*)	.56** (.45*)
Residualized Explicit Bias	.03 (.01)	.14 (.2)	.1 (.23)		.09 (.05)	.13	.28 (.35+)	.35+ (.41*)	.05 (.11)
Percent Women in STEM	.55* (.51**)	.09 (.02)	.59** (.5*)	.09 (.05)		.37+	.39 (.29)	.31 (.22)	.32 (.21)
Language IAT (Subtitles)		.48* (.43*)	.51* (.45*)	.03 (.01)	.55* (.51**)	.27	.42+ (.36+)	.39+ (.34+)	.26 (.18)
Language IAT (Wikipedia)	.48* (.43*)		.35+ (.22)	.14 (.2)	.09 (.02)	.3	.28 (.19)	.47* (.42*)	.22 (.12)
Prop. Gender-Neutral Labels	.26 (.18)	.22 (.12)	.56** (.45*)	.05 (.11)	.32 (.21)	.38+	.75** (.71**)	.66** (.61**)	
Occupation Bias (Subtitles)	.42+ (.36+)	.28 (.19)	.62** (.54**)	.28 (.35+)	.39 (.29)	.36		.8** (.77**)	.75** (.71**)
Occupation Bias (Wikipedia)	.39+ (.34+)	.47* (.42*)	.54** (.46*)	.35+ (.41*)	.31 (.22)	.31	.8** (.77**)		.66** (.61**)
Median Country Age	.27	.3	.62**	.13	.37+		.36	.31	.38+

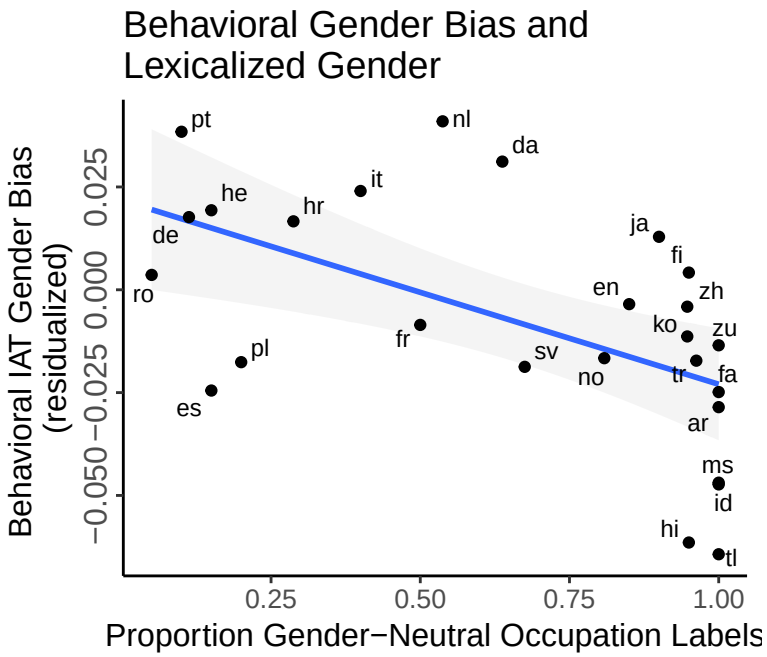

Figure 3. Residualized behavioral IAT gender bias as a function of the proportion of gender-netural labels for set of words referring to occupations. Error bands indicate standard error of the model estimate.

TO SORT OUT:

- Look at native vs. dictionary tranlsated languages
- remove low frequency words - based on google hits?
- deal with missing zh and ja translations for occupations, and others
- try mediation analysis?

General Discussion

- IAT is only behavioral in a weak sense
- socialization in development
- implicit vs. explicit difference
- pragmatics related to presense of distinction for kids - must be relevenat! - chesnut

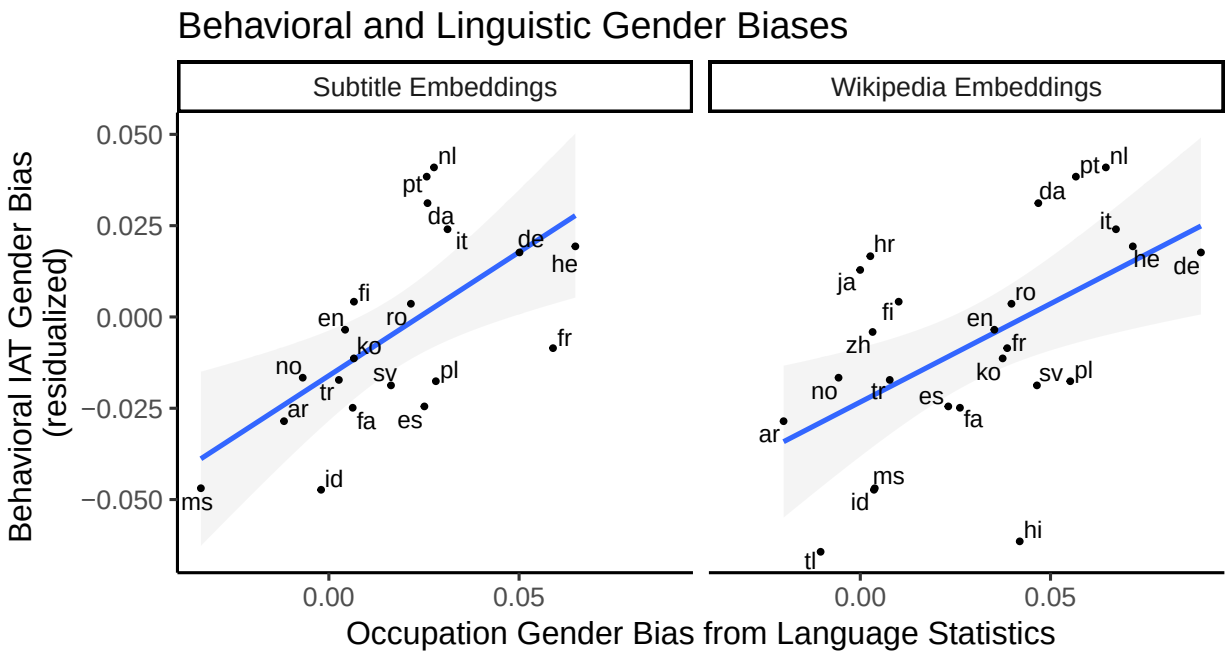

Figure 4. Residualized behavioral IAT gender bias as a function of mean gender bias of words referring to occupations, with each point corresponding to a language (Study 2). Linguistic biases are estimated from models trained on text in each language from a subtitle corpus (left) and a Wikipedia corpus (right).

stuff

- gender asymmetries in the IAT (women show bigger effect than men?)
- experimental work to get at causality more so

References

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391.

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

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 201014871.

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

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

Cimpian, A., Mu, Y., & Erickson, L. C. (2012). Who is good at this game? Linking an activity to a social category undermines children's achievement. *Psychological Science*, *23*(5), 533–541.

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

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

Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A.

(2016). A meta-analysis of change in implicit bias.

Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C., & Bigler, R. S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development*, i–142.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv Preprint arXiv:1607.01759*.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

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 (lrec 2016)*.

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

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of*

Educational Psychology, *107*(3), 631.

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., ... others. (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.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.

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

Rhodes, M., & Brickman, D. (2008). Preschoolers' responses to social comparisons involving relative failure. *Psychological Science*, *19*(10), 968–972.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 1–13.

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

Simons, G. F., & (eds.)., C. D. F. (Eds.). (2018). Ethnologue: Languages of the world. *Ethnologue: Languages of the World*.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593.

Van Paridon, J., & Thompson, B. (in prep.). Sub2Vec: Word embeddings from opensubtitles in 62 languages.