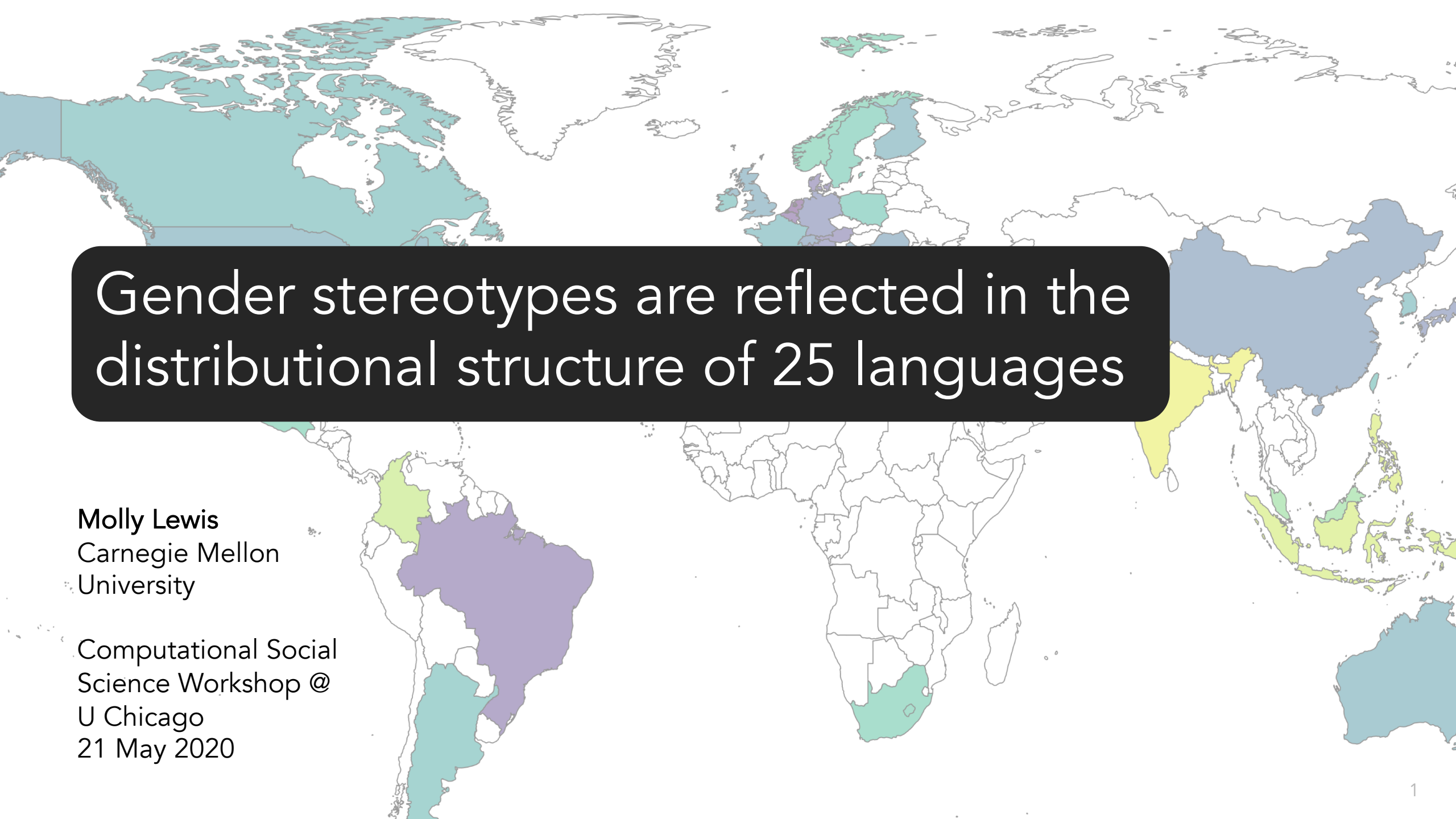


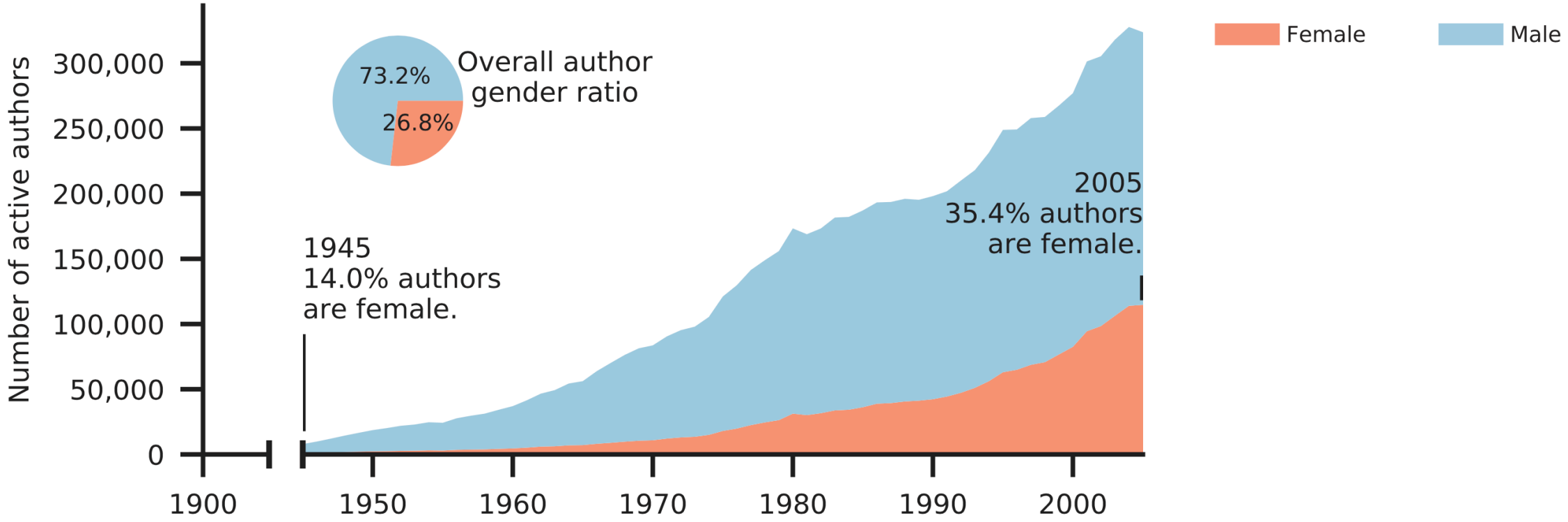
Gender stereotypes are reflected in the distributional structure of 25 languages

Molly Lewis
Carnegie Mellon University

Computational Social Science Workshop @ U Chicago
21 May 2020

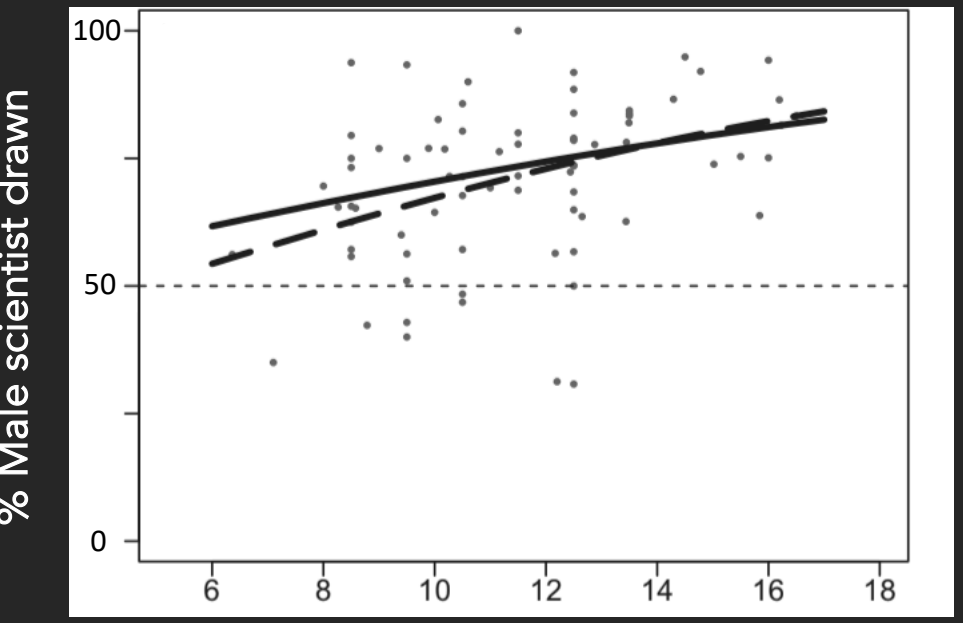
Gender disparities are pervasive

e.g., STEM fields



(Huang, et al., 2020)

Stereotypes develop in childhood…



(Miller et al, 2018)

Where do stereotypes come from?

Learn a lot from language:

The earth is round.

Mongolia is really cold.

Octopi have three hearts.

Boys are better at math than girls.

What about more **implicit messages** in language?

Semantic information from word co-occurrences

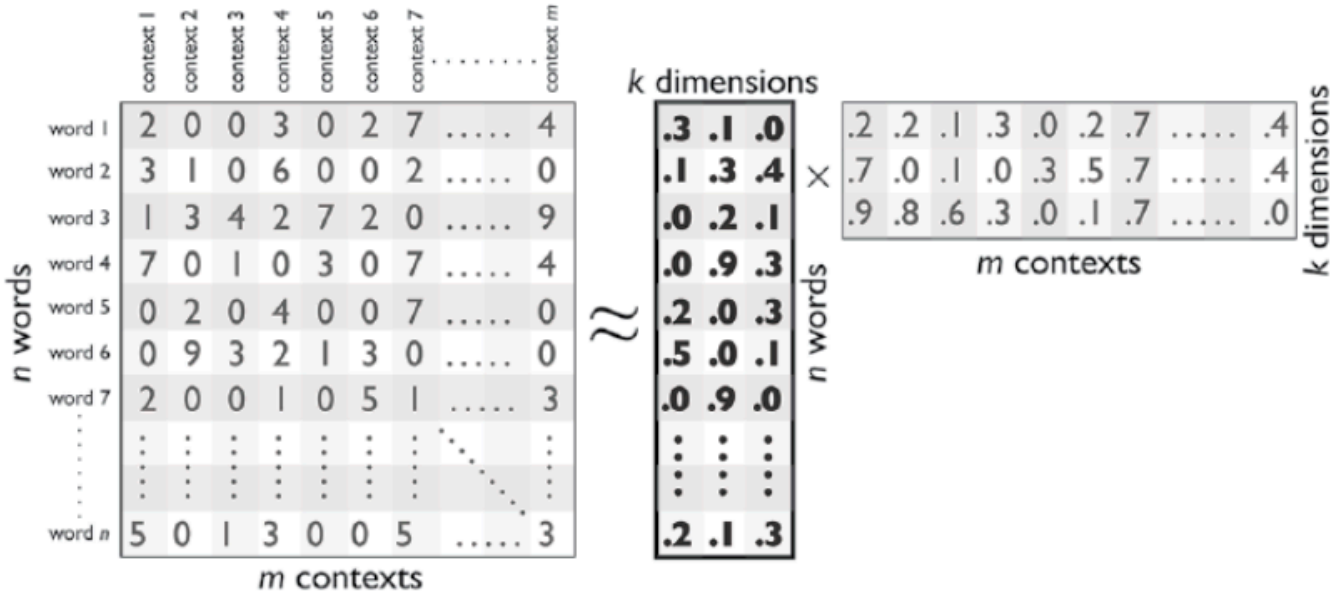
Distributional semantics: Semantic similarity between two words *A* and *B* is a function of the similarity of the linguistic contexts in which they appear.

Sam ate the red apple near the red barn...

	Sam	ate	the	red	apple	near	barn	...
Sam	0	1	0	0	0	0	0	
ate	1	0	1	0	0	0	0	
the	0	1	0	2	0	1	0	
red	0	0	2	0	1	0	1	
apple	0	0	0	1	0	1	0	
near	0	0	1	0	1	0	0	
barn	0	0	0	1	0	0	0	
⋮								

Reducing dimensionality of co-occurrence statistics extracts semantic information

- Represent all words from a corpus within a *k*-dimensional space
- Preserving distances between words in their local contexts
- Similar to factor analysis



k = 3

(Kozlowski, Taddy, & Evans, 2019)

Distributional models as *learning models*

Psychological Review
1997, Vol. 104, No. 2, 211–240

0033-295X/97/$3.00

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

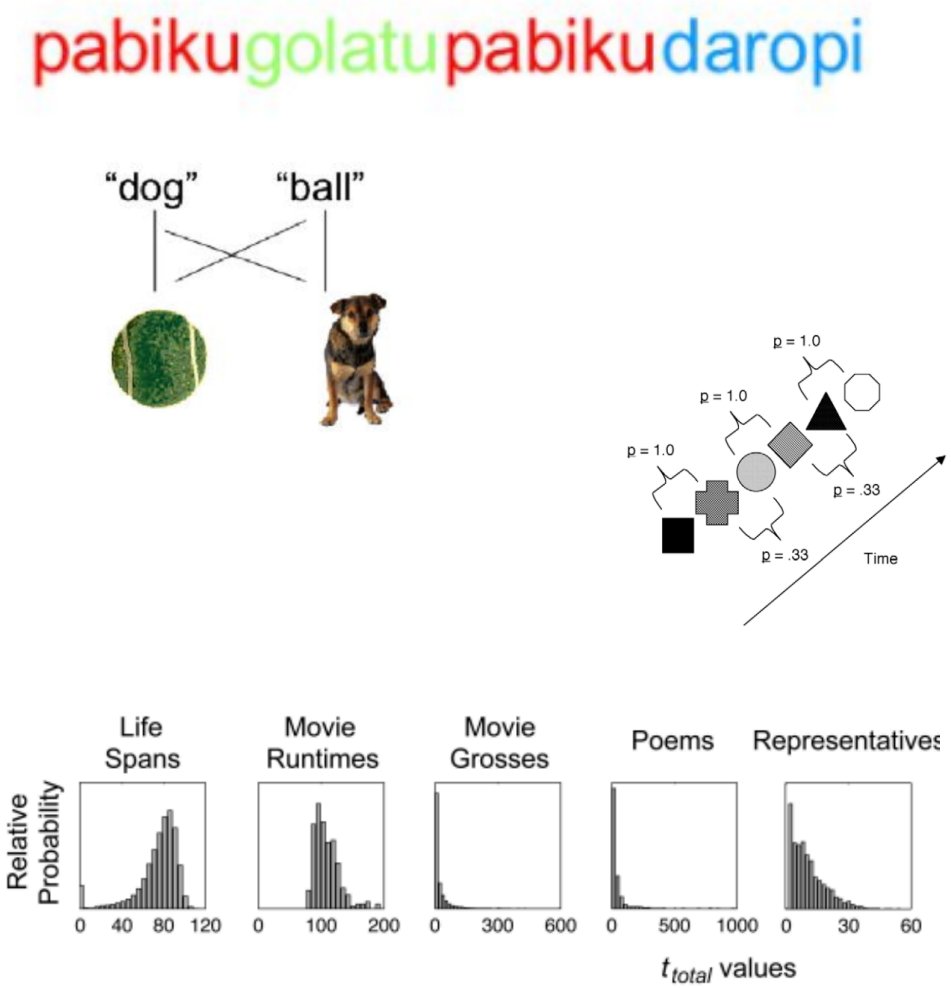
Susan T. Dumais
Bellcore

HAL (Lund & Burgess, 1996)
LSA (Landauer & Dumais, 1997)
Word2vec (Mikolov, Chen, Corrado, & Dean, 2013)
GloVe (Pennington, Socher, & Manning, 2014)
…

Cognitive Theory (Cognitive Science)

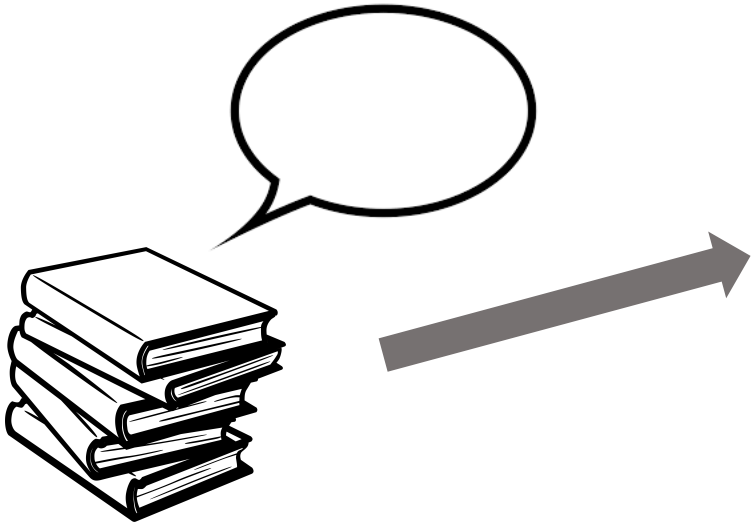
Solving language tasks (Machine Learning)

Humans are good at learning statistics



- Co-occurrence statistics to identify words (Saffran, Aslin, & Newport, 1996)
- Co-occurrence statistics to identify meanings (Smith & Yu, 2008)
- Co-occurrence statistics in the visual domain (Kirkham, Slemmer, & Johnson, 2002)
- Distributional statistics about everyday events (Griffiths & Tenenbaum, 2006)

Do humans learn social stereotypes by tracking distributional statistics in language?



	Sam	ate	the	red	apple	near	barn
Sam	0	1	0	0	0	0	0
ate	1	0	1	0	0	0	0
the	0	1	0	2	0	1	0
red	0	0	2	0	1	0	1
apple	0	0	0	1	0	1	0
near	0	0	1	0	1	0	0
barn	0	0	0	1	0	0	0

Gender stereotype

Men - career

Women - family

Implicit Association Test (IAT)

Categories

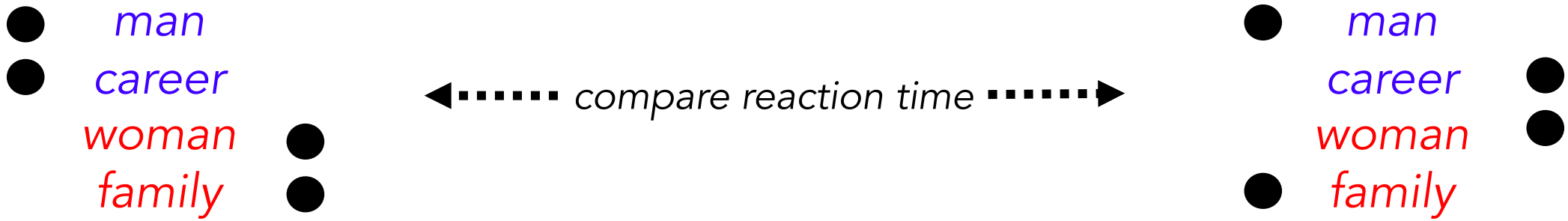
X = {man, male, he, him, boy}

Y = {woman, female, she, her, girl}

Attributes

A = {career, salary, office, business, professional}

B = {family, home, parents, children, cousins}



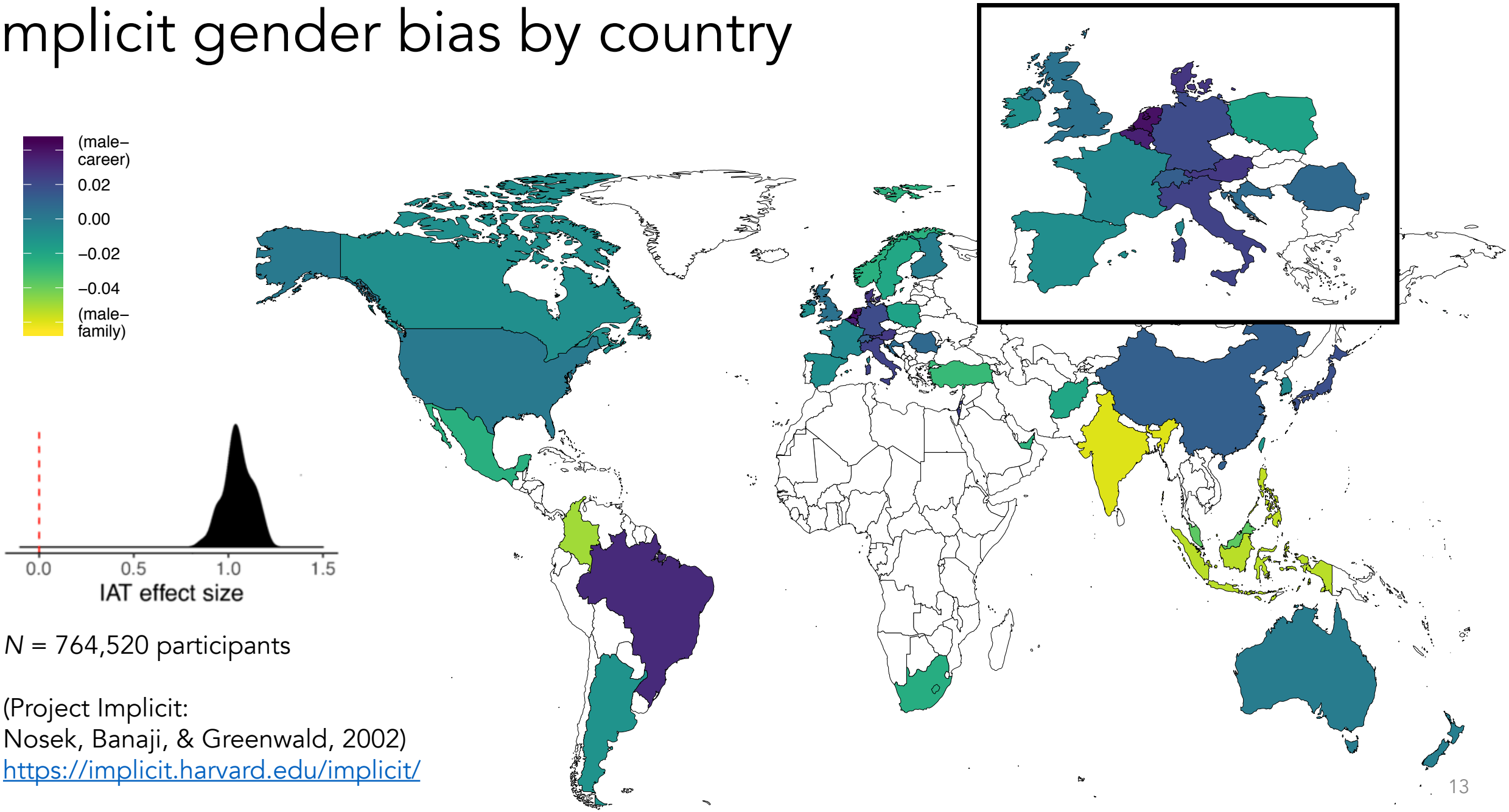
Participants slower for incongruent mapping (right), suggesting bias to associate men with career.

Quantifying the implicit bias: IAT effect size

Effect size = mean RT in incongruent condition – mean RT in congruent condition

Bigger effect size -> bigger bias

Implicit gender bias by country



N = 764,520 participants

(Project Implicit:
Nosek, Banaji, & Greenwald, 2002)
https://implicit.harvard.edu/implicit/

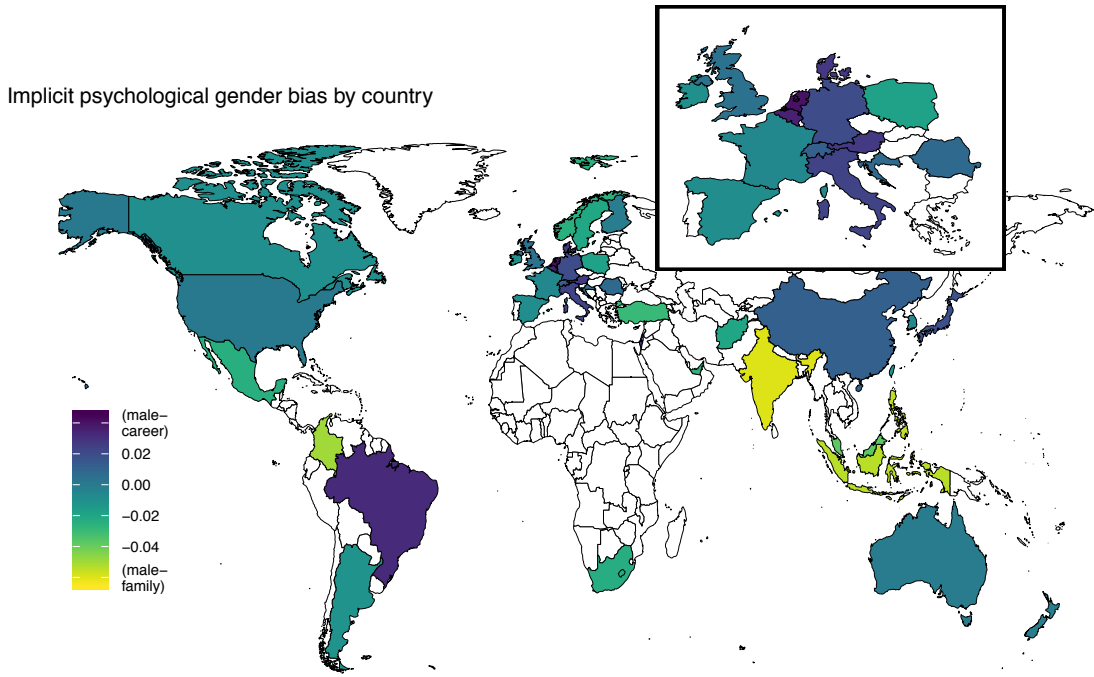
Does bias in language predict bias in IAT?

Language measure (word-occurrences)

Word embedding models trained on the 25 languages spoken by participants in the sample of countries

	Sam	ate	the	red	apple	near	barn
Sam	0	1	0	0	0	0	0
ate	1	0	1	0	0	0	0
the	0	1	0	2	0	1	0
red	0	0	2	0	1	0	1
apple	0	0	0	1	0	1	0
near	0	0	1	0	1	0	0
barn	0	0	0	1	0	0	0

Psychological measure (IAT)



Validating distributional statistics as encoding gender semantics

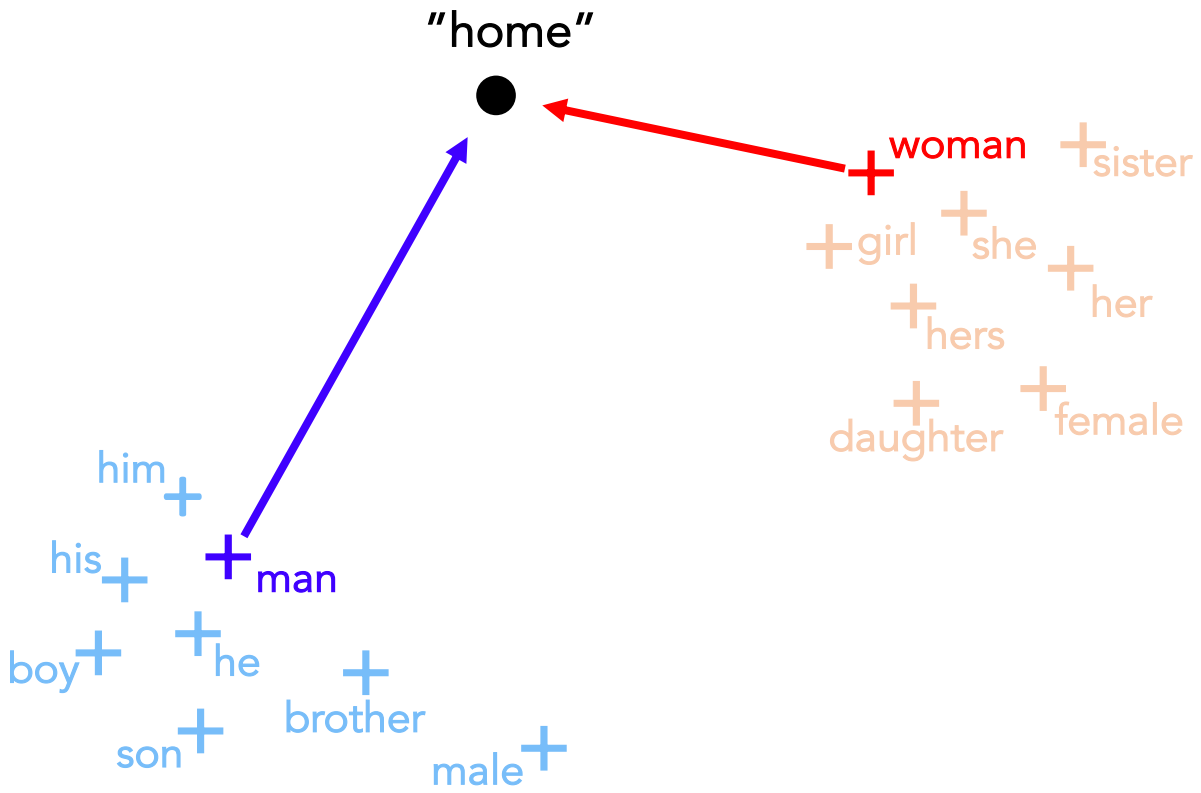
Measure gender bias from human judgements
(≈ 4,500 words; Scott et al., 2018)

Word embedding model trained on corpus of movie and TV subtitles in English (Lison & Tiedemann, 2016; Van Paridon & Thompson, in prep.).

Quantify gender association in semantic space as cosine distance

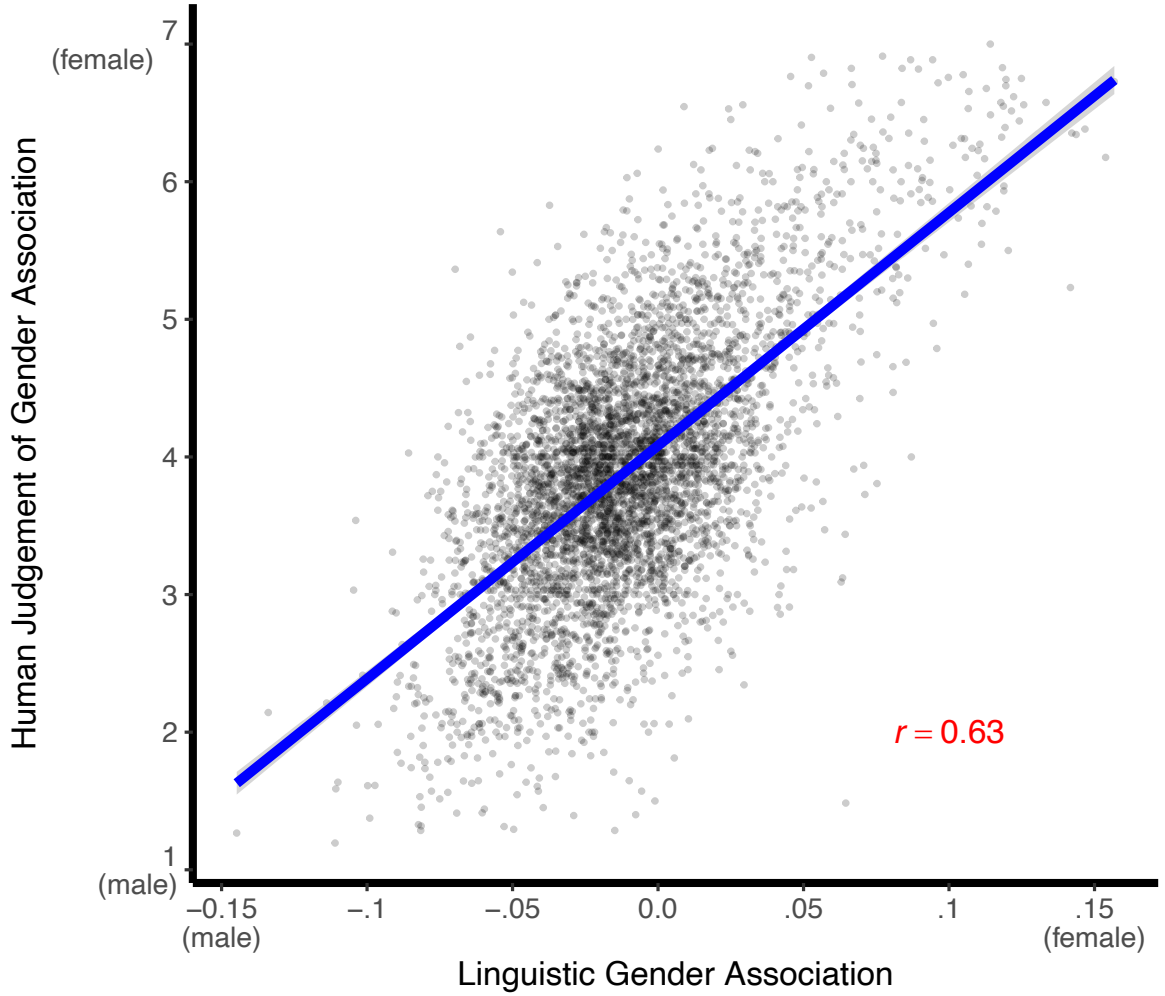
Rate the gender association of the word "**home**"

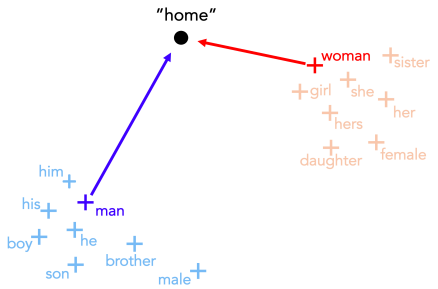
Very masculine ○ ○ ○ ○ ○ ○ ○ Very feminine



Rate the gender association of the word "home"

Very masculine ○ ○ ○ ○ ○ ○ ○ Very feminine





Replicated on model trained on English Wikipedia (r = .59)

Implicit Association Test (IAT)

…based on word co-occurrences

(using the same method as Caliskan, Bryson, & Narayanan, 2017)

Categories

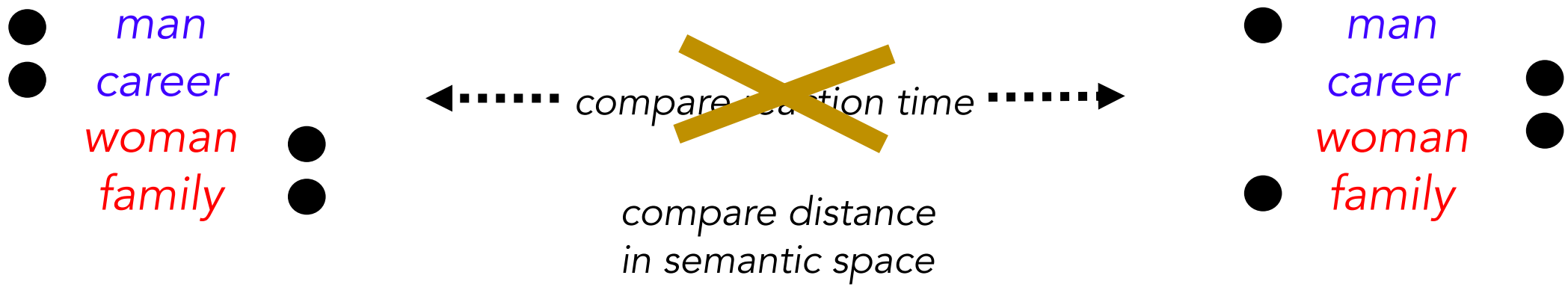
X = {man, male, he, him, boy}

Y = {woman, female, she, her, girl}

Attributes

A = {career, salary, office, business, professional}

B = {family, home, parents, children, cousins}



Language IAT effect size

** translated into each language*

Categories

*X** = {man, male, he, him, boy}

*Y** = {woman, female, she, her, girl}

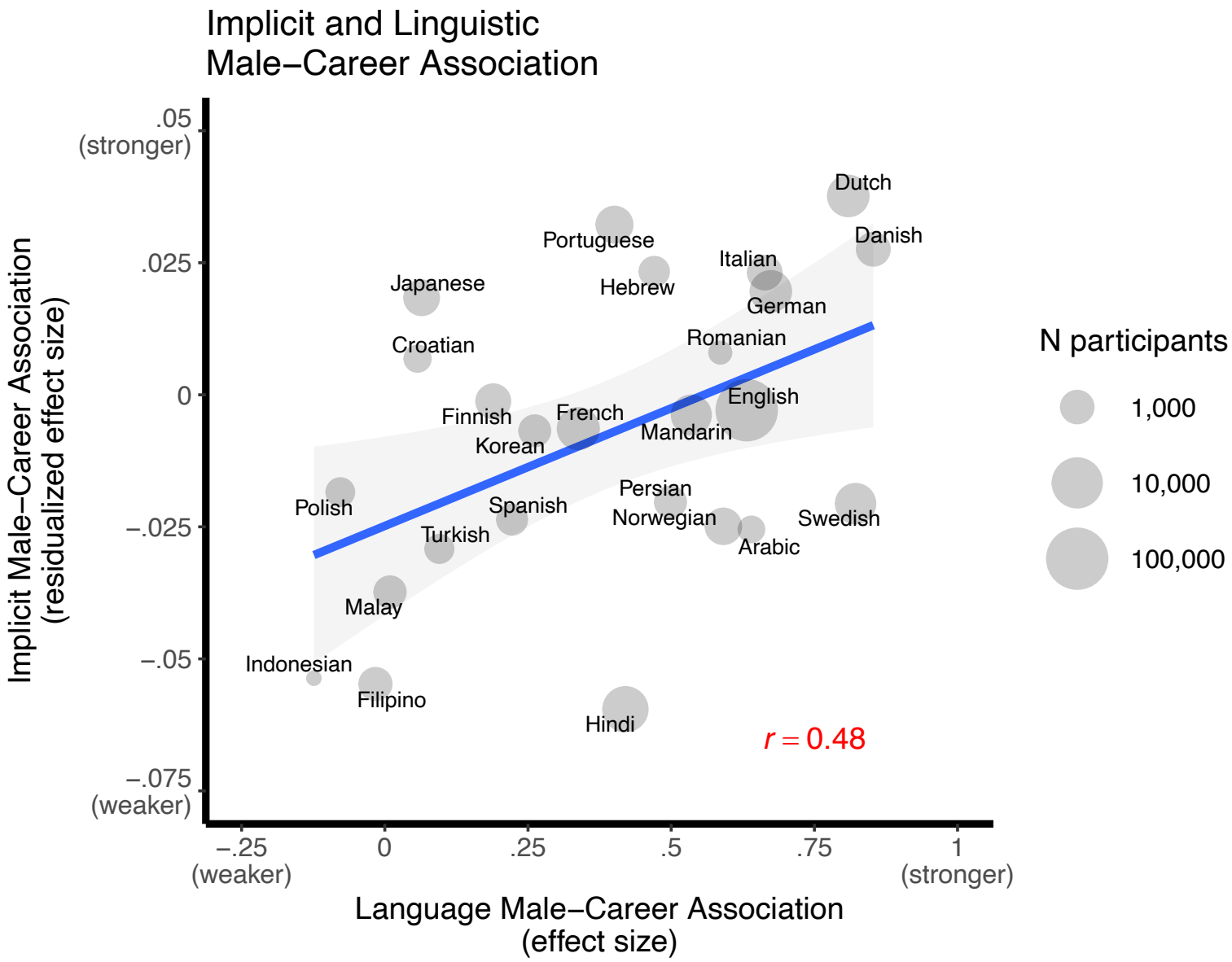
Attributes

*A** = {career, salary, office, business, professional}

*B** = {family, home, parents, children, cousins}

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

$$ES = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std\_ dev}_{w \in X \cup Y}\, s(w, A, B)}$$


Implicit and Linguistic
Male–Career Association
Implicit Male–Career Association
(residualized effect size)
.05
(stronger)
.025
0
−.025
−.05
−.075
(weaker)
−.25
(weaker)
0
.25
.5
.75
1
(stronger)
Language Male–Career Association
(effect size)
Dutch
Danish
Portuguese
Italian
Hebrew
Japanese
German
Romanian
Croatian
English
Finnish
French
Mandarin
Korean
Persian
Polish
Spanish
Norwegian
Swedish
Turkish
Arabic
Malay
Indonesian
Filipino
Hindi
r = 0.48
N participants
1,000
10,000
100,000

What about gender information encoded more explicitly in language?

Grammatical gender

"enfermero"

(nurse-MASC)

"enfermera"

(nurse-FEM)

Lexicalized gender

"waiter"

(waiter-MASC)

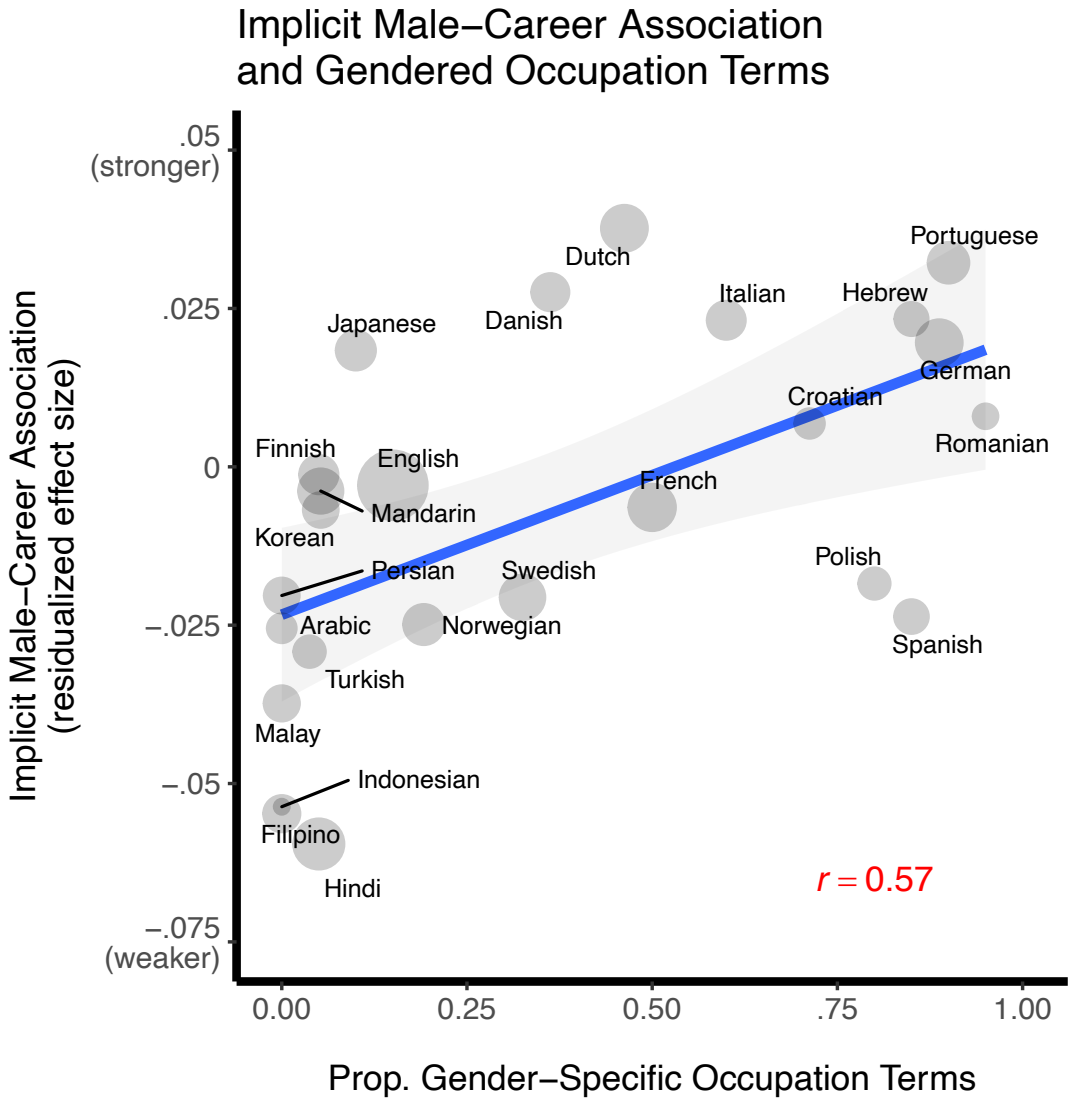
"waitress"

(waiter-FEM)

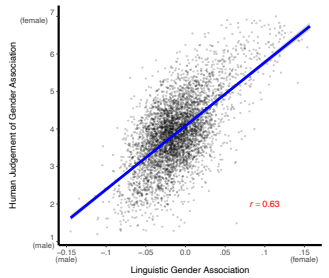
Explicit linguistic gender and implicit bias

- 12/25 languages encoded gender grammatical
- Languages that encoded gender grammatically did not have more implicit bias.
  (M_{diff} = 0.01 [-0.01, 0.03])

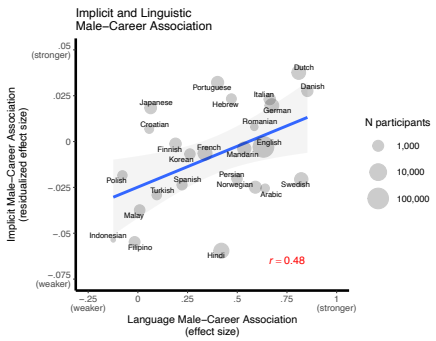
- In contrast, languages with more lexicalized gender forms have more implicit bias.



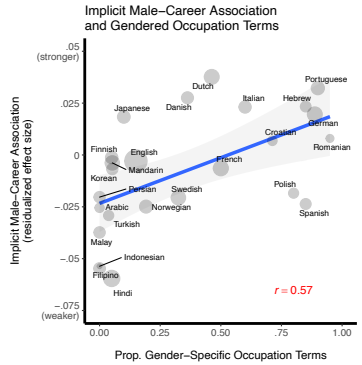
Summary



Statistical associations in language reflect human judgements of gender associations



Speakers of languages with stronger **statistical associations** between men-career and women-family tend to have a stronger bias when measured via the IAT



Speakers of languages with more **lexicalized gender distinctions** (but not grammatical gender) tend to have a stronger bias in the IAT.

Do humans learn social stereotypes by tracking distributional statistics in language?

Evidence for a correspondence between human semantic knowledge and distributional statistics (necessary but not sufficient)

Testing the causal question and other implications

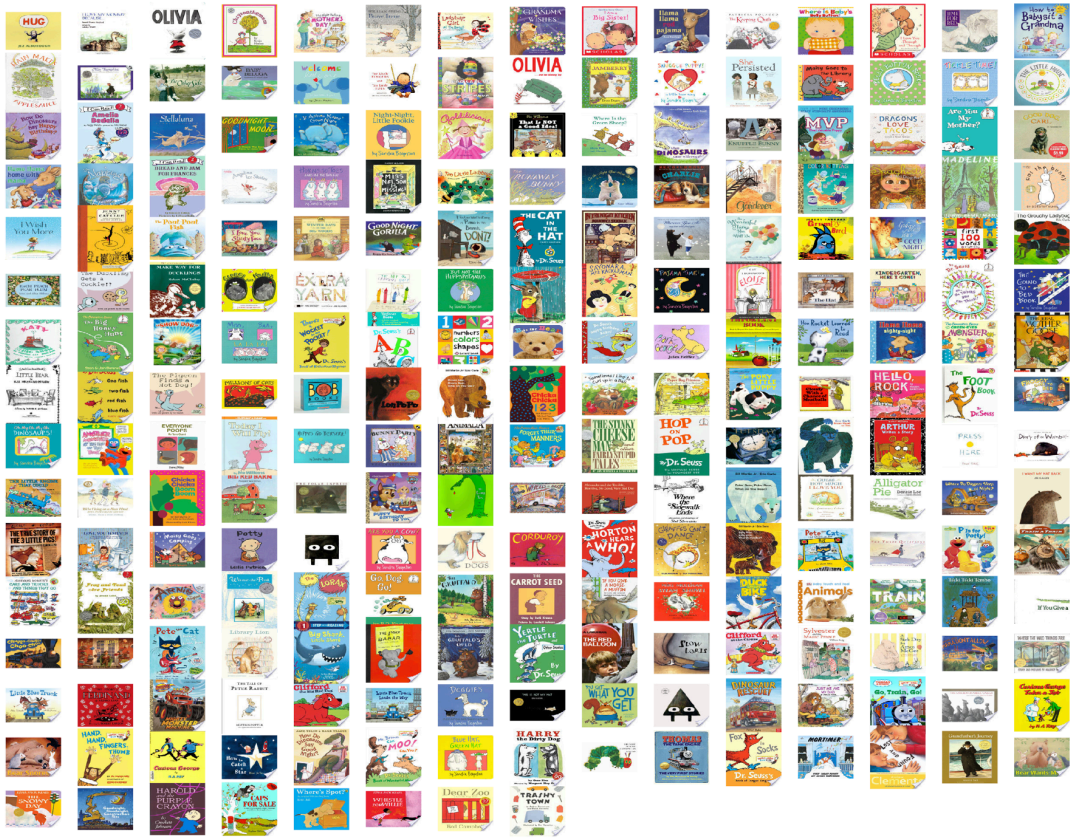
Is the link causal?

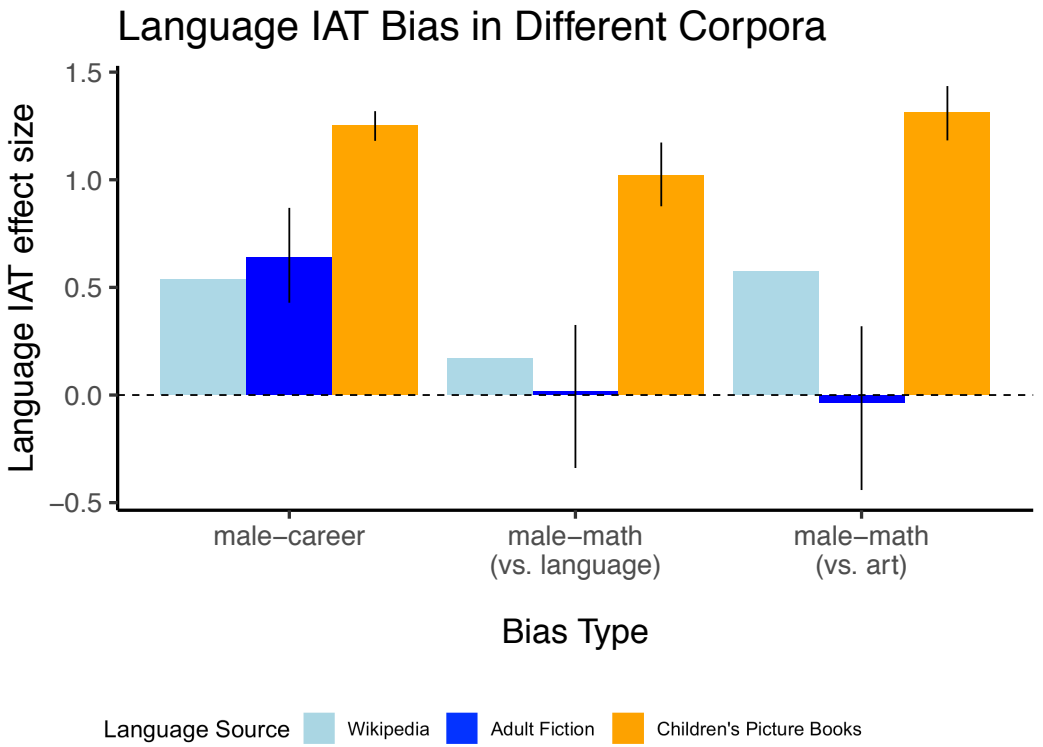
Distributional statistics → Human semantic representations

- All the evidence I've presented so far is correlational
- Likely bi-directional
- What kind of evidence might we bring to bear on this?
  - **Longitudinal analyses:** e.g., testing whether changes in language statistics predict or follow changes in measured implicit associations (Greenwald, 2017; Charlesworth & Banaji, 2019)
  - **Quasi-experimental tests:** e.g., measuring implicit associations in bilinguals using stimuli in languages that embed different linguistic associations
  - **Experimental designs:** measure the effect of manipulating language statistics on people's implicit associations.

Gender bias in children's books

If biases are learned from language, expect them to be present in the input to people who are learning the biases (i.e. children)





Language gender bias and other causal forces contributing to gender differences

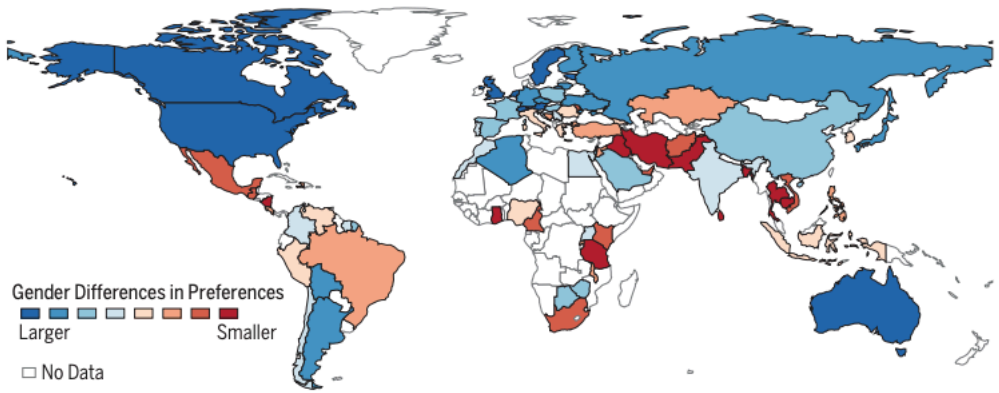
RESEARCH

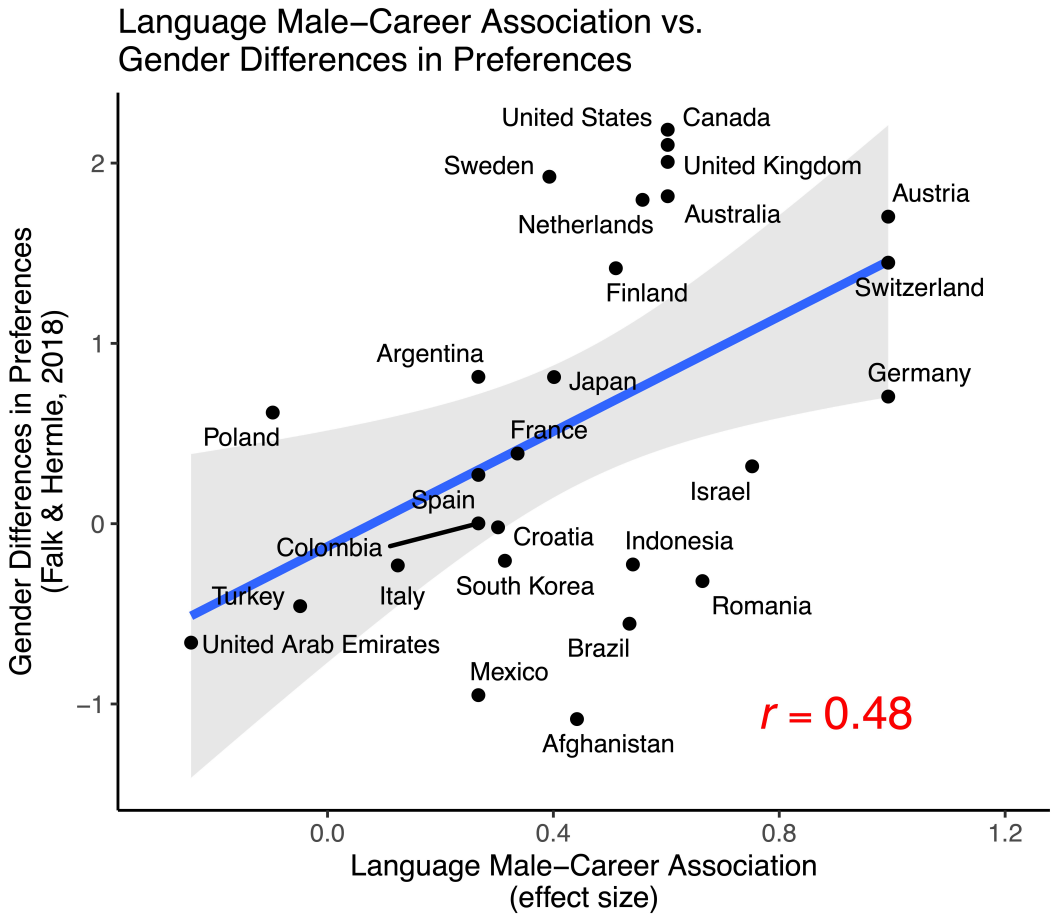
RESEARCH ARTICLE SUMMARY

SOCIAL SCIENCE

Relationship of gender differences in preferences to economic development and gender equality

Armin Falk* and Johannes Hermle*





Language gender bias and other causal forces contributing to gender differences

Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–13

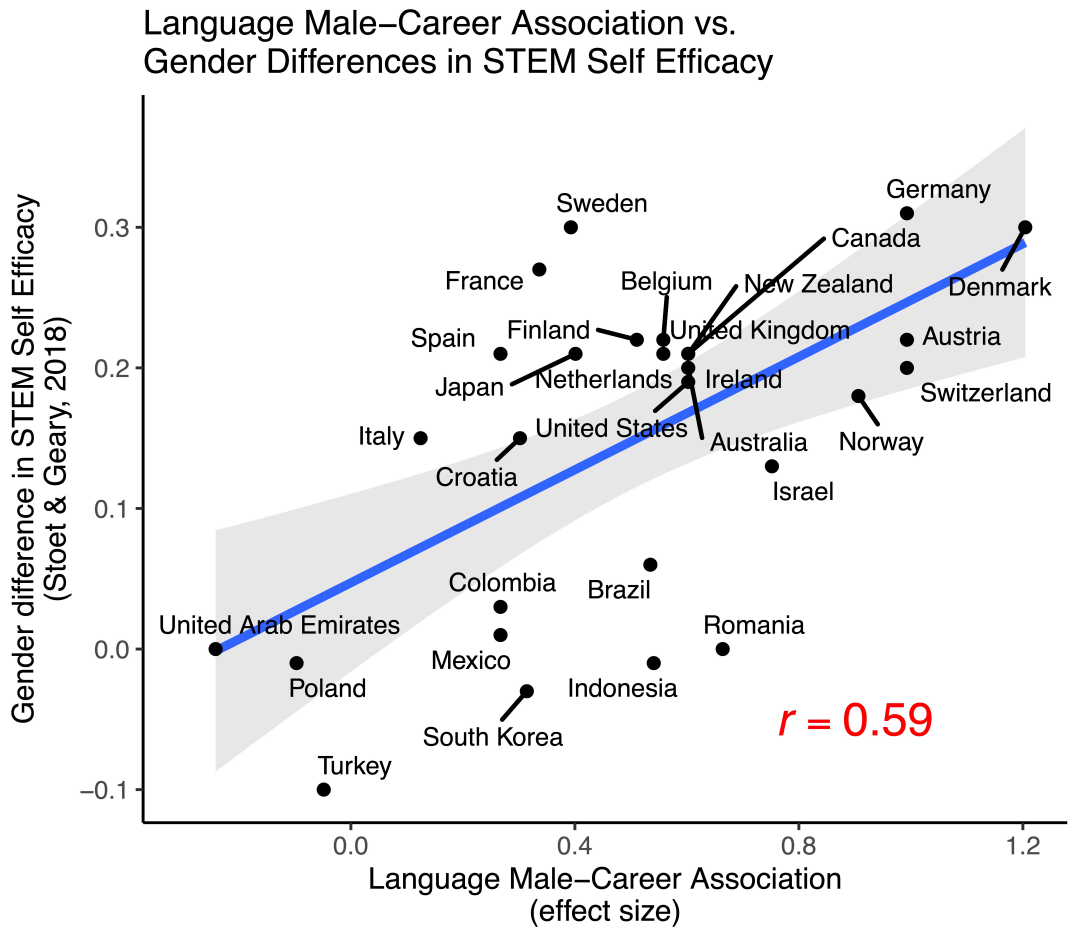
DOI: 10.1177/0956797617741719
www.psychologicalscience.org/PS
SAGE

The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education

Gijsbert Stoet[1] and David C. Geary[2]
[1]School of Social Sciences, Leeds Beckett University, and [2]Department of Psychological Sciences, University of Missouri

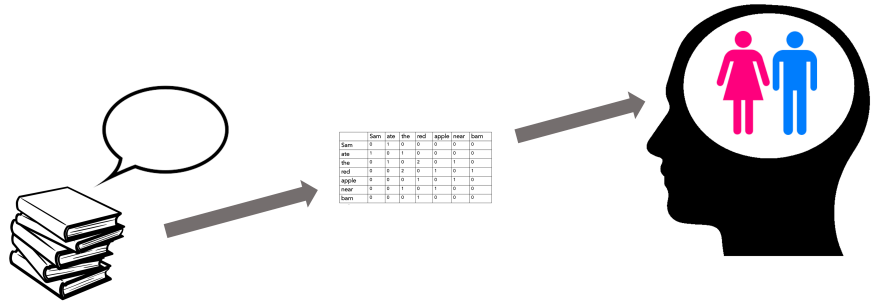
Students to report on how easy they thought it would be for them to:

- recognize the science question that underlies a newspaper report on a health issue
- explain why earthquakes occur more frequently in some areas than in others
- describe the role of antibiotics in the treatment of disease
- etc.



Do humans learn social stereotypes by tracking distributional statistics in language?

- Evidence for a close correspondence between human semantic knowledge and distributional statistics in the case of a particular stereotype: men-career; women-family



- Consistent with the idea that language is playing a causal role in shaping social stereotypes.
- Strongly correlated with other hypothesized “psychological” causal factors
- Additional work needs to be done to more directly test causality, and relationship to other causal forces
- Suggests that intervening on language input could change biases

Thanks!

Gary Lupyan
(U. of Wisconsin-Madison)

mollyllewis@gmail.com | mllewis | mollyllewis