Research Article

Psychological Science
2021, Vol. 32(2) 218–240

DOI: 10.1177/0956797620963619
www.psychologicalscience.org/PS

Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words

Tessa E. S. Charlesworth[1], Victor Yang[1], Thomas C. Mann[1], Benedek Kurdi[1,2], and Mahzarin R. Banaji[1]
[1]Department of Psychology, Harvard University, and [2]Department of Psychology, Yale University

Abstract
Stereotypes are associations between social groups and semantic attributes that are widely shared within societies. The spoken and written language of a society affords a unique way to measure the magnitude and prevalence of these widely shared collective representations. Here, we used *word embeddings* to systematically quantify gender stereotypes in language corpora that are unprecedented in size (65+ million words) and scope (child and adult conversations, books, movies, TV). Across corpora, gender stereotypes emerged consistently and robustly for both theoretically selected stereotypes (e.g., work–home) and comprehensive lists of more than 600 personality traits and more than 300 occupations. Despite underlying differences across language corpora (e.g., time periods, formats, age groups), results revealed the pervasiveness of gender stereotypes in every corpus. Using gender stereotypes as the focal issue, we unite 19th-century theories of collective representations and 21st-century evidence on implicit social cognition to understand the subtle yet persistent presence of collective representations in language.

Keywords
collective representations, gender stereotypes, machine learning, natural-language processing, word embeddings, open data, open materials

Received 10/21/19; Revision accepted 7/22/20

Psychological analyses of social-group stereotypes have most commonly asked participants to report on their own or other people's beliefs about social groups (e.g., indicate the degree to which *men–women* are associated with qualities of *agency–communion*). Across 50 years of research, variations of such questions have been the primary source of evidence about the presence and strength of stereotypes (Ellemers, 2018). Such methods are valuable indicators of individual, subjective reports of stereotypes, but they cannot reveal the presence and potency of stereotypes as *collective representations* (Durkheim, 1898/2009; Moscovici, 1988), the term used to refer to societal-level systems of meaning that pervade everyday social life.

To learn about the presence of stereotypes in social life, we must examine more natural expressions of human thought. Durkheim (1898/2009) and the social scientists who followed him argued that the primary place to seek such information is in the language of societal products (e.g., books, conversations, TV, movies, the Internet). Although Durkheim's view was ahead of its time, nobody could have conceived of the possibilities presented by today's computational approaches that rely on machine learning to analyze billions of words from the sociosphere of the Internet (e.g., Pennington, Socher, & Manning, 2014, analyzed 840 billion word tokens).

It is within the natural language of human conversations, books, and audiovisual media that the implicit

Corresponding Author:
Tessa E. S. Charlesworth, Harvard University, Department of Psychology
E-mail: tet371@g.harvard.edu

presence and potency of group stereotypes can be measured. Take, for example, innocuous child-directed statements such as "get mommy from the kitchen" or "daddy is still at the office." Such sentences do more than describe the physical locations or roles of mothers and fathers; they also reinforce attributes associated with those roles. That is, although the proximity between the words *mommy–kitchen* or *daddy–office* can describe the reality of gender-based roles, it also creates and perpetuates perceptions of the internal traits of individuals who occupy such roles (Eagly & Wood, 2012; Koenig & Eagly, 2014). When consistently expressed in spoken and written language, such perceptions can become collective "truths" that shape how children and adults think about and interact with the social world (e.g., Gaucher, Friesen, & Kay, 2011; Rhodes, Leslie, Yee, & Saunders, 2019).

But just how pervasive are stereotypes in natural language? Is there consistency even across language that varies in format (TV, movies, books, conversations), age groups, and time periods? To answer this question, we combined an unprecedented database of language corpora (65+ million words from seven corpora of child and adult text) with advances in natural-language processing (word embeddings) to quantify the prevalence of gender stereotypes in language.

We investigated language corpora that vary in theoretically meaningful ways, including (a) *format*, from ordinary conversations to books to audiovisual media; (b) *time*, from historical books to conversations of the late 20th century to contemporary TV and movies; and (c) *age groups*, from child to adult audiences and speakers. We also investigated gender stereotypes across domains, ranging from well-studied associations (e.g., women–home/men–work) to associations with more than 600 personality traits and more than 300 occupations. Given evidence that gender stereotypes vary in magnitude across age groups (e.g., Miller, Nolla, Eagly, & Uttal, 2018), time (e.g., Eagly, Nater, Miller, Kaufmann, & Sczesny, 2020), and domains (Martin & Ruble, 2010), the diversity of language sources and topics allows a rigorous test of the pervasiveness and potency of gender stereotypes.

Since Durkheim, scholars have argued that collective stereotypes are maintained through language that is subtle or indirect, often even more so than language that is explicit and direct (Moscovici, 2000). Echoes of such theories are heard in modern perspectives of implicit social cognition, which posit that indirectly assessed (implicit) stereotypes reflect and reveal collective, societal-level phenomena (Payne, Vuletich, & Lundberg, 2017) more so than directly assessed (explicit) measures. For instance, although elderly and young respondents show in-group preferences on explicit measures, both elderly and young respondents show consistent, anti-elderly/pro-young attitudes on implicit measures (Nosek et al., 2007). The only explanation is that implicit measures are particularly affected by societal-level representations of stigmatized groups, even overriding in-group preferences. Similarly, measuring the hidden, indirect structures of language (through word-embedding techniques elaborated below) is uniquely poised to reveal the collective stereotypes embedded in a society.

Statement of Relevance

Language permeates every aspect of our daily lives through conversations, books, TV, movies, and the Internet. A key role of language is to communicate social information, including stereotypes about social groups (e.g., which groups are delicate vs. strong or fast vs. slow). It is an intriguing aspect of social-group stereotypes that they are often hidden in plain sight; they are right there but rarely stated explicitly. In this research, we applied methods from natural-language processing (word embeddings) to systematically uncover and quantify the strength and prevalence of subtle gender stereotypes across child and adult language (conversations, books, TV, movies). Despite many differences across corpora (e.g., time periods, formats, age groups), gender stereotypes were surprisingly consistent and robust for widely studied stereotypes and lists of more than 600 traits and more than 300 occupations. The results underscore the pervasive and even obligatory role of language in sustaining stereotypes in mind and society.

The Present Research

With new data and methods, we have the opportunity to test whether large-scale natural language (from everyday conversations to formal writing) confirms the views from 19th-century theories and 21st-century research in implicit social cognition. If the present analyses show weak or inconsistent evidence of gender stereotypes (e.g., appearing only in some corpora, age groups, or time periods), we would be led to a limited view of gender stereotypes in language. If, on the other hand, evidence of gender stereotypes is strong across sources, we would conclude that language is a potent carrier of gender stereotypes and that it has a role in the propagation of collective representations.

Across three studies, we examined both well-studied stereotypes (e.g., women–home/men–work; Study 1)

and comprehensive lists of traits (Study 2) and occupations (Study 3) in diverse natural language of child-produced, child-directed, adult-produced, and adult-directed text. To identify stereotypes on this massive scale, we used *word embeddings* (see the Method section) to quantify the association between groups (male–female) and attributes (e.g., home–work; Caliskan, Bryson, & Narayanan, 2017). The idea underlying word embeddings is that patterns of word co-occurrences are modeled (with machine-learning algorithms) to quantify the semantic relationships between words (e.g., the semantic relationship between the words *women* and *home* vs. *men* and *home*). The method has already shown feasibility in documenting social-group representations, including gender biases (e.g., Caliskan et al., 2017; DeFranza, Mishra, & Mishra, 2020; Garg, Schiebinger, Jurafsky, & Zou, 2018; Lewis & Lupyan, 2020). However, such work has exclusively focused on adult-produced corpora and without direct comparisons across language sources (e.g., conversations vs. books). The current project thus offers the first test of consistency of gender stereotypes across child and adult natural language to quantify whether stereotypes are indeed widely shared collective representations.

Study 1: Well-Studied Gender Stereotypes

Study 1 was designed to investigate four well-studied gender associations: male–female with semantic attributes of home–work, arts–science, and math–reading and evaluative attributes of good–bad. Study 1 tested whether collective representations of gender exist in large-scale natural language, at consistent magnitudes across conversations, books, TV, and movies, from multiple decades and from both adults and children.

Method

Below, we describe the eight steps in our method procedure: Step 1, collect text data; Step 2, clean text data; Step 3, select the attributes to test (e.g., home–work, arts–science); Step 4, select word stimuli to represent categories and attributes; Step 5, generate word embeddings from text data using machine-learning algorithms; Step 6, perform the Word-Embedding Association Test (WEAT) and Single-Category WEAT (SC-WEAT) and calculate significance; Step 7, calculate meta-analytic estimates and meta-regressions to compare WEAT scores across age groups, sources, and domains; and Step 8, perform validation tests, including replicating the results with additional large-scale corpora and word-embedding algorithms. Our focus in this section is to provide a concise and accessible introduction to working with massive text data and word embeddings. We direct readers interested in the specifics to the Supplemental Material available online.

Step 1: collect text data. Child-produced and child-directed corpora were selected because they are, to our knowledge, the largest corpora of natural child-produced and child-directed language; the three adult corpora were subsequently chosen to best approximate the child corpora in data size, time period of data collection, breadth of topics, and linguistic style (e.g., dyadic speech, movie transcripts, or book text).

Child-produced and child-directed speech. Text of child-produced and child-directed speech (from parents and caregivers) was obtained from transcripts of English-language dyadic parent–child conversations documented through the Child Language Data Exchange System language corpus; most transcripts were collected between 1970 and 1990 (MacWhinney, 2000). Although these dates are historical (and therefore cannot provide insights into contemporary speech), the corpus remains an important product of study because (a) regardless of year, it can reveal whether children and adults from the same conversations are similarly communicating collective representations; (b) the corpus remains widely used to understand children's and adults' language; and (c) it is the largest known corpus of child speech. The prepared corpus consists of 6,518 conversations between children (age: $M = 2.92$ years, range = 0–12 years) and their caregivers, yielding 8,429,128 word tokens (i.e., individual words in the corpus, regardless of how many times the word is repeated).

Child-produced speech (i.e., child speaking to parent) and child-directed speech (i.e., parent speaking to child) were assessed independently by dividing the corpus according to whether the speaker was a child (indexed by a "CHI" tag in the corpus) or parent (indexed by a "MOT" or "FAT" tag, for mother or father, respectively). We therefore obtained two independent corpora with (a) utterances produced by child speakers (2,601,432 word tokens) and (b) utterances directed toward children by parents and caregivers (5,827,696 word tokens).

Child-directed books. Child-directed book text was retrieved from a subsample of English-language children's books obtained from Project Gutenberg, an open-source database of books (https://www.gutenberg.org/). This subsample of children's books was previously extracted from Project Gutenberg for machine-learning tests of language comprehension (Hill, Bordes, Chopra, & Weston, 2016). The current corpus consists of 98 books, published between 1820 (*The Legend of Sleepy Hollow* by Washington Irving) and 1922 (*Blacky the Crow* by Thornton Burgess),

and consists of 4,583,629 word tokens. Although these books are historical, we argue that they remain important societal-level products because they reflect classic texts that continue to be read by (and to) children, and they provide a comparison point against relatively more contemporary speech and audiovisual media to investigate possible influences of time period on the strength of gender stereotypes in language.

Child-directed audiovisual media. Transcripts from child-directed audiovisual media were retrieved from online transcripts, transcribed by volunteers, of English-language Disney movies, PBS Kids TV shows, and Nickelodeon TV shows, airing between approximately 1938 (Disney's *Snow White and the Seven Dwarfs*) and the present day (e.g., Disney's 2017 remake of *Beauty and the Beast*). The corpus of transcripts was created for this project and has been made available to other researchers at the project's OSF page (https://osf.io/kqux5/). The corpus consists of 1,078 movies, 4,309 TV episodes, and 6,747,208 word tokens.

Adult-produced speech. Adult-produced speech transcripts were retrieved from the Switchboard-1 Telephone Speech Corpus (Godfrey & Holliman, 1993), a database of English-language dyadic telephone conversations, recorded in 1990 and 1991, between 543 adult speakers (ages 17–68 years) on a set of 70 randomly selected topics. This corpus was chosen because it best approximated the size and time period of the child-produced and child-directed speech corpora (see above). The adult-produced speech corpus consists of approximately 2,400 conversations and 3,063,280 word tokens.

Adult-directed books. Texts of adult-directed books were obtained from a randomly selected subsample of 1,000 English-language books on Project Gutenberg. The subsample was determined using a random-number generator and including the text of the book indexed by each random number. The majority of texts in Project Gutenberg were published before 1923, matching the time period for children's books. Again, although these books are historical, they nevertheless continue to provide common cultural knowledge as well as a comparison point with more contemporary sources. The prepared adult-directed books corpus consists of 40,252,700 word tokens.

Adult-directed audiovisual media. Transcripts of adult-directed audiovisual media were retrieved from online transcripts, transcribed by volunteers, of popular English-language movies and TV shows for adult audiences across numerous broadcasting stations and production houses. The movies and TV shows aired between approximately the 1960s (e.g., *Doctor Who, The Addams Family*) and present day (e.g., *CSI, Breaking Bad*). Thus, the adult-directed audiovisual-media corpus matches the child-directed TV shows and movies corpus in time period (i.e., relatively more contemporary) and format (i.e., online transcripts provided by volunteer transcribers). As with the child-directed audiovisual corpus, the corpus of adult-directed audiovisual transcripts was created for this project and has been made available to other researchers at https://osf.io/kqux5/. The corpus consists of 2,056,384 word tokens.

Step 2: clean text data. Complete details on cleaning procedures for each corpus, including reproducible code and data, are available at OSF (https://osf.io/kqux5/) and described in the Supplemental Material. In brief, cleaning proceeded in two steps. First, all punctuation, meta-data (e.g., speaker gender, character name), and linguistic markings (e.g., notes about a speaker's tone) were removed from the text. Second, all words were "lemmatized," meaning that words were changed from any variant and inflection forms to their root form. For example, the words *running* and *ran* would be changed to the root form of *run* (for further details, see the glossary in the Supplemental Material). Reducing word variants to root forms increases the number of occurrences of each root word to improve the reliable computation of word embeddings (see below). In this case, lemmatization is particularly helpful because the target corpora are smaller than other natural-language corpora, such as the Common Crawl corpus (which we use for validation as described below) with more than 600 billion word tokens (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018).

Step 3: select categories and attributes to test. In Study 1, we aimed to provide a proof-of-concept test on whether the word-embedding method can replicate well-studied stereotypes in natural language. Only with such empirical grounding can we confidently extend the method to study consistency across more diverse stereotype topics and language corpora. Thus, in Study 1, we focused on gender stereotypes that have been robustly documented on both explicit and implicit measures and with both adults and children. Specifically, we examined the stereotypes of female–arts/male–science (for a review, see Charlesworth & Banaji, 2019), female–reading/male–math (e.g., Cvencek, Meltzoff, & Greenwald, 2011), and female–home/male–career (e.g., Croft, Schmader, Block, & Baron, 2014) and the attitude of female–good/male–bad (e.g., Dunham, Baron, & Banaji, 2016).

Although all four associations have been robustly documented, two of these associations warrant further discussion. First, the association of women–home/men–work may seem to challenge current evidence that, worldwide, approximately half of all women (52%)

Table 1. Word Stimuli Used to Represent Each Category and Attribute (Study 1)

Category	Word stimuli
Female	she, her, mommy, mom, girl, mother, lady, sister, mama, momma, sis, grandma, herself
Male	he, his, daddy, dad, boy, father, guy, brother, dada, papa, bro, grandpa, himself
Good	happiness, happy, fun, fantastic, lovable, magical, delight, joy, relaxing, honest, excited, laughter, lover, cheerful
Bad	torture, murder, abuse, wreck, die, disease, disaster, mourning, virus, killer, nightmare, stress, kill, death
Home	baby, house, home, wedding, kid, family, marry
Work	work, office, job, business, trade, activity, act, money
Art	art, dance, dancing, sing, singing, paint, painting, song, draw, drawing
Science	science, scientist, chemistry, physic, engineer, space, spaceship, astronaut, chemical, microscope
Reading	book, read, write, story, word, writing, reading, tale
Math	puzzle, number, count, math, counting, calculator, subtraction, addition

participate in the labor force (67% in the United States; 2019 estimates from The World Bank, 2020). Thus, women might be expected to be associated with the attribute *work*. However, the stereotype of women–home/men–work reflects the relative associations of women (vs. men) with work (vs. home). Thus, because men are more likely to participate in the labor force than women (75% of men worldwide and 68% of men in the United States), the relative association of men–work versus women–work will favor men–work. Additionally, because women are more likely to take on household responsibilities and more likely to participate in caregiving occupations even within the workforce (U.S. Bureau of Labor Statistics, 2019), the relative association of women–home versus men–home will favor women–home. Likely for these reasons, the women–home/men–work stereotype has been widely observed with typical psychological measures from both children and adults (Croft et al., 2014; Nosek et al., 2007) and is therefore also expected across the natural language of children and adults.

Second, the female–good/male–bad association may appear counterintuitive because higher status groups (in this case, men) are usually associated with "good" attributes. Nevertheless, psychologists have long documented a counterintuitive "women are wonderful" effect (Eagly & Mladinic, 1994). Indeed, contemporary data from both implicit and explicit measures reveal consistent evidence for a female–good/male–bad association among both children and adults (Dunham et al., 2016). In sum, previous research would predict that all four gender associations should be present, at least in adult language corpora. Whether the gender associations are also observed at consistent magnitudes across diverse child and adult corpora is the focus of the present study.

Step 4: select word stimuli to represent categories and attributes. As with any psychological experiment, a primary concern is how to best represent the construct of interest with both precision (i.e., low variance) and comprehensiveness (i.e., no obvious exclusions). This concern is also present when selecting the word stimuli to represent a given category or attribute in a word-embedding approach (e.g., selecting the words to represent *female*). Thus, to select word stimuli, we aimed to balance precision, comprehensiveness, and frequency of word occurrence to ensure that each category or attribute was accurately represented. Specifically, to select word stimuli, we began with the stimuli lists from Caliskan and colleagues (2017). If the stereotype was not tested in that study, we used the stimuli list from online Implicit Association Tests (IATs; http://implicit.harvard.edu). Next, we examined the frequency of these words in the child-produced speech corpus (the corpus least likely to include complex words). We retained those words that appeared in the child-produced speech corpus. Finally, we expanded the stimuli list by adding semantically related words (generated by the researchers) that were also present in the child-produced speech corpus. All final stimuli are reported in Table 1. Notably, we also performed a supplementary analysis to test the robustness of the results when using other (longer) word stimuli lists, obtained from a more recent application of the WEAT (DeFranza et al., 2020). All major conclusions held regardless of the choice of gender stimuli (see the Supplemental Material).

Step 5: create word-embedding vectors. To understand how word embeddings are created, it is useful to begin by imagining a "cloud" that represents all semantic meaning (formally, a high-dimensional semantic space). Each word in our language exists somewhere in this cloud of semantic meaning. To situate each word within this cloud, we can represent each word by a vector (a line that points in a specific direction). The goal of these vector representations is to represent words that are close in meaning (e.g., *mother* and *girl*) with vectors that point in similar directions and to represent words that are far in meaning (e.g., *mother* and *cactus*) with vectors that point in different directions. Projecting down into two-dimensional space, this would essentially

mean that words close in meaning have similar (*x, y*) coordinates and are therefore positioned close in space. A *word embedding* is the term for the vector representation of a word (the position of a word) within the cloud of semantic meaning.

Representing words as vectors (i.e., as word embeddings) is useful because one can use these numeric vectors in subsequent quantitative operations to understand the space of semantic meaning. In this case, we can use word embeddings to quantify the overlap in meaning between words (e.g., between *mother* and *girl*) by looking at the angle between the word-embedding vectors. Again, words that are close in meaning will have vectors pointing in similar directions and will therefore have a small angle between them and, consequently, a large cosine similarity (a measure of the strength of association between word vectors). Conversely, words far in meaning will have a large angle between them and a small cosine similarity. In the WEAT (described below), we used these cosine similarities as the basis for identifying stereotypes (i.e., associations between groups and attributes) in language.

The general approach to create an embedding for a word is to iteratively calculate the best set of real numbers that situates the word in semantic space according to its semantic meaning, so that it is situated close to words similar in meaning. To achieve this optimal representation, the word-embedding algorithm (in this case, the fastText algorithm; Mikolov et al., 2018) uses the target word's surrounding context to try to predict the target word (e.g., predict *dog* from within the context "the brown *X* wagged its tail"). At first, the accuracy of predicting the target word is low because the algorithm has received little feedback on the types of word co-occurrences that are most informative of a word's meaning (e.g., it may also predict *bag*, rather than *dog*, in the context "the brown *X* wagged its tail"). However, with each iteration, the accuracy of the predictions increases until the algorithm "understands" the contexts and co-occurrences of the target word.

Notably, because word embeddings are trained on the specific word contexts and co-occurrences in a given corpus, the embedding for a target word in corpus A may be different from the embedding for the same word in corpus B (e.g., the embedding for *dog* in children's books may be different from the embedding for *dog* in adult speech). For this reason, word embeddings can be used to identify the strength of stereotypes within a given corpus and also test consistency across corpora. In short, word embeddings document the traces of societal-level collective representations. From this perspective, debates on whether and how word embeddings (and vector-space models, more generally) reflect the operation of individual human cognition and semantic memory (e.g., Günther, Rinaldi, & Marelli, 2019) are not particularly applicable here because we are primarily using this method as an index of societal (rather than individual) phenomena.

In practice, one can calculate word embeddings from a variety of algorithms, including two of the most widely used algorithms, fastText and GloVe (for definitions, see the glossary in the Supplemental Material). More recently, word-embedding algorithms have also expanded to incorporate sentence-level contextual information, such as with the advent of ELMo, BERT, and RoBERTa embeddings (see the Supplemental Material). In this project, we used the fastText algorithm, an improvement from the widely used word2vec algorithm, for all main analyses because (a) at the time of analysis, it was the highest performing algorithm for single-word embedding vector creation (Mikolov et al., 2018); (b) it was similar in approach to previous studies using the WEAT that rely on single-word embeddings (i.e., Caliskan et al., 2017); and (c) it allowed us to maintain focus on the theoretical contributions of our findings rather than introduce a new class of sentence-level contextualized approaches (e.g., ELMo, BERT). Notably, in subsequent validation analyses, we also ensured the robustness of the results by using the GloVe algorithm, and the results remained generally consistent across both fastText and GloVe approaches.

Step 6: WEAT. To transform the individual word-embedding vectors into an effect size of the strength of gender stereotypes, we used the WEAT, which has begun to be widely applied in understanding social psychological phenomena (e.g., DeFranza et al., 2020; Kurdi, Mann, Charlesworth, & Banaji, 2019). The WEAT computes a standardized effect-size measure of the relative association between words representing group categories (in this case, male–female) and words representing attributes (in this case, home–work, math–reading, arts–science, good–bad). The degree of association is measured from the cosine similarities between category and attribute word-embedding vectors (see above). Again, large cosine similarities indicate large overlap between word vectors.

For an example of the WEAT computation, take the WEAT effect-size calculation for the stereotypical association of women–home/men–work. In this example, we represent each group and attribute by four individual word vectors: *she, her, mommy*, and *mom* for *women; he, him, daddy*, and *dad* for *men; house, home, baby*, and *family* for *home*; and *work, job, business*, and *money* for *work*. The WEAT computation can be described in six general steps.

First, we computed the association (i.e., the cosine similarity) between an individual *women* word vector (*she*) and all individual *home* word vectors (*house, home, baby, family*). The individual *she* to *house, she* to *home, she* to *baby*, and *she* to *family* associations

were then averaged to provide a mean *she* to *home* cosine similarity.

Second, we computed the association between that same individual *women* word vector (*she)* and all *work* word vectors (*work, job, business, money*). Again, the individual associations were averaged to yield a *she*-to-*work* mean cosine similarity.

Third, we took the difference between the *she*-to-*home* and *she*-to-*work* mean cosine similarities, providing a difference score for the individual word vector *she*. These three initial steps were then repeated for the other three *women* word vectors (*her, mommy, mom*) to get four individual word difference scores (*she* to *home* vs. *work, her* to *home* vs. *work, mommy* to *home* vs. *work*, and *mom* to *home* vs. *work*).

Fourth, we took the mean of these four individual word difference scores to provide a mean group difference score across all group word vectors (*women* to *home* vs. *work*). Steps 1 through 4 were then repeated to calculate the mean group difference score for the opposite group category (*men* to *home* vs. *work*).

Fifth, we took the difference between the two mean group difference scores (*men* to *home* vs. *work* minus *women* to *home* vs. *work*). This provided a "double-difference" score that reflects the relative semantic similarity between the groups (male–female) and attributes (home–work). Sixth and finally, we took this double-difference score (*women–home* vs. *work* minus *men–home* vs. *work*) and divided it by the standard deviation across all eight individual word-vector difference scores computed in Step 3 (*she* to *home* vs. *work, her* to *home* vs. *work, him* to *home* vs. *work*, etc.). This yielded an effect size—the WEAT *D* score—that is analogous in interpretation to an IAT *D* score in that it is a double-difference score normalized by a measure of variance.

SC-WEAT. A limitation of the WEAT, as well as the traditional IAT, is that the computations collapse two associations into a single relative measure of association. For instance, the stereotypical association of women–home/men–work represents both the association of *home* with *female* (over *male*) and the association of *work* with *male* (over *female*). Thus, finding a significant relative WEAT effect may be driven by one association being very large (e.g., a strong home–female association), whereas the second association is relatively small (e.g., a weaker work–male association). Fortunately, the present word-embedding approach can easily overcome this limitation by decomposing the relative association into two single associations, using the SC-WEAT (see also the Word-Embedding Factual Association Test, reported by Caliskan et al., 2017, and the SC-WEAT, described by Kurdi et al., 2019). Thus, to inform interpretation of the relative effects reported in Study 1, we also performed the SC-WEAT test for all stereotype associations. All results are reported in Table S3 in the Supplemental Material and are summarized in the main results below.

The SC-WEAT computation followed the same general steps as the above WEAT computation, except that (a) we did not repeat the procedures for a second attribute because we were interested only in a single attribute, and therefore, (b) we did not calculate a double-difference score but, rather, stopped after calculating the single-difference score.

To make this concrete, take the female/male–home SC-WEAT computation. First, we computed the association (i.e., cosine similarity) between each individual word representing the single attribute (e.g., the word *home* for the attribute *home*) and all individual words representing the *women* group category (*she, her, mommy, mom*). These associations are then averaged to give a *home*-to-*women* mean cosine similarity. Second, we computed the association between that same word (*home*) and all individual words representing the *men* group category (*he, him, daddy, dad*). Again, these individual associations were averaged to give a *home*-to-*men* mean cosine similarity. Third, we took the difference between the *home*-to-*women* and *home*-to-*men* mean cosine similarities, providing a difference score for the individual word vector *home*. These three initial steps were then repeated for the other three *home* word vectors (*house, baby, family*) to get four individual word difference scores. Fourth, we took the average of these four individual word difference scores to get the mean cosine similarity of *home* to *women* versus *men*, yielding a single-difference score. Fifth and finally, we divided this single-difference score by the standard deviation across all individual word-vector difference scores computed in Step 3 (*home* to *women* vs. *men; house* to *women* vs. *men*, etc.). Again, this provided an effect size—the SC-WEAT *D* score—analogous to a single-category IAT *D* score in that it is a difference normalized by the standard deviation.

Significance of WEAT and SC-WEAT. To perform significance tests for the WEAT and SC-WEAT effect sizes, we repeated the above computations 1,000 times after permuting the category word vectors across category boundaries (i.e., randomly shuffling the word vectors representing the categories *men* and *women*). This yielded an empirical null distribution of effect sizes across random permutations of categories. The two-tailed *p* value was then calculated as the proportion of WEAT (or SC-WEAT) effects in the empirical null distribution that are larger in absolute magnitude than the observed WEAT effect (or SC-WEAT effect).

Notably, the results from the permutation tests often produce relatively large standard errors and therefore

sometimes reveal individual effect sizes that are nonsignificant at an alpha of .05 (see Table 2). We nevertheless argue that when these nonsignificant effects are large in magnitude (e.g., a WEAT *D* score of 0.66 with a *p* value of .10), they can be interpreted descriptively, especially alongside the meta-analyses and meta-regressions (described below). Throughout the Results section, we focus attention on the meta-analyses and meta-regressions because these approaches combine data across corpora and thereby provide greater precision and statistical power in estimating the effect sizes of interest.

Why might individual effects sometimes be nonsignificant? In traditional psychological data, standard errors are larger when there are few (vs. many) observations, all else being equal. Similarly, in word embeddings, larger standard errors (and nonsignificant effects) can arise from multiple factors related to the frequency of observations, including (a) the number of stimuli words used to represent a given group category or attribute (e.g., whether 10 or 40 words are used to represent the female category), (b) the number of occurrences of a given stimuli word in each corpus (e.g., the number of times *mommy* appears in child speech), and (c) the number of co-occurrences between stimuli words (e.g., the number of times *mommy* and *kitchen* co-occur in child speech). Ongoing research is being conducted to investigate the multiple factors that contribute to the significance and sensitivity of WEAT results (e.g., Ethayarajh, Duvenaud, & Hirst, 2020).

Step 7: meta-analyses and meta-regressions. For succinct descriptions of effect sizes across corpora and stereotype domains, a fixed-effects meta-analysis was performed using the *meta* package in the R programming environment (R Version 3.6.1; Schwarzer, 2020). A fixed-effects approach was chosen over a random-effects approach because (a) the stereotypes (in Study 1) were not assumed to be a random draw from the true population of stereotypes but were specifically selected as well-studied topics; (b) in some cases, the number of studies used for the meta-analytic estimate was small (e.g., child-directed meta-analysis was based on $k = 3$), and therefore estimation of random effects may be biased; and (c) supplementary tests of meta-analytic heterogeneity indicated little to no significant cross-study heterogeneity (see Table S2.1 in the Supplemental Material). Nevertheless, to illustrate the robustness of the results, we also report the results from random-effects meta-analyses in the Supplemental Material (see Table S2.2). All major conclusions about stereotype consistency hold regardless of the fixed-effects or random-effects approach.

In Study 1, fixed-effects meta-analytic estimates were computed to provide (a) an overall effect size (collapsing across all seven corpora and all four stereotypes, $k = 28$; see Table 2), (b) a summary effect for each of the four stereotype domains (collapsing across corpora; see Table 2), and (c) a summary effect within each of the seven corpora (collapsing across stereotype domains; see Table 2).

Additionally, we performed meta-regressions to directly compare the strength of gender stereotypes across corpora and stereotype domains. Specifically, we predicted the magnitudes of the individual effect sizes from (a) stereotype or attitude domain (Study 1 only; the four dummy-coded domains were good–bad vs. home–work vs. arts–science vs. math–reading), (b) age group (all studies; the two dummy-coded age groups were child-produced/child-directed vs. adult-produced/adult-directed), and (c) corpora time period (all studies; the three dummy-coded time periods were "early" [books] vs. "middle" [speech] vs. "late" [AV media]). In Study 1, the power to detect significant effects in the meta-regressions was limited because the total number of effect sizes was relatively small ($k = 28$). Thus, the results are offered for illustration and interpreted alongside the descriptive patterns. In contrast, Studies 2 and 3 had greater power to detect significant meta-regression effects, with more than 1,000 effect sizes in Study 2 and more than 300 effect sizes in Study 3.

Step 8: validation and replication tests. Word embeddings have begun to be implemented more widely in psychology and adjacent fields and have been shown to be valid and reliable methods for capturing psychological and social phenomena (Caliskan et al., 2017; DeFranza et al., 2020; Garg et al., 2018; Kurdi et al., 2019; Lewis & Lupyan, 2020). However, because this article introduces multiple novel corpora as well as a relatively less-used algorithm (i.e., fastText), we provide four further tests for validation and replication.

First, to assess whether the chosen word-embedding vectors cohesively represented the categories/attributes of interest (i.e., were valid indicators of the category/attribute), we compared the similarity between individual word vectors *within* a given category (e.g., within the categories *men, women, good, bad*, etc.) with the similarity between individual word vectors *across* categories. Specifically, we compared the average within-category cosine-similarity score between words (e.g., the average cosine similarity between *she, her, mommy, mom*) with the null distribution of all cross-category pairwise cosine similarities, computed through permutation tests (e.g., the pairwise cosine similarities between *she, work, he, home*, etc.). The *p* value was computed as the proportion of cosine similarities from the null distribution that were greater than the average within-category cosine similarity. If the word vectors are indeed cohesive within their category, the *p* value should be less than .05, indicating that less than 5% of

Table 2. Gender Associations Occurring in Child and Adult Natural-Language Corpora (Study 1)

	Overall results			Female–good, male–bad			Female–home, male–work			Female–arts, male–science			Female–reading, male–math		
Corpus	*D*	*SE*	*p*	*D*	*SE*	*p*	*D*	*SE*	*p*	*D*	*SE*	*p*	*D*	*SE*	*p*
Meta-analytic estimate	0.57	0.08	< .001	0.53	0.16	< .001	0.76	0.16	< .001	0.44	0.16	.005	0.55	0.15	< .001
Child-directed combined (speech, audiovisual, books)	0.46	0.12	< .001	0.42	0.24	.08	0.61	0.24	.01	0.26	0.24	.27	0.56	0.24	.02
Adult-produced/directed combined (speech, audiovisual, books)	0.66	0.12	< .001	0.49	0.25	.05	0.94	0.24	< .001	0.54	0.24	.02	0.67	0.24	.005
Child-produced speech (children to parents)	0.61	0.20	.002	0.96	0.41	.02	0.69	0.39	.08	0.66	0.40	.10	0.19	0.38	.58
Child-directed speech (parents to children)	0.44	0.20	.03	0.79	0.40	.05	−0.13	0.40	.80	0.68	0.39	.09	0.43	0.41	.30
Child-directed books	0.76	0.23	< .001	0.77	0.46	.12	1.05	0.45	.02	0.69	0.46	.14	0.55	0.46	.22
Child-directed audiovisual media	0.27	0.20	.17	−0.19	0.40	.66	0.97	0.38	.008	−0.44	0.38	.27	0.69	0.38	.07
Adult-produced speech	1.02	0.21	< .001	0.46	0.45	.31	1.07	0.43	.01	0.91	0.42	.03	1.62	0.43	< .001
Adult-directed books	0.67	0.21	.001	0.63	0.41	.13	1.09	0.42	.01	0.90	0.44	.03	0.17	0.39	.64
Adult-directed audiovisual media	0.32	0.21	.12	0.36	0.42	.43	0.69	0.40	.10	−0.11	0.41	.77	0.36	0.43	.38

Note: *D* is the Word-Embedding Association Test (WEAT) *D*-score effect size (analogous to the Implicit Association Test *D* score), providing a standardized effect size of the overlap between categories (female–male) and attributes (e.g., good–bad). The standard error of the WEAT *D*-score effect size was computed as the standard deviation of the permutation distribution of WEAT effects. The first row reports the meta-analytic estimates collapsed across all corpora, yielding summaries at the level of stereotypes. The "overall results" column reports the meta-analytic estimates collapsed across all associations, yielding summaries at the level of corpora. The second and third rows report the meta-analytic estimates for the combination of child sources and adult sources, respectively.

the null cosine similarities are greater than the actual within-category cosine similarity. Testing for cohesiveness also increases confidence that no single word stimulus is overwhelmingly driving the observed associations because all word stimuli from a category/attribute are taken to be similarly representative of the category/attribute.

Second, to assess whether the trained word embeddings accurately captured semantic associations that are known to be strong and consistent in psychological data, we tested the strength of a nonsocial association, *musical-instrument–good/weapon–bad*. If the WEAT effect for the musical-instrument–good/weapon–bad association is significant and strong within a corpus, it can be inferred that the word vectors have accurately learned the expected semantic associations (for a similar approach, see Caliskan et al., 2017).

Third, to assess replicability of the observed results, we performed the same analyses with external data sets from data from (a) stereotypes aggregated at the societal level (i.e., stereotypes measured through the IAT taken at the Project Implicit demonstration website) as well as (b) the largest-known corpus of natural language (i.e., the Common Crawl corpus, consisting of more than 600 billion words from all Internet text). If the results are replicated in the Project Implicit data set, it suggests that stereotypes measured through word embeddings are consistent with a very different form of measuring aggregate societal-level stereotypes. Additionally, if the results are replicated in the Common Crawl corpus, it suggests that the observed findings are unlikely to be an artifact of idiosyncratic features in the relatively smaller corpora but, rather, are consistent even across the majority of Internet text.

Fourth, to assess the robustness of the results to the choice of word-embedding algorithm, we retrained all word-embedding vectors using the GloVe algorithm (Pennington et al., 2014). In brief, the GloVe algorithm differs from fastText most notably in (a) representing words only as whole words (e.g., the word *cat* is represented as the whole word *cat*) rather than also representing words with subword information (e.g., also representing *cat* as a sum of "ca" and "at," as in fastText) and (b) working directly on the word–word co-occurrence matrix (for further details, see the glossary in the Supplemental Material). If the results are replicated despite these differences, it indicates that the findings are robust to word-embedding training.

Results

Meta-analyses across corpora and stereotypes. Collapsing across all seven corpora and all four stereotype associations (k = 28), we found that the meta-analytic estimate revealed a significant and large overall WEAT D score in the stereotypically expected direction (overall WEAT D = 0.57, p < .001; see Table 2 and Fig. 1). Additionally, collapsing across all corpora revealed that WEAT D scores were significant and large for each of the four stereotype domains (see Table 2; all Ds > 0.44, all ps < .005). Finally, collapsing across stereotypes, we found that WEAT effects were significant and large for five out of the seven corpora (see Table 2; all Ds > 0.44, all ps < .03). The consistency across corpora is remarkable in that even child-produced speech (from children with mean age of ~3 years) and child-directed speech (from parents) were communicating gender stereotypes that have not yet been robustly documented at such young ages (Martin & Ruble, 2010).

The two nonsignificant corpora were the child-directed audiovisual media (D = 0.27, p = .17) and the adult-directed audiovisual media (D = 0.32, p = .12), although the effect sizes of these corpora are moderate in the expected direction (see Table 2). These results are discussed below in terms of the possible role of the relatively more contemporary time period of these two corpora. Nevertheless, with these two exceptions, the meta-analytic estimates suggest surprising strength and consistency in the magnitude of gender stereotypes across stereotype domains and corpora in children's and adult's natural language.

SC-WEAT scores across corpora and stereotypes. Decomposing the relative WEAT D-score effect sizes into the SC-WEAT D scores revealed that no single-category association appeared to be driving the relative effects (see Table S3 and Fig. S1 in the Supplemental Material). In other words, the SC-WEAT scores were approximately parallel in magnitude: The stereotypically male-typed attribute (e.g., *bad, work, science, math*) was always men-associated, and the stereotypically female-typed attribute (e.g., *good, home, arts, reading*) was always women-associated. Given the parallel results across the SC-WEAT scores, we have greater confidence in reporting and interpreting the more succinct relative results for all other analyses.

Meta-regressions across stereotype domain. To directly examine the consistency across stereotype topics, we performed a meta-regression predicting the individual WEAT effect sizes (k = 28) from stereotype domain (good–bad, home–work, arts–science, math–reading). No significant differences were found across stereotype domains (all bs = −0.09 to 0.23, zs = −0.35 to 0.91, ps > .36; see Table S7.1 in the Supplemental Material), reinforcing that these domains are similarly and consistently expressed throughout child and adult language.

Nevertheless, descriptive trends suggest the strongest meta-analytic effect for home–work stereotypes, followed, in order, by associations with math–reading,

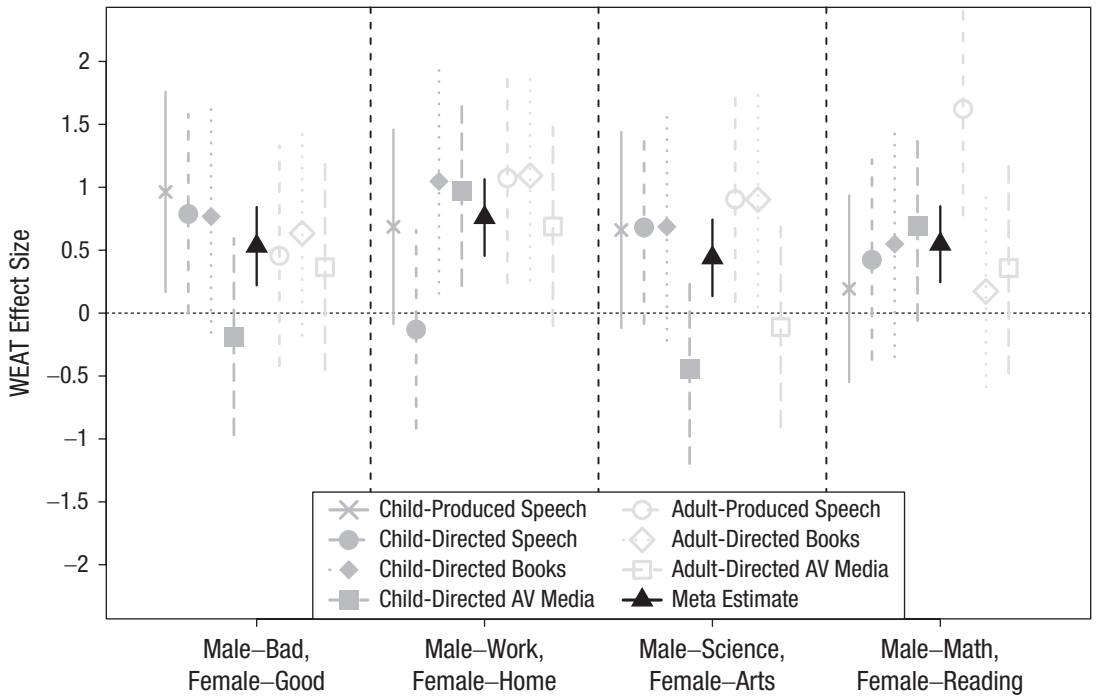

Fig. 1. Gender associations in child and adult language (Study 1). Word-Embedding Association Test (WEAT) *D*-score effect sizes are shown as a function of gender association (stereotypes and attitudes), separately for each type of child-directed/child-produced and adult-directed/adult-produced speech, books, and audiovisual (AV) media. Also shown is the meta-analytic estimate, which was computed from a fixed-effects meta-analysis across all sources. Error bars represent 95% confidence intervals computed from the standard error (i.e., the standard deviation of the permutation distribution of WEAT effect scores).

good–bad, and arts–science (see Table 2). That the home–work stereotype stands out as the descriptively strongest stereotype warrants further examination. It is possible that the domain of home–work may have greater observability than other stereotypes because gender distribution in caregiving versus labor roles can be widely observed and directly experienced by both children and adults. In contrast, distributions of specific occupational subfields (arts–science), capacities (math–reading), and especially more general evaluative associates (good–bad) may be less observable. Thus, it is possible that the direct experience and observability of home–work roles may lead to particularly strong home–work gender stereotypes being reflected in language (Eagly & Wood, 2012; Koenig & Eagly, 2014).

It is also possible, however, that language is not merely reflecting observable role distributions but also being used to draw attention to stereotypes that are deemed most important for maintaining social order, such as the separation of men and women into agentic ("breadwinning") versus communal ("caregiving") roles. From this perspective, language may be used as a pedagogical tool to provide indirect experience and perpetuate stereotypes about who should (or should not) occupy certain roles. Continued research is needed to examine the role of language as reflecting (from direct experience) or creating and perpetuating (from indirect experience) the collective representations of gender stereotypes in society.

Meta-regressions across corpora by age groups. There was no significant difference in the magnitude of WEAT *D*-score effect sizes between the baseline of child-directed/child-produced corpora and adult-directed/adult-produced corpora (b = 0.16, 95% confidence interval [CI] = [−0.18, 0.51], z = 0.91, p = .36). Despite previous work suggesting variation across children and adults in their magnitude of gender stereotypes and attitudes (Dunham et al., 2016; Martin & Ruble, 2010; Miller et al., 2018), the meta-regression suggests that, at least for these four well-studied domains, the language produced by and directed toward children and adults may be largely consistent.

Meta-regressions across corpora by time period. A small (and barely significant) difference emerged across corpora divided by time period, operationalized as early (i.e., child-directed books, adult-directed books) versus middle (i.e., child-directed speech, adult-directed speech, child-produced speech) versus late (i.e., child-directed

audiovisual media, adult-directed audiovisual media) corpora. Specifically, compared with the baseline of early corpora, later corpora were marginally significantly weaker in their expression of gender stereotypes (b = −0.42, 95% CI = [−0.85, 0.01], z = −1.91, p = .06). Early and middle time-period corpora did not differ from one another (b = −0.04, 95% CI = [−0.44, 0.36], z = −0.18, p = .86). The small difference between early and late corpora suggests that the magnitude of effect sizes, at least for these four well-studied stereotypes in language, may be slightly decreasing across time. Such decreases are in line with trends observed on other aggregated psychological measures (e.g., on male–science/female–arts associations; Charlesworth & Banaji, 2019; Miller et al., 2018) but stand in contrast to trends of increasing (or stable) stereotypes for other domains (e.g., female–communion associations have increased over time; Eagly et al., 2020).

It is important to highlight that the corpora compared across time also differ in other respects (e.g., early corpora are books whereas late corpora are transcripts of audiovisual media; middle corpora are conversations, which are more spontaneous than books or media). Ideally, temporal comparisons would be performed within a single language format (e.g., within books or within TV and movies). The present corpora are not sufficiently large for within-corpus comparisons and are thus offered as a first step in understanding change. Future research examining within-corpus change (Garg et al., 2018) will be beneficial to draw firm conclusions about patterns of gender stereotypes over time in both child and adult language.

Additional results. The main takeaway from the meta-analytic estimates and meta-regressions is that gender stereotypes are consistent in magnitude, even across stereotype domains and even across language from both children and adults. Nevertheless, we also identify a selection of surprising and potentially informative differences to guide future research.

Within child-produced speech, we note that, unlike the large effects observed for associations of home–work, arts–science, and good–bad, the effect for the math–reading stereotype was small in magnitude and nonsignificant (see Table 2). Although a female–reading/male–math stereotype has been documented among children (ages 6–10 years) with laboratory-based implicit and explicit measures (Cvencek et al., 2011), it is possible that the stereotype may emerge only after 6 years old and not at the young average age of the speakers in this corpus (3 years old). Additionally, it is possible that the female–reading/male–math stereotype may be observed on psychological tests that reinforce categorical distinctions (i.e., focusing children on categorizing by groups and attributes) but may not yet be reliable in spontaneous language that does not focus children's attention on these group boundaries.

Also, within child-directed speech (from parents to children), the home–work stereotype was small in magnitude and nonsignificant (see Table 2), unlike the relatively large effects for the three other associations. Given that this corpus is largely composed of mothers speaking with their children about daily life in the home, it may be a unique context in which both mothers and fathers are equally likely to co-occur with *work* and *home*. Phrases such as "mommy needs to work" and "daddy comes home soon" could be more common in this corpus than others and thereby lead to more neutral associations. Future research using parent-to-child speech from contexts outside the home, such as in educational or work settings, could reveal how the environment in which speech is produced also shapes the content of the speech.

Finally, within child-directed audiovisual media, two domains— arts–science and good–bad (see Table 2)— revealed nonsignificant WEAT effects. In general, this corpus may sometimes show weaker gender stereotypes because it is more contemporary than other child corpora (see the meta-regression results above). Additionally, the audiovisual media corpus may be more likely to be the focus of gender-equitable interventions, such as the United Kingdom's 2019 ban on gender stereotypes in advertisements. In line with this explanation, the good–bad association may be weak because that association could be seen as particularly harmful and important to address (i.e., associating men or women with "bad" is perceived as particularly harmful). Future research looking at changes within children's audiovisual media will be helpful in testing these explanations.

Validation and replication tests. The corpora and word vectors passed all four tests of validation and replication, indicating that the results are reliable and interpretable. First, word vectors were significantly more cohesive within category than across categories (all ps < .05; with the exception of four out of 70 effects, or 6%, which were p = .06; see Tables S1.1 and S1.2 in the Supplemental Material), indicating that the word vectors are cohesive representations of their underlying latent group/attribute categories. Second, as expected, the WEAT D score for the musical-instrument–good/weapon–bad association was strong and consistent in all seven corpora (D range = 1.32–1.50, all ps < .003; see Table S5 in the Supplemental Material), indicating that the novel data sets are sufficiently large to identify linguistic associations at the expected magnitudes. Third, the magnitude (and significance) of WEAT effect sizes was generally replicated in the available data from IATs at the Project Implicit

website, indicating consistency with a very different method of capturing societal-level, aggregated representations of gender stereotypes (see Table S7 in the Supplemental Material). The magnitude of WEAT effect sizes was also generally replicated in vectors trained on the Common Crawl corpus (see Table S7), indicating that the results appear to be consistent even with a corpus that captures nearly all Internet text. Fourth, the magnitude (and significance) of WEAT effect sizes was generally replicated with word embeddings trained using the GloVe algorithm (see Table S6.1 in the Supplemental Material). The current findings—showing consistent gender stereotypes in child and adult language across sources and stereotype topics—are therefore not dependent on any one method of representing word meaning.

Study 2: Gender–Trait Stereotypes

Experiments documenting stereotypes typically test only a subsample of topics because of concerns of interpretability, theoretical precedent, practice effects, and resource limitations. In this vein, we used a subsample of gender stereotypes in Study 1 to align with theoretical precedent. However, using a subsample of stereotypes risks misestimation if the sample does not represent the full population. Thus, having shown that word embeddings capture well-studied stereotypes, we used word embeddings to test entire populations of stereotypes with more than 600 traits (Study 2) and more than 300 occupations (Study 3). In Study 2, we focused on traits because they are a fundamental input to person perception (Fiske, Cuddy, Glick, & Xu, 2002) and are even made spontaneously without instruction. Examining hundreds of gender–trait stereotypes can reveal the consistency of such fundamental stereotypes throughout language.

Method

Data sources and procedures for cleaning, lemmatizing, and creating word embeddings were identical to those in Study 1. Only the SC-WEAT (defined above) was used in Study 2 to measure the association between a single attribute (i.e., a trait) and the group categories (i.e., male–female).

Single trait words were taken from a list of 627 traits (Peabody, 1987), providing the most comprehensive sample space of traits that were not a priori assumed to be associated with men or women. Because of the large number of effect sizes coming from the more than 600 traits across seven corpora (yielding more than 4,200 possible individual effect sizes that would be impossible to succinctly describe), effects were summarized with fixed-effects meta-analyses and meta-regressions. For reporting and describing the meta-analysis results, we retained only trait words that were present in five out of seven of the primary corpora in the final meta-analysis summary, resulting in 170 trait words.

For these 170 target trait words, we also identified five synonyms (from online thesaurus searches; see the Supplemental Material) that were specific to the *trait* meaning of the word. This ensured that words with sometimes ambiguous meanings (e.g., *frank* referring to both the trait of "being honest or direct" and the common male name "Frank") were grouped together with other trait words that clearly denoted the semantic trait meaning (e.g., *frank* was represented as the average association with "frank," "candid," "direct," "forthcoming," "honest," and "straightforward"). This increased the likelihood that the effect size was capturing the intended semantic meaning of the trait, rather than some other usage of the word.

Additional analyses with single trait words (rather than traits and their synonyms) as well as using different cutoffs for retaining trait words (e.g., appearing in one corpus, retaining 541 trait words, or appearing in all seven corpora, retaining 54 trait words) are provided in the Supplemental Material. Overarching conclusions are consistent regardless of the number of traits retained (see the Supplemental Material).

Results

Prevalence of gender–trait stereotypes across all corpora. Across the 170 trait words (aggregated with their five synonyms), 72% of traits revealed meta-analytic SC-WEAT *D*-score effects beyond [–0.1, 0.1], 47% revealed effects beyond [–0.2, 0.2], and 29% revealed effects beyond [–0.3, 0.3] (see Fig. 2). These SC-WEAT effect-size cutoffs correspond roughly to Cohen's *d* cutoffs of small, small to medium, and medium to large effects (because SC-WEAT effect sizes correspond to roughly half of a Cohen's *d*). Thus, these results can be interpreted as showing the pervasiveness of gender–trait associations: 72% of traits reveal meaningful (greater than small) effect sizes associating a trait word with male or female.

Additionally, the majority of traits (76%) were associated with women (i.e., had SC-WEAT effect sizes < 0), a proportion that is significantly more likely than would be expected if traits were equally likely to be male or female ($P = .76$, 95% CI = [.69, .83], $p < .001$). Perhaps traits are more likely to be associated with women because women is the "nondefault" social category and therefore more likely to be described and labeled (Bailey, LaFrance, & Dovidio, 2019). In contrast, men, as the default social category, is seen as synonymous with the

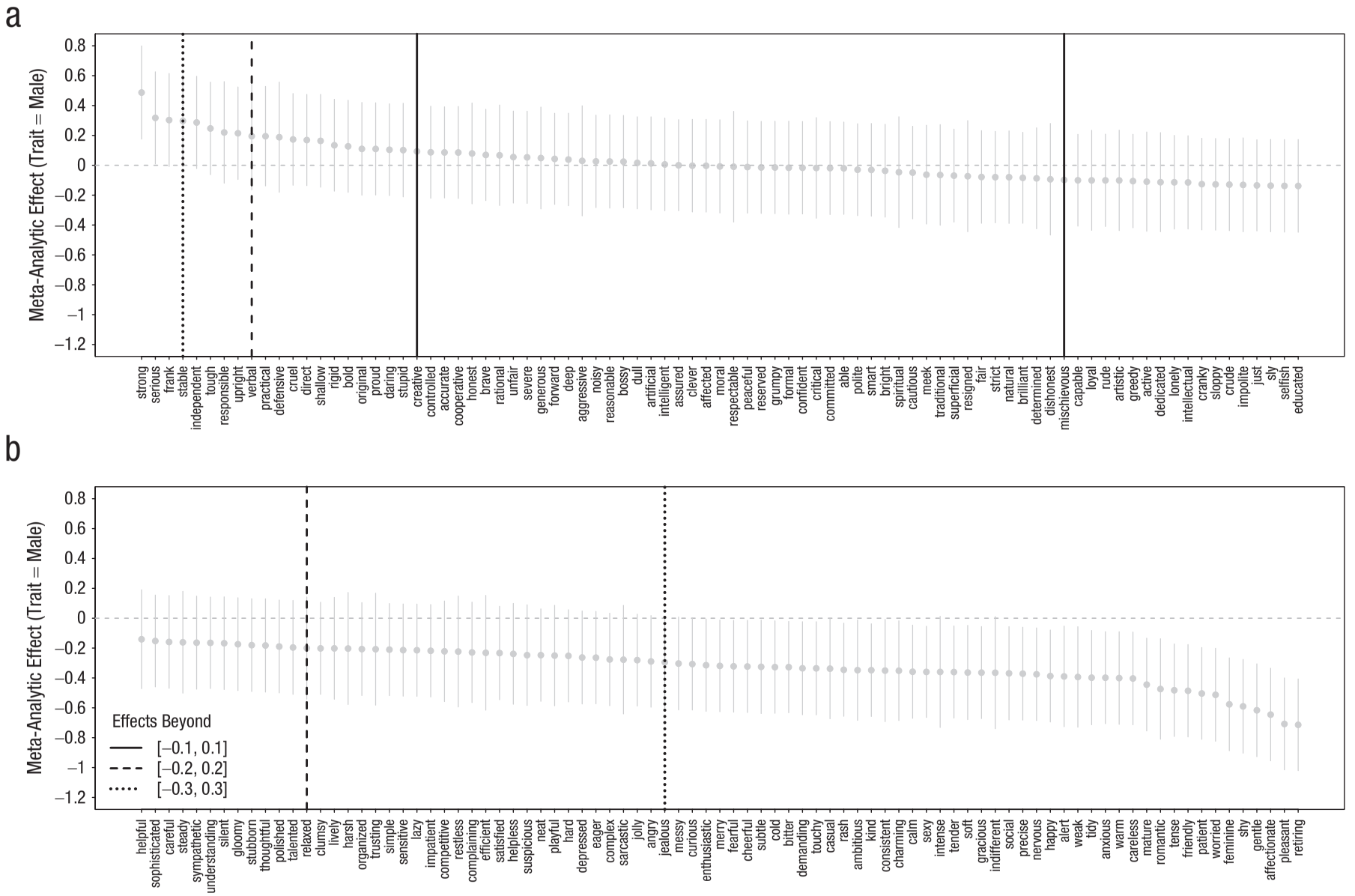

Fig. 2. Gender–trait stereotypes in child and adult language: Traits 1 to 85 (a) and 86 to 170 (b) ranked from most male to most female (Study 2). The Single-Category Word-Embedding Association Test (SC-WEAT) effect size is shown for each trait (higher scores indicate that the trait label is associated with male more than female). Traits further toward the right-hand side of each plot are the most strongly female typed; traits further toward the left-hand side of each plot are the most strongly male typed. Error bars represent 95% confidence intervals computed from the standard error (i.e., the standard deviation of the permutation distribution of SC-WEAT effect scores). Effects to the left and right of the solid black lines are greater than small effects (i.e., outside [–0.1, 0.1] SC-WEAT *D* scores), effects to the left and right of the dashed black lines are greater than medium effects (i.e., outside [–0.2, 0.2] SC-WEAT *D* scores), and effects to the left and right of the dotted black lines are greater than large effects (i.e., outside [–0.3, 0.3] SC-WEAT *D* scores).

general "human" or "person" and therefore may not be labeled with as many trait descriptors. If so, then the greater frequency of women-typed traits could be interpreted as a form of implicit androcentrism.

Despite the meaningful effect sizes for the SC-WEAT trait associations, it is notable that the standard errors of the effect sizes were large and, thus, the majority of traits were not significantly associated with male or female categories (see Fig. 2). Perhaps within the "noisy" environment of spontaneous natural language, and the polysemy inherent in many trait words, traits may not always reveal clear signals of gender stereotypes, even when the effect sizes are large. However, we reemphasize that the majority of traits revealed medium to large effect sizes, suggesting that gender–trait stereotypes are widely communicated through language.

Meta-regressions across corpora by age group. SC-WEAT effect sizes (total $k = 1{,}133$) were compared across child-directed/child-produced and adult-directed/adult-produced corpora. A significant difference emerged by age group, with adult corpora indicating significantly more female–trait associations than child corpora (indicated by more negative effect sizes; $b = -0.16$, 95% CI = [−0.21, −0.11], $z = -6.47$, $p < .001$). Notably, because the child corpora already indicated a baseline toward female–trait associations ($b = -0.08$, 95% CI = [−0.11, −0.05], $z = -4.73$, $p < .001$), this indicates that adult corpora are expressing stronger gender–trait stereotypes than children, in that they were significantly further below the neutral point. This means that children may express (and be exposed to) language that indicates more gender equity in trait stereotypes.

Meta-regressions across corpora by time period. SC-WEAT effect sizes for traits were next tested across corpora divided by time period into early (i.e., books), middle (i.e., speech), and late (i.e., audiovisual media) corpora. Significant differences were found across corpora by time period, indicating movement toward more equitable trait stereotypes over time: Compared with the baseline of early corpora showing that traits were, on average, female typed ($b = -0.23$, 95% CI = [−0.27, −0.18], $z = -9.46$, $p < .001$), both middle corpora ($b = 0.11$, 95% CI = [0.05, 0.17], $z = 3.68$, $p < .001$) and later corpora ($b = 0.10$, 95% CI = [0.03, 0.16], $z = 2.94$, $p = .003$) revealed significantly more gender-balanced trait stereotypes (i.e., more positive effect sizes). Thus, with the caveat that these temporal comparisons were confounded by other differences, the result suggests that gender–trait associations (similar to the four well-studied domains of Study 1) may be decreasing in bias, in this case decreasing in androcentric bias. However, we encourage caution in interpreting this finding because meta-regression analyses with single trait words did not reach significance, although effects were in the same direction (see the Supplemental Material).

Content of gender–trait stereotypes across corpora. In addition to the quantitative examination of gender–trait stereotypes, word embeddings can also begin to shed light on the more qualitative content of the trait stereotypes associated with men (i.e., male typed) and women (i.e., female typed). Descriptively, the top male-typed and female-typed traits across corpora can be seen to communicate the trait stereotypes that women are generally "pleasant" and "affectionate," whereas men are "strong" and "serious" (see Table 3 and Fig. 2). The general content reflected in these qualitative descriptions appears to align with adults' explicit reports that men are agentic and competent, whereas women are communal and warm (Abele, Uchronski, Suitner, & Wojciszke, 2008; Fiske et al., 2002). Future research would benefit from empirically testing the agency–communion dimension further to answer questions such as whether SC-WEAT gender–trait associations correlated with the ratings of those traits on agency–communion. Are communion stereotypes stronger than agency stereotypes? And has the strength of the communion stereotype (as represented in traits) increased across time while agency stereotypes have remained stable (Eagly et al., 2020)?

Exploratory analyses of gender–trait associations. As in Study 1, some results of gender–trait stereotypes were occasionally surprising and may call into question the validity of the analyses. We therefore report two additional exploratory analyses to show that gender–trait associations in language are indeed meaningful representations of gender stereotypes. First, we tested external validity by correlating SC-WEAT scores with actual data from child and adult participants' masculinity/femininity ratings or categorizations for a subset of traits, collected within the same decades as the speech data (obtained from Powlishta, 1995, for children and Williams & Bennett, 1975, for adults). The SC-WEAT effect sizes for traits (from traits and their synonyms) were significantly correlated with both children's ratings of a trait's masculinity/femininity, $r = .50$, $t(18) = 2.45$, $p = .02$, and the percentage of adults categorizing a trait as masculine/feminine, $r = .72$, $t(21) = 4.77$, $p < .001$ (see Tables S12.1 and S12.2 in the Supplemental Material). That is, if a trait was strongly associated with *female* in natural language, children and adults also explicitly reported that trait to be strongly feminine, and if a trait was strongly associated with *male* in natural language, children and adults also explicitly reported that trait to be strongly masculine.

Second, we tested construct validity by calculating the primary dimensions (i.e., principal components) that characterize the SC-WEAT scores and examined

Table 3. Top Female–Trait and Male–Trait Associations Occurring Across Corpora (Study 2)

Overall	Child-produced speech	Child-directed speech	Child-directed books	Child-directed audiovisual media	Adult-directed speech	Adult-directed books	Adult-directed audiovisual media
			Female-typed traits				
Retiring	Careless	Worried	Mature	Affectionate	Pleasant	Tender	Sarcastic
Pleasant	Shy	Cold	Feminine	Polite	Happy	Feminine	Friendly
Affectionate	Lazy	Helpless	Charming	Feminine	Playful	Affectionate	Pleasant
Gentle	Retiring	Suspicious	Consistent	Careless	Messy	Pleasant	Worried
Shy	Tense	Consistent	Romantic	Social	Sloppy	Gracious	Shy
Feminine	Sly	Pleasant	Tense	Crude	Casual	Gentle	Jolly
			Male-typed traits				
Strong	Rigid	Strong	Independent	Deep	Polished	Responsible	Grumpy
Serious	Controlled	Serious	Noisy	Sarcastic	Stable	Competitive	Creative
Frank	Tough	Responsible	Sly	Meek	Resigned	Accurate	Proud
Stable	Formal	Clever	Strong	Generous	Strong	Creative	Rigid
Independent	Polite	Independent	Unfair	Proud	Serious	Cranky	Noisy
Tough	Independent	Gloomy	Careful	Helpful	Unfair	Practical	Artificial

Note: Trait results were computed from aggregated trait synonyms and were ranked according to magnitude of effect sizes. Results for overall corpora were computed from meta-analytic estimates across the 170 traits that were present in at least five (out of seven) corpora. Thus, the overall results were determined on the basis of the magnitude of individual effect sizes, as well as the standard errors of the individual effect sizes and the range or variability in the magnitude of individual effect sizes. A top overall result therefore reflects both that the trait had a high magnitude of effect size, on average, and that it had low variability in the magnitude of effect sizes.

the correlations between these principal components and ratings of masculinity/femininity. The principal component analysis indicated that a one-factor solution provided the best fit to the SC-WEAT scores, with the first principal component explaining 49% of the variance (see the Supplemental Material). Moreover, the loadings on the first principal component were strongly and significantly correlated with the percentage of respondents categorizing that trait as masculine/feminine (Williams & Bennett, 1975), $r = .73$, 95% CI = [.45, .88], $t(20) = 4.82$, $p < .001$. Thus, the primary underlying component of the SC-WEAT gender–trait scores indeed appears to be the gender typing of the traits.

Study 3: Gender–Occupation Stereotypes

Societal-level stereotypes about social groups are also grounded in associations between groups and occupations: The occupations/roles that groups occupy (or are expected to) fundamentally shape the traits and qualities ascribed to those groups (Eagly & Wood, 2012). Additionally, such gender–occupation stereotypes are of interest because occupations, unlike unobservable traits, have observable real-world data on gender distributions. The strength of gender–occupation stereotypes can thus be compared with real-world gender–occupation distributions to understand the relationship between reality and stereotypes.

Method

All data and procedures for data preparation and analysis were identical to those in Study 2, except that occupation stimuli were used in place of trait stimuli.

Occupation stimuli were obtained from a list of 306 occupation titles used by the U.S. Bureau of Labor Statistics (1998). The year 1998 was chosen to match the time period of many of the corpora (e.g., child-produced speech, child-directed speech, and adult-directed speech, as well as the majority of child and adult audiovisual media) and was the earliest year available online with statistics on occupational gender distributions. Occupational–gender distribution data were obtained from the same 1998 Bureau of Labor Statistics report.

As in Study 2, because of the large number of occupations across seven corpora (yielding more than 2,000 possible individual effect sizes), the effects were summarized with a fixed-effects meta-analysis as well as meta-regressions. Occupation titles that appeared in at least one out of the seven primary corpora were retained (yielding a final sample of 82 occupations). Only single occupation titles were used (without synonyms). Additional analyses using more strict limits (appearing in five out of seven corpora, retaining 39 occupations; or in all seven corpora, retaining 17 occupations) are reported in the Supplemental Material. Overarching conclusions remain consistent regardless of the number of occupations retained (see the Supplemental Material).

Results

Prevalence of gender–occupation stereotypes across all corpora. Out of 82 occupation titles present in at least one corpus, 79% revealed SC-WEAT effects beyond [–0.1, 0.1], 57% revealed effects beyond [–0.2, 0.2], and 44% revealed effects beyond [–0.3, 0.3]. As with gender–trait associations, these results show that gender–occupation associations are strong and pervasive in child and adult natural language. Additionally, the majority (62%) of gender–occupation trait associations were associated with male (P = .62 [.51, .73], p = .04), aligning with the fact that, in the 1998 Bureau of Labor Statistics report, the workforce was 60% male.

Finally, although the majority of occupations showed large effect sizes, only a subset revealed significant effects. As with gender–trait stereotypes in Study 2, this may suggest that single labels of occupations are not always clearly gendered in the "noise" of spontaneous natural language, even when their effect sizes are large.

Meta-regressions across corpora by age group. SC-WEAT effect sizes for gender–occupation stereotypes (total k = 344) were compared between child-directed/child-produced and adult-directed/adult-produced corpora. No significant difference emerged by age group (b = −0.08, 95% CI = [−0.19, 0.04], z = −1.27, p = .20). Unlike gender–trait stereotypes—where adult corpora revealed stronger female–trait associations than child corpora—the similarity across children's and adults' gender–occupation stereotypes may emerge because these associations are more likely to be grounded in direct experiences and real-world observations (Eagly & Wood, 2012; Koenig & Eagly, 2014). That is, the gender distributions across occupations are arguably more visible than any minor gender differences that may emerge in the expression traits. Thus, to the extent that children and adults have similar direct experiences with distributions in their environments, children and adults would also be expected to show similar magnitudes of gender–occupation stereotypes.

Meta-regressions across corpora by time period. A meta-regression predicting SC-WEAT effect sizes by time period indicated movement toward weaker male–occupation associations over time: Compared with the baseline of early corpora, which indicated a significant baseline of male–occupation associations (b = 0.20, 95% CI = [0.10, 0.31], z = 3.84, p < .001), both middle corpora (b = −0.27, 95% CI = [−0.41, −0.13], z = −3.75, p < .001) and late corpora (b = −0.21, 95% CI = [−0.35, −0.07], z = −2.98, p = .003) moved toward more gender-equal occupation associations. With the caveat that comparisons by time period are likely confounded by other differences across corpora, the result suggests that, as more women have entered the workforce over the past century (Charlesworth & Banaji, 2019), children's and adults' natural-language corpora also express increasingly female–occupation associations.

Content of gender–occupation associations across sources. The occupations that revealed large effect sizes were descriptively consistent across corpora (see Fig. 3 and Table 4). For instance, *nurse* was among the top six female-typed occupations in six out of seven corpora, whereas *maid* and *teacher* were strongly female typed in five out of seven corpora; *pilot* was strongly male typed in five out of seven corpora, and both *guard* and *excavator* were strongly male typed in three out of seven corpora. This qualitative consistency aligns with the finding that children and adults did not differ in their quantitative magnitude of gender–occupation stereotypes. Moreover, the content of these gender–occupation stereotypes aligns with the occupations rated as most feminine/masculine by children and adults (e.g., Liben, Bigler, & Krogh, 2002).

Relationship between gender–occupation stereotypes and occupational gender distributions. The strength of gender–occupation stereotypes in language was significantly and positively correlated with real-world occupational gender distributions, r = .53, 95% CI = [.35, .67], $t(80)$ = 5.59, p < .001 (see Fig. 4). The more that men were represented in a given occupation in the real world, the stronger the association between *men* and the occupation in language. Although similar results have been reported for large-scale Internet text produced by and for adults (Caliskan-Islam et al., 2016; Garg et al., 2018), the current analyses extend such findings to child-produced speech, r = .46, 95% CI = [.11, .71], $t(26)$ = 2.65, p = .01, as well as across all other child and adult corpora (rs = .21–.78, all ps < .10; see the Supplemental Material). The relationship between gender–occupation stereotypes expressed in language and real-world occupational gender distributions is therefore consistent and robust, regardless of the language source, age group, or time period.

The bidirectionality of this relationship will be of interest for future research. In one direction, it is possible that gender–occupation stereotypes are collective representations that shape how men and women participate in different occupations (Gaucher et al., 2011). In the other direction, the distribution of men and women into occupations likely shapes how observers talk about, describe, and perceive those occupations and the people in those occupations (Eagly & Wood, 2012; Koenig & Eagly, 2014). For now, the current data merely reveal a coupling between language and the real world that is present even in the language of young children.

Of note, this coupling between real-world occupation distributions and stereotypes in language is moderate in

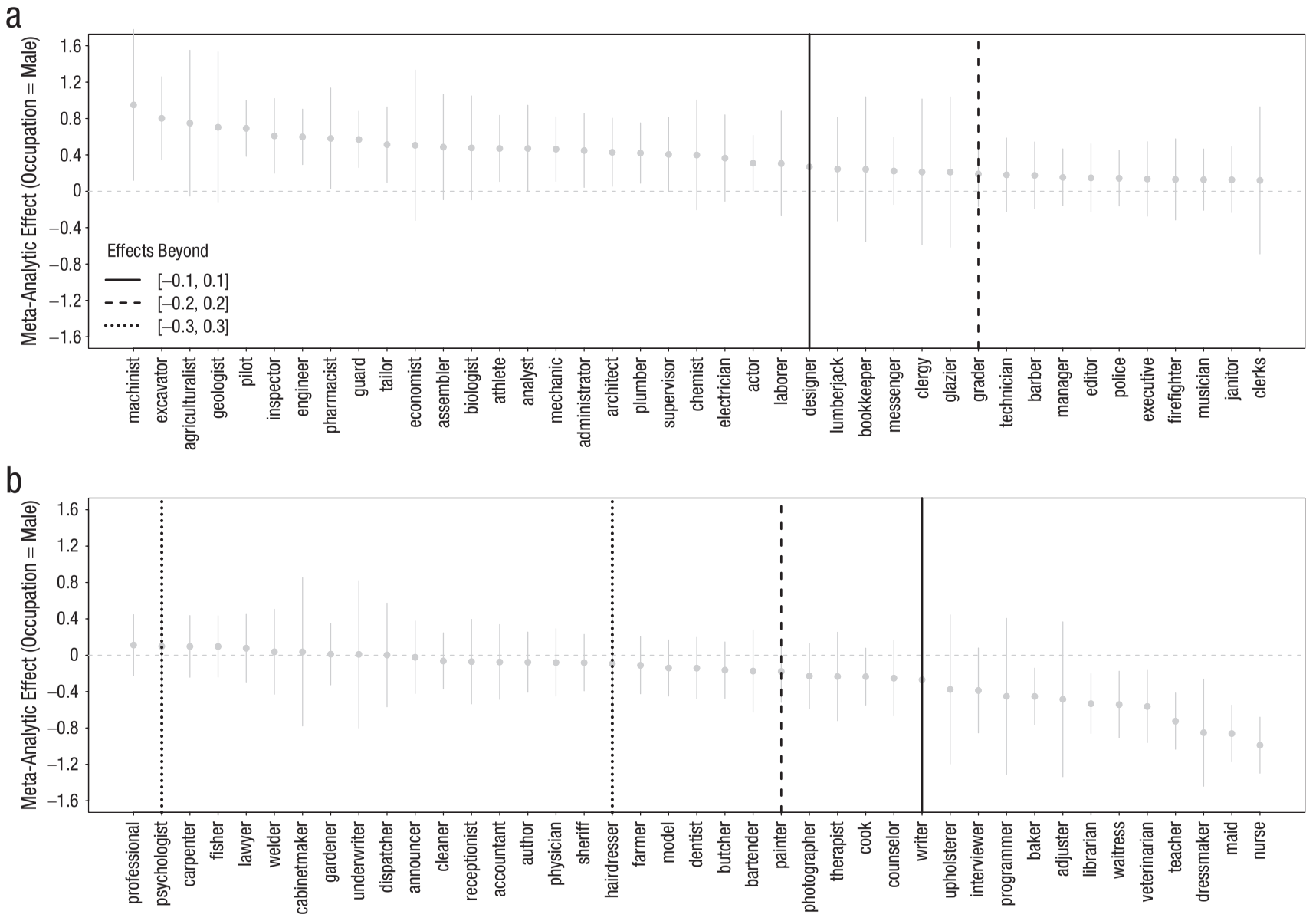

Fig. 3. Gender–occupation stereotypes in child and adult language: Occupations 1 to 41 (a) and 42 to 82 (b) ranked from most male to most female (Study 3). The Single-Category Word-Embedding Association Test (SC-WEAT) effect size is shown for each occupation (higher scores indicate that the occupation is associated with male more than female). Occupations further toward the right-hand side of the plot are the most strongly female typed; occupations further toward the left-hand side of the plot are the most strongly male typed. Error bars represent 95% confidence intervals computed from the standard error (i.e., the standard deviation of the permutation distribution of SC-WEAT effects). Effects to the left and right of the solid black lines are greater than small effects, effects to the left and right of the dashed black lines are greater than medium effects, and effects to the left and right of the dotted black lines are greater than large effects.

Table 4. Top Female–Occupation and Male–Occupation Associations Occurring Across Corpora (Study 3)

Overall	Child-produced speech	Child-directed speech	Child-directed books	Child-directed audiovisual media	Adult-directed speech	Adult-directed books	Adult-directed audiovisual media
			Female-typed occupations				
Nurse	Maid	Nurse	Dressmaker	Veterinarian	Librarian	Maid	Interviewer
Maid	Nurse	Librarian	Maid	Librarian	Nurse	Waitress	Cook
Dressmaker	Teacher	Cook	Nurse	Nurse	Waitress	Nurse	Teacher
Teacher	Sheriff	Maid	Teacher	Farmer	Teacher	Dressmaker	Model
Veterinarian	Cleaner	Veterinarian	Model	Baker	Musician	Adjuster	Maid
Waitress	Cook	Teacher	Baker	Gardener	Editor	Upholsterer	Announcer
			Male-typed occupations				
Machinist	Manager	Athlete	Police	Actor	Mechanic	Machinist	Analyst
Excavator	Pilot	Plumber	Engineer	Cook	Guard	Administrator	Guard
Agriculturalist	Excavator	Gardener	Barber	Pilot	Inspector	Excavator	Pilot
Geologist	Plumber	Excavator	Guard	Janitor	Athlete	Editor	Messenger
Pilot	Grader	Firefighter	Musician	Inspector	Chemist	Engineer	Assembler
Inspector	Guard	Announcer	Pilot	Architect	Pilot	Agriculturalist	Tailor

Note: Occupations were ranked according to magnitude of effect sizes. Results for overall corpora were computed from meta-analytic estimates across the 82 occupations that were present in at least one (out of seven) corpora. Thus, the overall results were determined on the basis of the magnitude of individual effect sizes, as well as the standard errors of the individual effect sizes and the range or variability in the magnitude of individual effect sizes. A top overall result therefore reflects both that the occupation had a high magnitude of effect size, on average, and that it had low variability in the magnitude of effect sizes.

magnitude, perhaps suggesting that stereotypes in language are also shaped by inputs other than direct experience and observation of the real world. That is, because people are not always accurate at noticing and discussing real-world statistics on gender (e.g., they underestimate the gender pay gap; Beyer, 2018), even direct experience with real-world statistics is unlikely to be a perfect predictor of how occupations are represented and stereotyped in language. It is possible that indirect experiences from language itself (which may overemphasize or underemphasize gender differences) can further create and perpetuate stereotypes in language. However, the moderate correlation may also be suppressed as a result of less theoretically interesting features of the data, including noise in the estimation of the Bureau of Labor Statistics data, noise in the SC-WEAT scores from language corpora, and the different time periods and populations of the Bureau of Labor Statistics data and language corpora. Thus, although the moderate correlation may point to the possibility of multiple sources of input to occupation stereotypes (i.e., both direct and indirect inputs), further research will be needed to support this interpretation.

General Discussion

The study of collective representations of social-group stereotypes has a long history of theory (Durkheim, 1898/2009; Moscovici, 1988), yet it has remained short on empirical evidence. In this project, we provided comprehensive and quantitative evidence that gender stereotypes are indeed collective representations, consistently expressed across different language formats, age groups, and time periods. Our ability to conduct this analysis is a function of unprecedented availability of language corpora and the emergence of machine-learning algorithms to systematically analyze such data. More than any individual finding, this project stands as a signal of the vast possibilities that lie ahead.

Gender stereotypes in language are surprisingly consistent

Across three studies, yielding thousands of effect sizes from hundreds of stereotypes and seven corpora, results revealed surprising consistency in the strength of gender stereotypes in natural language. First, four well-studied domains (e.g., female–home/male–work) all revealed large and significant meta-analytic estimates (Study 1). Indeed, consistency was observed in the magnitude of effect sizes across all four domains and across children and adults. Small differences emerged across language sources divided by time period (1800s to present day), with late corpora expressing weaker stereotypes than early corpora, perhaps suggesting movement toward more equitable gender stereotypes over time (Charlesworth & Banaji, 2019). Although the number of effect sizes in Study 1 was limited, the trends suggest that these gender stereotypes are consistently communicated collective representations.

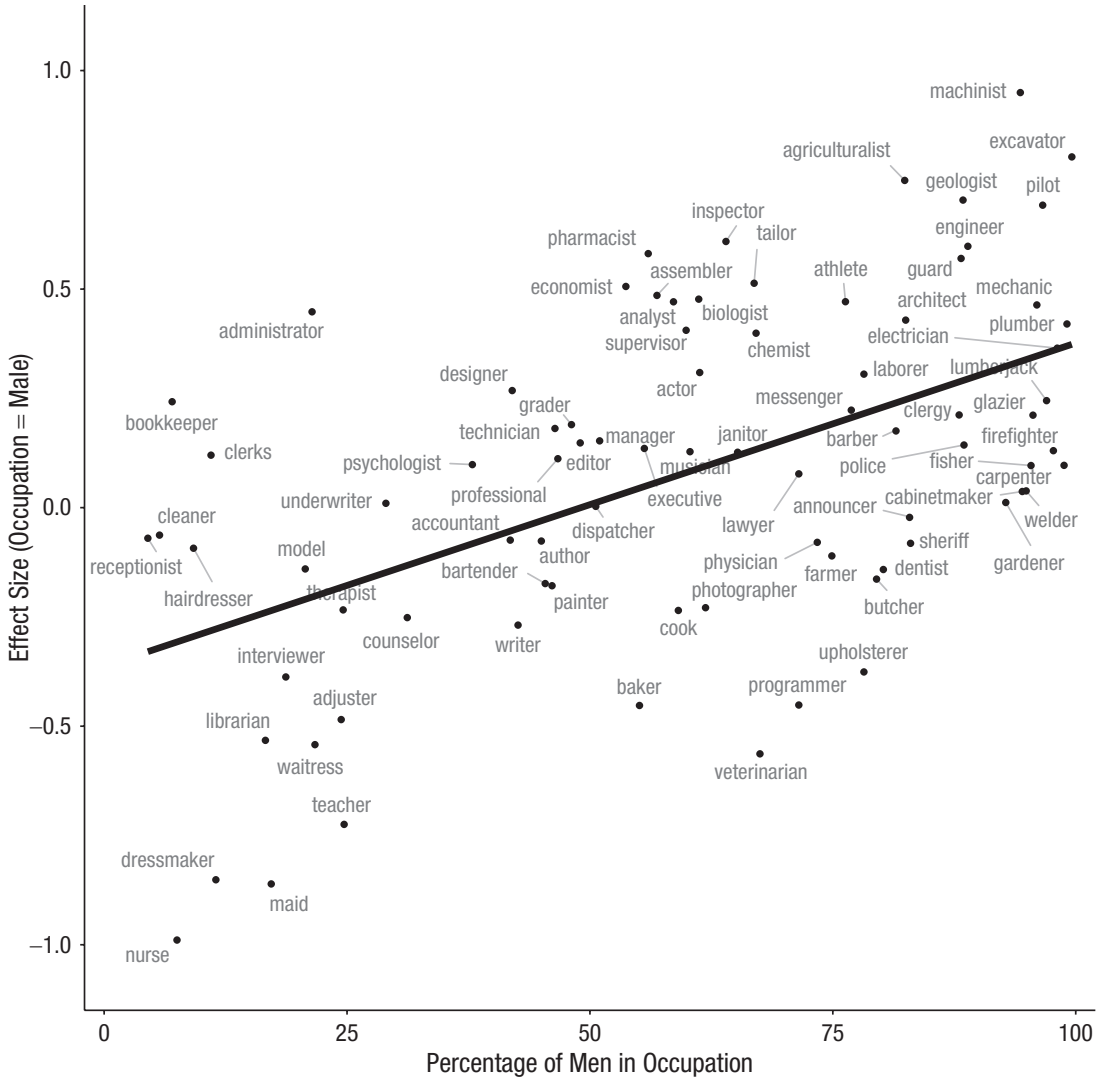

Fig. 4. Gender–occupation stereotypes in language and occupational gender distributions (Study 3). The Single-Category Word-Embedding Association Test effect size (meta-analyzed across all seven corpora) is shown for the percentage of men in each occupation. Higher scores on the *y*-axis indicate that the occupation label is associated with men more than women). Values on the *x*-axis are taken from the Bureau of Labor Statistics (1998). The line shows the best-fitting slope for a simple linear regression.

Second, gender–trait stereotypes were also found to be pervasive, even across the largest sample of traits ever simultaneously tested: 72% of traits showed meaningful associations with gender (Study 2). These pervasive gender–trait associations nevertheless indicated some change over time, with recent corpora showing more gender-equal trait associations. Additionally, gender–trait associations significantly differed across age groups: Child corpora indicated more gender-equal trait associations than adult corpora. Although children and adults may be similar in how they express widely communicated stereotypes (Study 1), they may nevertheless differ in how they express and understand more nuanced trait associations (Martin & Ruble, 2010).

Third, pervasiveness also extended to gender–occupation stereotypes, in which 79% of occupations showed meaningful associations with gender (Study 3). Gender–occupation stereotypes revealed significant differences across corpora by time period, moving toward more female–occupation associations over time, perhaps in concert with an increasingly female workforce (Charlesworth & Banaji, 2019). Indeed, the strength of gender–occupation stereotypes was significantly correlated with real-world gender distribution of occupations, suggesting a coupling between language and direct experience of the real world (Eagly & Wood, 2012). Reinforcing this interpretation is the finding that gender–occupation stereotypes were consistent across age groups, perhaps

because children and adults observe similar occupational gender distributions. Finally, across all studies, we performed multiple supplementary analyses to test robustness to methodological variations including stimuli choice, corpus selection, and word-embedding algorithms. General conclusions held throughout, indicating that the results reflect stable features of how children and adults use language to express collective gender stereotypes.

Limitations

Theoretical, empirical, and methodological advances notwithstanding, this project is limited in several ways. First, the corpora captured only a subset of children's and adults' linguistic repositories. Future research will benefit from other sources, including the language of siblings, peers, teachers, advertising, and social media. Second, the text was all English; including non-English languages, each associated with differing cultures, will advance theories of how culture and language interact in shaping collective representations (DeFranza et al., 2020). Third, although we provided preliminary analyses of patterns of change over time, we were limited in our ability to look at change within a corpus. Knowing that stereotypes are dynamic, future researchers must seek to document changes within child (and adult) language.

Conclusion

With seven corpora of more than 65 million words, this project used word embeddings to quantify the presence and magnitude of hundreds of gender stereotypes in adult and child language. Associations of gender (male–female) with well-studied attributes of home–work, arts–science, math–reading, and good–bad, as well as with hundreds of traits and occupation labels, emerged with consistent magnitude across child and adult language. These results underscore that gender stereotypes, expressed subtly through patterns of word co-occurrences in language, are deeply embedded in the social ether. We take this as the first empirical evidence for stereotypes as collective representations with a strong presence in our language and with the potential to shape how society thinks about and treats social groups.

Transparency

Action Editor: Rebecca Treiman
Editor: D. Stephen Lindsay
Author Contributions

All the authors developed the study concept and drafted the manuscript. T. E. S. Charlesworth, V. Yang, and T. C. Mann analyzed the data, and all the authors interpreted the data. All the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the Harvard University Dean's Competitive Fund for Promising Scholarship awarded to M. R. Banaji and by the Institute for Quantitative Social Sciences Undergraduate Research Scholars program.

Open Practices

All data and analysis scripts have been made publicly available via OSF and can be accessed at https://osf.io/kqux5. The design and analysis plans for this study were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges. The studies were not formally preregistered.

ORCID iD

Tessa E. S. Charlesworth https://orcid.org/0000-0001-5048-3088

Acknowledgments

We thank Aylin Caliskan for guidance and comments on the manuscript.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797620963619

References

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*, 1202–1217. doi:10.1002/ejsp.575

Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2019). Is man the measure of all things? A social cognitive account of androcentrism. *Personality and Social Psychology Review*, *23*, 307–331. doi:10.1177/1088868318782848

Beyer, S. (2018). Low awareness of occupational segregation and the gender pay gap: No changes over a 16-year span. *Current Psychology*, *37*, 373–389. doi:10.1007/s12144-016-9521-4

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186. doi:10.1126/science.aal4230

Charlesworth, T. E. S., & Banaji, M. R. (2019). Gender in science, technology, engineering, and mathematics: Issues, causes, solutions. *The Journal of Neuroscience*, *39*, 7228–7243. doi:10.1523/jneurosci.0475-18.2019

Croft, A., Schmader, T., Block, K., & Baron, A. S. (2014). The second shift reflected in the second generation: Do

parents' gender roles at home predict children's aspirations? *Psychological Science*, *25*, 1418–1428. doi:10.1177/0956797614533968

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. doi:10.1111/j.1467-8624.2010.01529.x

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, *119*, 7–22. doi:10.1037/pspa0000188

Dunham, Y., Baron, A. S., & Banaji, M. R. (2016). The development of implicit gender attitudes. *Developmental Science*, *19*, 781–789. doi:10.1111/desc.12321

Durkheim, E. (2009). *Sociology and philosophy*. New York, NY: Taylor & Francis (Original work published 1898).

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, *5*, 1–35. doi:10.1080/14792779543000002

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*, 301–315. doi:10.1037/amp0000494

Eagly, A. H., & Wood, W. (2012). Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 458–476). Thousand Oaks, CA: SAGE.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*, 275–298. doi:10.1146/annurev-psych-122216-011719

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2020). Understanding undesirable word embedding associations. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1696–1705). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/p19-1166

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902. doi:10.1037//0022-3514.82.6.878

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, USA*, *115*, E3635–E3644. doi:10.1073/pnas.1720347115

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*, 109–128. doi:10.1037/a0022530

Godfrey, J., & Holliman, E. (1993). *Switchboard-1 Release 2* (Catalog No. LDC97S62). Retrieved from https://catalog.ldc.upenn.edu/LDC97S62

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*, 1006–1033. doi:10.1177/1745691619861372

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016, May). *The Goldilocks principle: Reading children's books with explicit memory representations*. Paper presented at the 4th International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico. Retrieved from https://arxiv.org/abs/1511.02301

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*, 371–392. doi:10.1037/a0037215

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences, USA*, *116*, 5862–5871. doi:10.1073/pnas.1820240116

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*, 1021–1028. doi:10.1038/s41562-020-0918-6

Liben, L. S., Bigler, R. S., & Krogh, H. R. (2002). Language at work: Children's gendered interpretations of occupational titles. *Child Development*, *73*, 810–828. doi:10.1111/1467-8624.00440

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Martin, C. L., & Ruble, D. N. (2010). Patterns of gender development. *Annual Review of Psychology*, *61*, 353–381. doi:10.1146/annurev.psych.093008.100511

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In N. Calzolari (Ed.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 52–55). Retrieved from https://www.aclweb.org/anthology/L18-1008.pdf.

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of U.S. Draw-A-Scientist studies. *Child Development*, *89*, 1943–1955. doi:10.1111/cdev.13039

Moscovici, S. (1988). Notes towards a description of social representations. *European Journal of Social Psychology*, *18*, 211–250. doi:10.1002/ejsp.2420180303

Moscovici, S. (2000). *Social representations: Explorations in social psychology*. Cambridge, England: Polity Press.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., . . . Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88. doi:10.1080/10463280701489053

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*, 233–248. doi:10.1080/1047840X.2017.1335568

Peabody, D. (1987). Selecting representative trait adjectives. *Journal of Personality and Social Psychology*, *52*, 59–71. doi:10.1037/0022-3514.52.1.59

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In B. Pang & W. Daelemans (Chairs), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). doi:10.3115/v1/D14-1162

Powlishta, K. K. (1995). Gender bias in children's perceptions of personality traits. *Sex Roles*, *32*, 17–28. doi:10.1007/BF01544755

Rhodes, M., Leslie, S.-J., Yee, K. M., & Saunders, K. (2019). Subtle linguistic cues increase girls' engagement in science. *Psychological Science*, *30*, 455–466. doi:10.1177/0956797618823670

Schwarzer, G. (2020). *Package 'meta': General package for meta-analysis*. Retrieved from https://cran.r-project.org/web/packages/meta/meta.pdf

U.S. Bureau of Labor Statistics. (1998). *Labor force statistics from the current population survey: 1995–1999 annual averages - household data - tables from employment and earnings (Table 10)*. Retrieved from https://www.bls.gov/cps/cps_aa1995_1999.htm

U.S. Bureau of Labor Statistics. (2019). *American time use survey—2019 results (Table A-1)*. Retrieved from www.bls.gov/tus/a1-2019.pdf

Williams, J. E., & Bennett, S. M. (1975). The definition of sex stereotypes via the adjective check list. *Sex Roles*, *1*, 327–337. doi:10.1007/BF00287224

The World Bank. (2020). *Labor force participation rate, female (% of female population ages 15-64) (modeled ILO estimate)*. Retrieved from https://data.worldbank.org/indicator/SL.TLF.ACTI.FE.ZS