This manuscript is currently under media embargo.

Title: Based on Billions of Words on the Internet, PEOPLE = MEN

Authors: April H. Bailey[1†*], Adina Williams[2†], Andrei Cimpian[1]

In press, *Science Advances.*

Affiliations:

[1]Department of Psychology, New York University; 6 Washington Place, New York, NY 10003

[2]Facebook Artificial Intelligence Research, Facebook; 770 Broadway, Floor 7, New York, NY 10009

[†]Equal author contribution.

*Corresponding author. Email: ab9490@nyu.edu

Abstract:

Recent advances have made it possible to precisely measure the extent to which any two words are used in similar contexts. In turn, this measure of similarity in linguistic context also captures the extent to which the concepts being denoted are similar. When extracted from massive corpora of text written by millions of individuals, this measure of linguistic similarity can provide insight into the *collective concepts* of a linguistic community, concepts that both reflect and reinforce widespread ways of thinking. Using this approach, we investigated the collective concept PERSON/PEOPLE, which forms the basis for nearly all societal decision- and policy-making. In three studies and three preregistered replications with similarity metrics extracted from a corpus of over 630 billion English words, we found that the collective concept PERSON/PEOPLE is not gender-neutral but rather prioritizes men over women—a fundamental bias in our species' collective view of itself.

One-Sentence Summary: Based on billions of words on the internet, the concept of a PERSON is not gender-neutral but instead prioritizes men.

INTRODUCTION

Recent advances in natural language processing have enabled cognitive scientists to use large corpora of naturally produced language to characterize the content of, and relations between, human concepts at a scale that is unprecedented in the history of the field. The assumption underlying this language-based approach to the study of concepts is surprisingly simple: Words that are used in similar contexts express concepts that are similar in content (*1*, *2*). The development of sophisticated tools for computing word-usage similarity from massive corpora of language (*3*-*7*) has thus opened the door for the study of what we call *collective concepts*—representations extracted from the aggregated linguistic output of millions of individuals that both reflect and reinforce widespread ways of thinking (*8*-*10*; for a recent discussion, see *11*). Here, we apply this approach to a corpus of over 630 billion words to characterize perhaps the most basic concept in human psychology, the concept of PERSON (or PEOPLE). How do collective concepts represent the human species? Are certain groups privileged over others in these representations? In three studies and three preregistered replications, we find a fundamental bias: The collective concept PERSON is more similar to MAN than it is to WOMAN. Given the fact that women and men each make up ~50% of our species (*12*), the finding that people are conflated with men at the level of collective concepts has many problematic consequences, not just cognitively but also with respect to societal decision- and policy-making.

Language and collective concepts

In this research, we used a natural language processing tool called *word embeddings*. Briefly, a word embedding is a high-dimensional vector that represents, in a compressed format, a word's patterns of co-occurrences with the other words in a given corpus. Thus, the similarity between word embeddings, computed as the cosine of the angle between them in vector space,

reveals the extent to which the corresponding words tend to be used in similar ways (i.e., in similar linguistic contexts; *6*). For instance, the embeddings for words that are used almost interchangeably ("scientist" and "researcher") are more similar than the embeddings for words that are only occasionally used in the same linguistic contexts ("scientist" and "smart"), which in turn are more similar than the embeddings for words that occur in very different contexts ("scientist" and "instead"). Precisely, "scientist" is more similar to "researcher" (0.767) than it is to "smart" (0.204) and to "instead" (0.036), where the highest possible similarity score is 1 (based on cosine similarity and fastText word embeddings, *13*). By allowing us to measure similarity in word use, word embeddings provide a linguistic tool for approximating the similarity between the concepts being denoted.

The claim that similarity in word use can be used to measure similarity in concepts is motivated by the *distributional hypothesis* of word meaning, according to which words that occur in similar linguistic contexts have similar meanings (*1*; see also *2*, *14*). Linguist J. R. Firth summarized this hypothesis as, "You shall know a word by the company it keeps" (*15*, p. 11). To make the intuition behind this hypothesis concrete, consider a hypothetical situation in which a speaker uses the unfamiliar word "balak" (*16*). While a listener might not be familiar with this word, they can start to understand its meaning by paying attention to the linguistic context in which this word is used. For example, if the speaker says, "Each morning, Joe boiled water in the balak for tea," the listener might start to guess that "balak" means something similar to "kettle" because the words alongside "balak"—"tea," "boiled," and "water"—also frequently co-occur with "kettle" in other contexts. Essentially, this is the principle that motivates the use of word embeddings. Word embeddings capture a word's patterns of co-occurrences with other words to represent word meaning (broadly construed; see *2*, *14*). In addition, because words denote

concepts, word embedding vectors can be described equally validly as proxies for word meaning and as proxies for the concepts denoted by words.

When extracted from massive corpora of billions of words written by millions of individuals, word embeddings can be used to investigate *collective* concepts—concepts that both reflect and reinforce shared ways of thinking among a linguistic community. The notion of a collective concept, as we use it here, draws heavily on sociological theories about *collective* (8) or *social representation*s (*9*). These are systems of concepts, values, and practices that characterize a community and that also go beyond (rather than being wholly reducible to) just what individuals in that community think. Our term *collective concept* thus refers to a collective or social representation that pertains to a concept (e.g., PERSON).

This simple, language-based method of investigating collective concepts has already produced some remarkable results (*17-19*). For instance, using nothing more than similarity computations over word embeddings, researchers have been able to reconstruct the taxonomic structure of collective concepts (e.g., that WRIST and ANKLE are the same kind of thing, and different kinds of things than DOG or HAWAII; *20*) and the social biases embedded in them (e.g., that SCIENCE is more similar to MEN than to WOMEN; *11*, *21-23*). Here, we apply this powerful technique to a massive linguistic corpus in order to investigate the collective concept of PERSON and its relation to its gender-specific counterparts, WOMAN and MAN.

The PEOPLE = MEN hypothesis

Theories in philosophy, sociology, and linguistics have long argued that men are treated as the "default" humans, whereas women are treated as a gendered deviation from this male default (e.g., *24-27*). Using the terminology of the present research, this argument can be translated into an empirical claim that the similarity between the collective concepts of PEOPLE

and MEN, which we will denote as Sim(PEOPLE, MEN), is greater than the similarity between the collective concepts of PEOPLE and WOMEN, which we will denote as Sim(PEOPLE, WOMEN).

Empirical investigations in psychology have tended to support this PEOPLE = MEN claim at the level of individuals' concepts. For instance, lay participants describe more men than women when asked to think of examples of a person (*29*-*30*), select men more often than women to represent humanity as a whole (*31*), and are faster to associate men than women with words for PEOPLE (*32*; for a review, see *33*). However, considering that the samples in these studies generally consisted of no more than a few hundred participants (and often fewer), the extent to which they provide insight into the *collective* concept of PERSON is unclear.

Some larger-scale investigations, involving thousands to millions of participants, are relevant to our question. For instance, "he" occurs more often than "she" in the linguistic output of millions of individuals in news coverage and in published books (*34*, *35*). This overrepresentation of "he" is consistent with the PEOPLE = MEN hypothesis. However, "he" may also appear more often than "she" because of the linguistic practice of referring to a person of unknown gender using "he" rather than "she"—that is, due to grammatical conventions rather than due to gender biases (*27*). Thus, previous large-scale investigations do not speak directly to biases in the collective concept PERSON (and indeed they did not set out to do so) because they rely on simple frequency comparisons (e.g., does "he" occur more often than "she"?), whose interpretation is ambiguous. In contrast, word embeddings capture nuances in the typical linguistic contexts of words—including co-occurrences and higher-order co-occurrences (e.g., do "he" and "person" occur alongside the same words more often than "she" and "person"?)—and are thus ideally suited to investigate whether the collective concept of a PERSON is more similar to MAN than it is to WOMAN.

The present studies provide a direct investigation of the collective concept PERSON—a concept that is not only central to the human experience but also the basis for nearly all health, safety, and workplace policy-making enacted in modern societies (*36*-*38*). Despite the importance of this concept, there has been far less research—and no large-scale research we know of—on gender bias in the concept of PEOPLE. In contrast, other forms of gender bias (e.g., that SCIENCE is more associated with MEN than with WOMEN) have been the focus of numerous large-scale studies involving thousands to millions of participants (e.g., *39*), as well as several meta-analyses (e.g., *40*). The present studies fill this gap and investigate the collective concept PEOPLE based on the aggregated linguistic output of millions of individuals. We hypothesize that the similarity between PEOPLE and MEN will be greater than the similarity between PEOPLE and WOMEN.

RESULTS

To test whether Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN) at the level of collective concepts, we used word embeddings (*13*) extracted from the May 2017 Common Crawl corpus (CC-MAIN-2017-22; *41*), which contains a large cross-section of the internet: over 630 billion words from 2.96 billion web pages and 250 uncompressed TiB of content. Although the Common Crawl is not accompanied by documentation about its contents, it likely includes informal text (e.g., blogs, discussion forums) written by many individuals, as well as more formal text written by the media, corporations, and governments, mostly in English (*42*, *43*). Using word embeddings extracted from this massive corpus, we computed the similarity in linguistic context between words—a proxy for the similarity between the concepts denoted—as the cosine of the angle between corresponding embeddings in vector space, or *cosine similarity*.

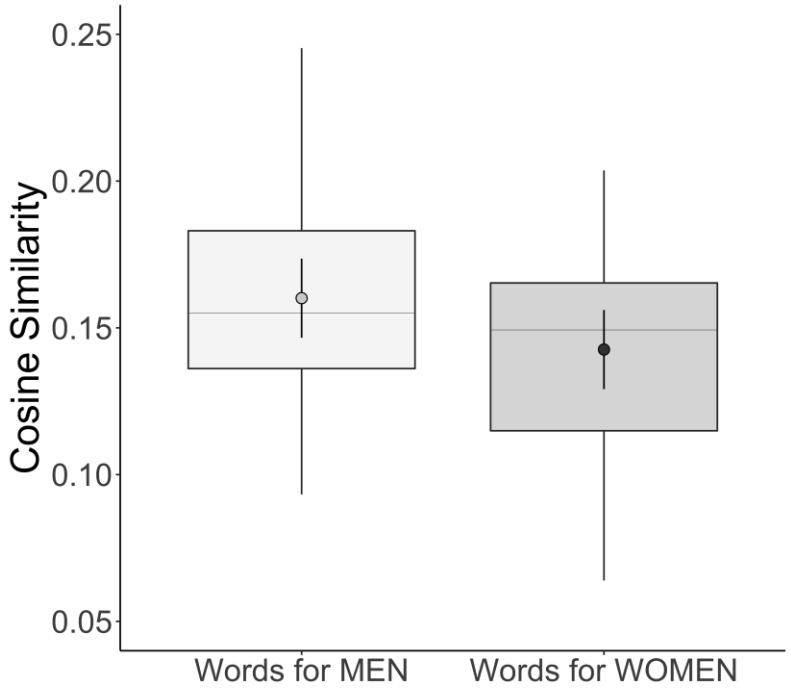
Study 1: Comparing words for PEOPLE with words for WOMEN and MEN

In Study 1, we conducted a straightforward test of the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN). We compared the similarity in linguistic context between words for PEOPLE and words for MEN to the similarity in linguistic context between words for PEOPLE and words for WOMEN. To do so, we first created suitable lists of words that denote the concepts PEOPLE (e.g., "individual," "humanity"; $n = 30$), WOMEN (e.g., "she," "female"; $n = 38$), and MEN (e.g., "he," "male"; $n = 36$; for examples, see Table 1; for full lists, see supplementary materials). Second, we retrieved the word embeddings extracted by a standard algorithm (fastText with 300 dimensions, *13*) and computed the cosine similarities between the embeddings for (a) the words for PEOPLE and the words for MEN and (b) the words for PEOPLE and the words for WOMEN.

We found that words for PEOPLE were more similar in their use to words for MEN than to words for WOMEN, $B = 0.017$, $SE = 0.004$, $p < .001$, $d = 0.465$ (Fig. 1). Differences of this magnitude ($d = 0.465$) are considered "medium" by conventional standards for effect sizes ($d = 0.50$, *44*; $d = 0.36$, *45*), and by comparison, some gender-stereotypical associations found in collective concepts are larger (e.g., SCIENCE = MEN / ARTS = WOMEN, $d = 1.24$; *21*). In summary, the collective concept PEOPLE—measured with word embeddings extracted from a large cross-section of the internet—overlaps more with the concept MEN than with the concept WOMEN.

Fig. 1

Cosine Similarity Between Words for PEOPLE, WOMEN, *and* MEN



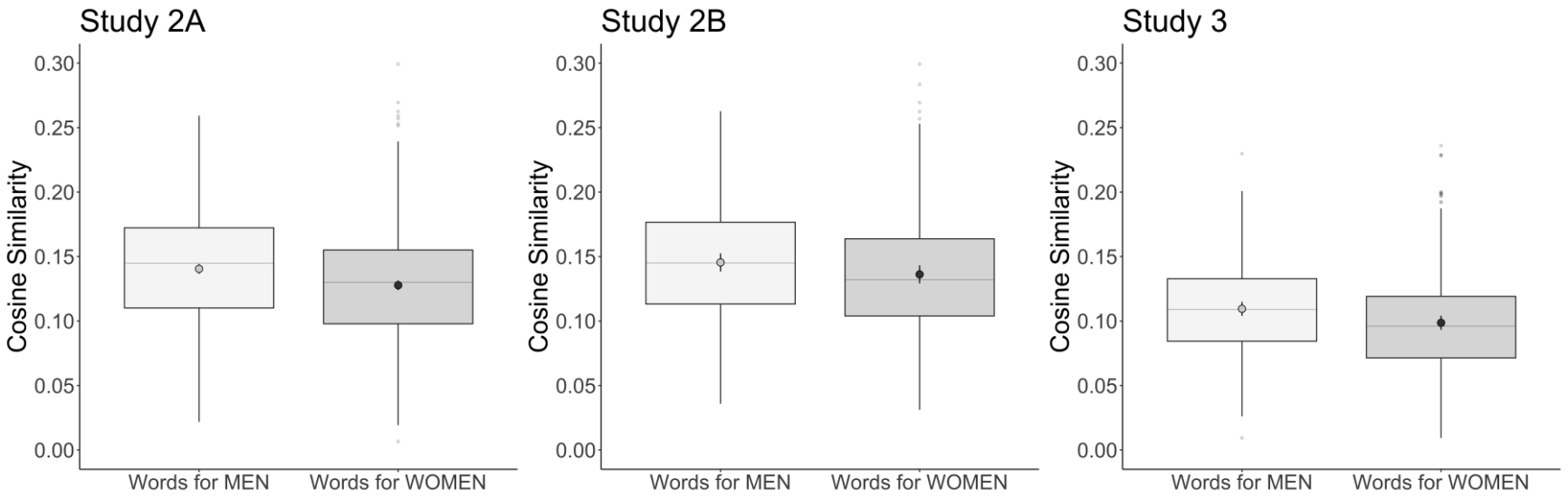
Note. Words for PEOPLE were used in more similar contexts to words for MEN than to words for WOMEN, as indicated by the cosine similarities between the corresponding word embeddings. Word embeddings for words that are always used in the same context approach a cosine similarity of 1, and word embeddings for words that are never used in the same context approach a cosine similarity of 0. Boxplots show the full range of the raw data as well as the 25th and 75th percentiles (the bottom and top edges of the boxes, respectively), and the median is a horizontal gray line. Dots are the fitted means, and error bars are 95% confidence intervals based on the fitted standard errors.

Study 2A: Comparing trait words descriptive of PEOPLE with words for WOMEN and MEN

Study 2 took a different approach to testing the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN). Instead of focusing on words for PEOPLE, we investigated words denoting *features central to this concept*—specifically, words for traits that commonly describe what people are like. In Study 2A, we compared 538 trait words identified in prior work as common descriptors of people (e.g., "extroverted"; *46*) to the same lists of words for WOMEN and words for MEN from Study 1. We found that the linguistic contexts of these common person-descriptors were overall more similar to those of words for MEN than to those of words for WOMEN, $B = 0.013$, $SE = 0.001$, $p < .001$, $d = 0.286$ (Fig. 2 left). This difference is smaller than in Study 1—likely because the trait words are more varied in meaning than the words for PEOPLE—but is nevertheless statistically reliable and provides further evidence for the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN).

Fig. 2

Cosine Similarity Between Words for WOMEN and MEN and Trait Words in Study 2A, Trait Words in Study 2B, and Verbs in Study 3



Note. Traits and verbs that describe what people are like and what they do were used in more similar linguistic contexts to words for MEN than to words for WOMEN. Word embeddings for words that are always used in the same context approach a cosine similarity of 1, and word embeddings for words that are never used in the same context approach a cosine similarity of 0. Boxplots show the full range of the raw data as well as the 25th and 75th percentiles (the bottom and top edges of the boxes, respectively), and the median is a horizontal gray line. Dots are the fitted means, and error bars are 95% confidence intervals based on the fitted standard errors.

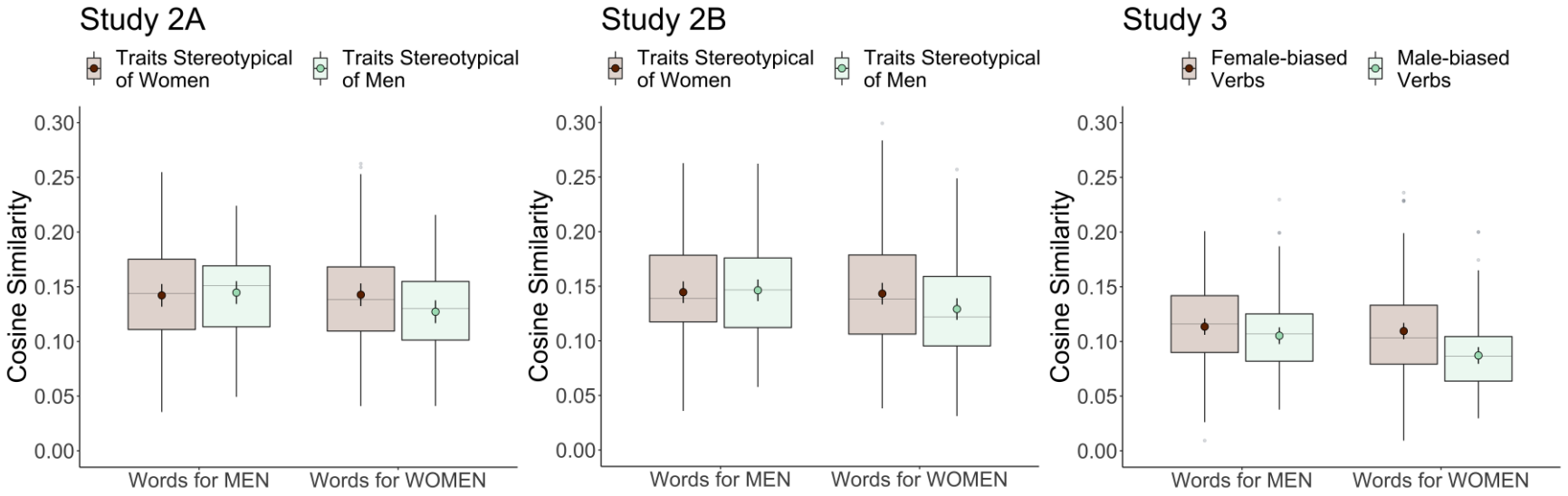
The hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN) also licenses a striking prediction about gender-stereotypical associations. In prior work in on *individuals*' psychological stereotypes about women and men, gender stereotypes are often found to be symmetrical (*39, 40*, *47-49*). For example, women are stereotyped to possess communal traits such as compassionate more than agentic traits such as brave; whereas, conversely, men are stereotyped to possess agentic traits more than communal traits (*40*). But in collective concepts, we predicted that gender-stereotypical associations would be *asymmetrical*. Our reasoning was as follows. If the collective concept of PEOPLE is conflated with MEN (as in Study 1), then words for MEN may appear in contexts that are similar to those of words for *any* trait that a person can display. Correspondingly, if the collective concept of WOMEN has less overlap with PEOPLE (as in Study 1), then words for WOMEN may appear in contexts that are similar to traits *that are specifically stereotypical of women*. That is, words denoting MEN may be similar in their usage to a wide range of common person-descriptor traits (e.g., both "brave" and "compassionate"), whereas words denoting WOMEN may be similar in their usage to a more specific set of person-descriptor traits that are stereotypical of women (e.g., "compassionate" rather than "brave").

To test our prediction in Study 2A, we first classified each trait word as stereotypical of women, men, or neither. Three raters who were unaware of our hypotheses rated the 538 traits; of these, 145 traits were rated by all three raters as more stereotypical of either women or men. Focusing on these 145 traits, we found an interaction between which gender was denoted (words for MEN vs. words for WOMEN) and which gender the traits were rated as stereotypical of (stereotypical of men vs. stereotypical of women), $B = 0.018$, $SE = 0.004$, $p < .001$. Specifically, the similarity in linguistic context between words for MEN and traits did not differ based on which gender the traits were rated as stereotypical of, $B = 0.003$, $SE = 0.007$, $p = .733$, $d = 0.056$.

In contrast, words for WOMEN appeared in more similar linguistic contexts to trait words rated as stereotypical of women than to trait words rated as stereotypical of men, $B = –0.016$, $SE = 0.007$, $p = .039$, $d = –0.344$ (Fig. 3 left). Thus, we found an asymmetry in the gender-stereotypical associations embedded in collective concepts, as we predicted based on the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN).

Fig. 3

Cosine Similarity Between Words for WOMEN and MEN and Trait Words in Study 2A, Trait Words in Study 2B, and Verbs in Study 3 As a Function of Gender-Stereotypicality



Note. The cosine similarity between words for MEN and a wide range of traits and verbs did not differ based on prior gender stereotypicality designation, but words for WOMEN were used in more similar contexts to traits and verbs stereotypical of women than to traits and verbs stereotypical of men. Word embeddings for words that are always used in the same context approach a cosine similarity of 1, and word embeddings for words that are never used in the same context approach a cosine similarity of 0. Boxplots show the full range of the raw data as well as the 25th and 75th percentiles (the bottom and top edges of the boxes, respectively), and the median as a horizontal gray line. Dots are the fitted means, and error bars are 95% confidence intervals based on the fitted standard errors.

Study 2B: Conceptual replication of Study 2A with a different set of trait words

The preceding study (Study 2A) relied on person-descriptor traits rated for gender-stereotypicality by just three raters. In Study 2B, we extracted a list of 178 person-descriptor traits directly from the gender stereotyping literature in psychology (*40*, *47-50*). All 178 traits had been designated as stereotypical of either women or men based on ratings from thousands of participants. As in Study 2A, these 178 person-descriptors were used in linguistic contexts that were overall more similar to those of words for MEN than to those of words for WOMEN, $B = 0.009$, $SE = 0.002$, $p < .001$, $d = 0.194$ (Fig. 2 center).

In addition, we again found an interaction between which gender was denoted (words for MEN vs. words for WOMEN) and which gender the traits were rated as stereotypical of (stereotypical of men vs. stereotypical of women), $B = 0.016$, $SE = 0.004$, $p < .001$. That is, the gender-stereotypical associations reflected in collective concepts were again asymmetrical: The linguistic contexts of words for MEN did not differ in their similarity to the contexts of words for traits rated as stereotypical of women vs. men, $B = 0.002$, $SE = 0.007$, $p = .807$, $d = 0.036$, but words for WOMEN were used in contexts that were more similar to words for traits rated as more stereotypical of women (vs. men), $B = –0.014$, $SE = 0.007$, $p = .049$, $d = −0.295$ (Fig. 3 center).

Study 3: Comparing verbs descriptive of PEOPLE with words for WOMEN and MEN

As a final test of the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN), Study 3 followed the same logic as Studies 2A and 2B but investigated verbs rather than trait words. If the collective concept PEOPLE overlaps more with the concept MEN than with the concept WOMEN, then words that describe what people do and what is done to them (e.g., “love,” “annoy”) may also appear in more similar linguistic contexts to words denoting MEN than to words denoting WOMEN. We compared the cosine similarities between embeddings for 252 verbs that take words

for PEOPLE as syntactic arguments (*51*) and embeddings for words for MEN vs. words for WOMEN. Overall, these "person verbs" were more similar in their usage to words for MEN than to words for WOMEN, $B = 0.011$, $SE = 0.001$, $p < .001$, $d = 0.264$ (Fig. 2 right). This result provides additional support for the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN).

The person verbs in this sample had been previously tagged as showing either a "female bias" or a "male bias" (to use the original authors' terms) with respect to their syntactic arguments, based on whether they tended to modify women (e.g., the verb "giggle") or men (e.g., the verb "kill") on Wikipedia (*51*). We used this syntactic tagging for an additional test of whether the gender-stereotypical associations reflected in collective concepts are asymmetrical, as was the case for trait words in Studies 2A and 2B. Indeed, we found an interaction between which gender was denoted (words for MEN vs. words for WOMEN) and the gender bias of the verb (male-biased vs. female-biased), $B = 0.014$, $SE = 0.002$, $p < .001$. The words for MEN did not differ in how similar their linguistic contexts were to the contexts of male- and female-biased person verbs, $B = -0.008$, $SE = 0.005$, $p = .128$, $d = -0.202$, but words for WOMEN were more similar in their linguistic contexts to female-biased verbs than to male-biased verbs, $B = -0.022$, $SE = 0.005$, $p < .001$, $d = -0.544$ (Fig. 3 right).

Replication studies, control analyses, and robustness checks

Across Studies 1–3, our findings were robust to a variety of checks (for details, see supplementary materials). First, they were not specific to a particular set of word embeddings: We replicated our results in three preregistered replication studies using an entirely different set of word embeddings (GloVe with 300 dimensions, *7*). Second, our findings were not specific to a particular corpus: We replicated our results using word embeddings trained on a corpus of biomedical research text and clinical notes (*52*) instead of general-purpose text on the internet

(i.e., the Common Crawl, which was the focus of the main studies). This biomedical corpus is of particular interest in part because biases in biomedical research have direct implications for gender (in)equity in health (*37*). Third, our findings were not explained by the fact that some of the words in our list of words for MEN are *masculine generic* words, meaning that English speakers sometimes use these words (e.g., "he") to refer to a person of unknown gender (*27*). When these words were removed from the analyses, we observed the same pattern of results. Fourth, more generally, our findings were not contingent on any particular word: We found similar results when we iteratively re-computed all of our analyses, each time removing a single word from our word lists (i.e., "leave one out" analyses).

Fifth, we built confidence in our finding of an asymmetry in gender-stereotypical associations by replicating seemingly symmetrical patterns of association from previous work on collective concepts (*11*, *21*, *53*). Previous work has used a word-embedding association test (WEAT) to study gender-stereotypical associations in word embeddings (*21*). We applied this test to our data and replicated previous evidence for gender-stereotypical associations. However, because the WEAT was designed to mimic an influential test of human biases (the Implicit Association Test, *54*), it relies on a double difference score. That is, in the present case, the cosine similarity of each trait/verb and words for WOMEN is subtracted from the cosine similarity of that trait/verb and words for MEN and then this difference score for traits/verbs designated as stereotypical of WOMEN is subtracted from the difference score for traits/verbs designated as stereotypical of MEN (for formulas, see supplementary materials). Difference scores hide any asymmetry, if present, precluding the possibility of observing the asymmetry in gender-stereotypical associations that we predicted and found.

In a sixth and final robustness check, we considered the possibility that disproportionately more text on the internet may be written about men than women, which could contribute to a the PEOPLE = MEN bias in collective concepts. The overrepresentation of men in text on the internet may itself be due to men being construed as the "default" person, but it could also be due to a variety of other factors (e.g., historic barriers to women's participation in public roles; *55*). Nevertheless, in the corpus from which the word embeddings we used were extracted, words for MEN did not occur significantly more often than words for WOMEN (for details, see supplementary materials). Thus, frequency differences cannot explain the present finding that the collective concept of PEOPLE is more similar to MEN than WOMEN. Even if words for MEN *were* in fact more frequent than words for WOMEN in our corpus, that would not necessarily explain our findings. Word embeddings tend to be more accurate for words that are more frequent (*56*), but a difference in precision between the embeddings for words for WOMEN and MEN would not, by itself, explain why the words for MEN were systematically more similar in usage to words for PEOPLE. Put differently, the extra "noise" in the embeddings for words for WOMEN would have to be *directional* to explain our results. But to reiterate, we did not find evidence that words for MEN occurred at higher frequencies than words for WOMEN in the present corpus.

DISCUSSION

We investigated the collective concept of PERSON/PEOPLE using computational tools applied to language from a large cross-section of the internet (630+ billion words) and found that this concept is not gender-neutral but instead prioritizes men over women. A key contribution of these large-scale studies is to demonstrate that the PEOPLE = MEN bias is embedded in our species' collective view of itself and is thus likely to be pervasive. Based on the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN), we also predicted and found that the gender-

stereotypical associations in collective concepts are asymmetrical. Whereas words for WOMEN were semantically closer to words for traits and actions stereotypical of women (vs. men), words for MEN did not show the corresponding difference. That is, the collective concept of WOMEN is specifically associated with the traits and actions stereotypical of women, but MEN is associated with a broader range of person-descriptive traits and actions.

The present results contribute to the extensive literature on stereotypes in psychology. Gender stereotypes are often found to be symmetrical: Men are thought to be agentic (e.g., brave) more than communal, and women are thought to be communal (e.g., compassionate) more than agentic (e.g., *40*). But we find that gender-stereotypical associations reflected in collective concepts are asymmetrical. What explains this difference?

One possibility is suggested by the fact that stereotypes and collective concepts are distinct types of representations. According to a definition common among psychologists, stereotypes are individuals' beliefs that a certain social group possesses or lacks a certain attribute (e.g., *40*). In contrast, while a collective concept reflects to some extent the beliefs of individuals in the relevant community, it is also by definition not *just* the sum of these beliefs (*8*, *9*). Collective concepts measured in word embeddings likely capture individuals' beliefs to some extent, but they also capture ideas that transcend individuals and are enmeshed in broader social systems and historical traditions. In summary, one reason why collective concepts and stereotypes show different patterns of gender-stereotypical associations (respectively, asymmetrical and symmetrical patterns) may be because they are two distinct types of representations.

In addition, the ways in which collective concepts and stereotypes are measured may help explain their different patterns of gender-stereotypical associations. Conventional ways of

measuring gender stereotypes make gender salient to participants by asking questions that directly contrast women and men: for example, “In general, do you think each of the following characteristics is more true of women or men, or equally true of both?” (*40*). In turn, the salience of gender may prompt participants to assign traits to women and men in a mutually exclusive fashion, resulting in more symmetrical patterns of gender stereotypes than might otherwise be observed. Even indirect measures of stereotypes (e.g., the Implicit Association Test; *39*, *54*) make gender salient to participants by having them sort women and men by gender group—these measures also tend to rely on double difference scores that hide any asymmetry, if present. In contrast, here, collective concepts were extracted from language produced in a broad range of real-world contexts, and in all likelihood, many of these naturalistic contexts did not make gender salient. Under these conditions, we found an asymmetrical pattern with greater gender-stereotypical associations concerning words for WOMEN than words for MEN. It will be important for future research to consider, and empirically test, whether this asymmetry in gender-stereotypical associations in collective concepts may in fact *also* characterize individual-level gender stereotypes if they are measured without making gender salient to participants.

The present work suggests several additional avenues for future research as well. Here, we showed that women are less central than men to the collective concept PEOPLE, but gender non-binary individuals may be even more marginalized in this collective concept, given that the very existence and legitimacy of these identities has been questioned (*57*, *58*; but see *59*). Further, words for WOMEN and MEN (e.g., “female” and “male”) apply to individuals with a range of other social identities besides gender, such as race, ethnicity, age, nationality, etc. (*60*, *61*). Future research should consider possible intersections between gender (including non-binary identities) and other key dimensions of identity in collective concepts. This could be done by

examining embeddings for words that simultaneously encode information about gender and, for instance, race (e.g., first names). Such research could reveal whether the PEOPLE = MEN bias is more pronounced about certain subgroups of PEOPLE than about others.

In addition to examining variation in the PEOPLE = MEN bias *about* various subgroups, it would also be worthwhile to examine variation of this bias *among* different groups and subgroups of speakers (e.g., men vs. women; English-speakers vs. Spanish-speakers; adults vs. children; people from the UK vs. people from the US). This could be done by examining word embeddings trained on a smaller corpus of language produced exclusively by members of a certain subcommunity. Such investigations of different subcommunities could also help address two open questions about the present phenomenon, which we discuss next.

First, is it possible that the PEOPLE = MEN bias is driven largely by men? Men may write disproportionately more to text on the internet compared to people with other gender identities, and men are also particularly likely to prioritize their own gender group in their individually held PERSON concept (*32*). As a result, men's linguistic output may be largely responsible for an overall PEOPLE = MEN bias in the collective concept of a PERSON. One of our robustness checks makes this possibility somewhat unlikely. Recall that we found virtually the same amount of PEOPLE = MEN bias in word embeddings trained on a corpus of biomedical text. Given the overrepresentation of men as authors in the biomedical domain (*62*), this corpus presumably includes an even greater proportion of text written by men compared to undifferentiated text on the internet (i.e., the Common Crawl corpus). The fact that this (presumably) greater imbalance in the gender of the individuals who produced the text did not result in any appreciable change in the extent of PEOPLE = MEN bias goes against the possibility that men alone are driving the patterns we observed here. Nevertheless, future research on smaller, more differentiated corpora

(i.e., produced by women vs. men) would be informative about the role of speakers' own gender identity in the PEOPLE = MEN bias.

A second open question is the following: Is it possible that the PEOPLE = MEN bias documented here is driven by particular features of the English language? Languages differ in the extent to which their grammars encode information about gender. Some languages specify gender information on nouns, pronouns, verbs, and adjectives (e.g., Spanish); other languages do not include any information about gender in that way (e.g., Turkish); English falls somewhere in between. This variation across languages is potentially relevant to the PEOPLE = MEN bias: The more a language encodes information about gender, the less likely it is to include suitable gender-neutral terms, and the more it may then license using male terms when referring to a person of unknown gender (e.g., "he" in English, "él" in Spanish; *27*). The practice of using such masculine generic terms may be part of what causes the PEOPLE = MEN bias to develop in collective concepts. It is noteworthy that the presence of masculine generic terms in our word lists did not explain the PEOPLE = MEN bias in our own data; this bias was observed even when masculine generic terms were excluded from the analysis (see the supplementary materials). Nevertheless, it is possible that the very existence of masculine generics in a language exacerbates the PEOPLE = MEN bias in collective concepts because masculine generics suggest to speakers of that language that one gender (i.e., men) can stand in for the generic PERSON category. Variation in this aspect of language could thus correspond to variation in the PEOPLE = MEN bias across different linguistic communities. Future research could systematically compare different linguistic communities while also accounting for other cultural-level variation in gender attitudes and norms to test this possibility. Such research would also contribute to a more

complete view of who is privileged in the collective concept PEOPLE among different linguistic communities around the world.

Implications

Collective concepts do not just reflect but also *instill* and *reinforce* widespread ways of thinking about women and men (*8*, *9*). Thus, the present findings have broad implications for society.

First, the conflation of PEOPLE with MEN at the level of collective concepts likely helps to instill a PEOPLE = MEN cognitive bias in each new generation of individuals. In the present investigation of collective concepts, we found the PEOPLE = MEN bias in large-scale statistical regularities in the linguistic environment. Children are sensitive to the statistical structure of their linguistic environments (*16*, *63*, *64*). It is thus likely that children are able to infer how others in their linguistic community conceive of the concept PEOPLE without receiving any explicit input on this topic. In this way, the PEOPLE = MEN bias is maintained across generations, perpetuating decision-making that advantages men with negative consequences for women's health, safety, and workplace well-being (*36*-*38*).

Second, the PEOPLE = MEN bias in word embeddings likely spills over into the wide range of downstream artificial intelligence applications that utilize word embeddings, including machine translation, automatic answering of user-generated questions, automatic recommendations on a range of topics (e.g., in the financial or legal system), and content ranking systems (e.g., Google Search and Twitter feed ranking; *65*, *66*). Previous research has documented social biases in virtually all applications that are reliant on word embeddings (e.g., *67*-*70*). Consider machine translation, for example. When "the doctor" in the English sentence "The doctor asked the nurse to help her in the procedure" is translated into Spanish, this noun is

automatically assigned masculine gender even though the pronoun "her" in the original sentence clearly indicates that the doctor was a woman ("El doctor le pidio a la enfermera que le ayudara con el procedimiento"; *71*). Such gender biases in machine translation have been documented in currently active commercial systems that rely on word embeddings (*72*). Ongoing efforts to "debias" word embeddings to prevent them from replicating such biases have yielded mixed results (*56*, *73*, *74*) and have yet to consider the fundamental PEOPLE = MEN bias we uncover here. This raises a key point. Even if every single individual's own cognitive bias to conflate PEOPLE with MEN were to suddenly disappear, there would still be PEOPLE = MEN bias in our culture because it is embedded in our artificial intelligence systems and applications that are built on the linguistic output of previous generations. We hope the present work guides future efforts to debias natural language processing algorithms.

To conclude, we investigated the collective concept of PEOPLE using word embeddings distilled from billions of words on the internet. We found that speakers write (and to some extent presumably, think) about PEOPLE and MEN more similarly relative to how they write (and think) about PEOPLE and WOMEN, indicating that the collective concept PEOPLE privileges men over women.

MATERIALS AND METHODS

In all studies, our methods proceeded in three steps. In Step 1, we created suitable lists of words for the concepts of interest. In Step 2, we extracted word embeddings for each word on these lists. In Step 3, we computed cosine similarity scores—a standard metric of similarity in word embeddings. Steps 2 and 3 are the same across studies and are thus only described in detail under Study 1. Note that throughout, we use small caps to distinguish concepts from words, following a longstanding convention in cognitive psychology (e.g., PEOPLE is the concept denoted by the word "people"). We also assume that singular and plural versions of the same word (e.g., "person" and "people") denote the same substantive concept. We thus use the singular and plural words interchangeably when referring to concepts (e.g., PERSON and PEOPLE).

Study 1

Word Lists (Step 1)

We first generated lists of words for the concepts PEOPLE, WOMEN, and MEN. For PEOPLE, a preliminary list was developed by the research team. For WOMEN and MEN, we used the gender dictionaries (i.e., word lists) supplied by the Linguistic Inquiry and Word Count software (LIWC2015; *75*) as a starting point. We removed gender words that pertained to specific domains with gender-stereotypical connotations (e.g., personal relationships, leadership), focusing as much as possible on words for MEN and words for WOMEN as generic constructs. Note that the present investigation focuses only on the gender concepts of WOMEN and MEN. Our methodology does not isolate representations of gender non-binary individuals (*76*), nor does it differentiate between biological and social aspects of sex and gender (see *gender/sex*; *77*). Our three lists of words for the concepts PEOPLE, WOMEN, and MEN were further augmented with synonyms and highly related words by inputting each word into WordNet (*78*). This process resulted in preliminary lists of 28 words for PEOPLE, 33 words for WOMEN, and 32 words for MEN.

Six coders who were unaware of our hypotheses rated these preliminary lists. Each list was presented in a separate block, with the order of the blocks randomized, although the gender blocks were always completed back-to-back. For each of the three types of words, coders were provided with a description of the underlying concept and then rated each word in terms of its fit with this concept (1 = *not a good fit* to 9 = *a good fit*). The order of the words on each list was randomized. Intra-class correlations treating both raters and words as random effects indicated moderate consistency among coders, *ICC* = .65 (*79*). Ratings were generally high—no words were rated below the scale midpoint—and thus all words were retained. Coders were also asked to generate additional words that were a good fit for the concept but were not already included in the lists they rated. We added the three words that were generated by two or more coders (i.e., "beings" and "group" for PEOPLE and "femme" for WOMEN).

Finally, we again examined the resulting lists of words. At this stage, we added seven gender words that had an obvious other-gender counterpart but that the previous steps had not produced. For instance, the gender word list included "male's" but not "female's," so we added "female's" at this stage along with: "guys," "gentleman's," "manhood," and "laddie" to words for MEN (to parallel "lady's," "womanhood," and "lassie") and "schoolgirls," "womens," and "shes" to the words for WOMEN (to parallel "schoolboys," "mens," and "hes"). This resulted in our final list of 30 words for PEOPLE, 38 words for WOMEN, and 36 words for MEN. Several examples of each type of word are provided in Table 1; the full lists are available in the supplementary materials.

Word Embeddings (Step 2)

We used fastText—an unsupervised predictive learning algorithm—word embeddings that had been trained on the May 2017 Common Crawl corpus (*13*). Although fastText word embeddings are available for other, smaller corpora, we chose the Common Crawl because the present study investigated the PEOPLE = MEN hypothesis in culture broadly rather than in a specific domain, so the largest available corpus was the best fit for our research aims. We extracted fastText embeddings with 300 dimensions for each word on our three lists.

The May 2017 Common Crawl is a large collection of over 630 billion tokens (roughly, words) and contains 2.96+ billion web pages and over 250 uncompressed TiB of content (*41*). Recent investigations of the Common Crawl suggest the majority of this corpus is written in English and based on webpages generated within a year or two of their inclusion in the corpus (*43*). The most prevalent 25 websites in the 2019 version include websites on patents filings, news coverage, and peer-reviewed scientific publications (*43*), but more informal content such as travel blogs and personal websites are also represented (*42*).

Cosine Similarity (Step 3)

To measure similarity between word embeddings, we computed the cosine similarity between each word for PEOPLE and each gender word (as in *21*). Cosine similarity is the cosine of the angle between two vectors—in this case, two word embeddings. Similarity scores range from –1 to 1, and can be thought of as being conceptually similar to a correlation coefficient. A cosine similarity score of 1 indicates that the two words are used in identical contexts; a similarity score of 0 indicates that the two words are orthogonal and used in unrelated contexts; and a score of –1 indicates that the two words are used in exactly opposite contexts.

Following the analytic strategy of references *21* and *22*, we computed two averages for each word for PEOPLE: (1) the average across the word's cosine similarity scores with all words for WOMEN and, separately, (2) the average across the word's cosine similarity scores with all words for MEN. This process resulted in two scores for any given word for PEOPLE (e.g., "person"): One score captured the average similarity between this word and words for WOMEN and the other score captured the average similarity between this word and words for MEN. These scores allowed us to test the hypothesis that Sim(PEOPLE, MEN) > Sim(PEOPLE, WOMEN).

Study 2A

The methods and materials were similar to Study 1 and again proceeded in three steps. In Step 1, we created a suitable list of person-descriptor trait words (*46*). The list of words for MEN and words for WOMEN was the

same from Study 1. In Step 2, we extracted word embeddings for each word on these lists, using fastText word embeddings with 300 dimensions trained on the Common Crawl corpus. In Step 3, we computed the average cosine similarity between each trait word and words for WOMEN and, separately, words for MEN.

To create a suitable list of common trait words that describe what people are like, we drew on the literature in personality psychology. An influential paper (*80*) developed several lists of traits that capture a range of basic aspects of people's personalities. These lists have subsequently been used widely to study personality, including a list of 587 traits that was recently used by reference *46*. Following precedent (*46*), we removed 47 amplifications (e.g., "overambitious") from this list. We also removed the trait words "masculine" and "feminine" because these words were also in our list of words for WOMEN and words for MEN. For the present study, this process resulted in a final list of 538 traits.

Next, we determined which gender (if any) each trait was stereotypical of. By necessity, we made this determination using conventional methods that make gender salient to coders (see Discussion). Six coders who were unaware of our hypotheses rated the 538 traits as stereotypical of either women or men. Coders also had the option to say that a given trait was not specifically stereotypical of either women or men or that the word was unfamiliar to them. Because of the large number of traits, each coder only coded half of the traits, meaning that each trait was coded by three of the six coders. To be conservative, we designated traits as stereotypical of women or men only if there was consensus among the three coders. This occurred for 145 traits. Several examples of each type of trait are provided in Table 1; the full lists are available in the supplementary materials.

Study 2B

The methods and materials were the same as in Study 2A, except we used a different list of person descriptive trait words. To create this list, we drew on the gender stereotyping literature in psychology. Several investigations of gender-stereotypical beliefs both about the self and about others have identified lists of common descriptors—often traits—that are considered particularly characteristic of women or men. These designations are based on large-scale polling data as well as lab-based studies with U.S. and international participants.

We examined five such lists to extract an initial list of 316 words (*40*, *47-50*). Many traits appeared on multiple lists—as would be expected given how these lists are created—so we removed repetitions. Because our focus was on traits and trait-like descriptors, we also removed occupation nouns. For the purpose of extracting word embeddings, we removed multi-word phrases or, whenever possible, split them into single word descriptors; for

instance, we changed "polite and well-mannered" into "polite" and "well-mannered" (*40*). Finally, we removed the traits "masculine" and "feminine" because these words were in our list of words for WOMEN and words for MEN. This process resulted in a final list of 178 traits. The list of words for WOMEN and words for MEN was the same from Study 1. Several examples of each type of trait are provided in Table 1; the full lists are available in the supplementary materials.

Study 3

The methods and materials were the same as in Studies 1 and 2, except we compared the cosine similarity of words for WOMEN and, separately, words for MEN with a list of person-descriptive verbs. To create a suitable list of verbs, we drew on the natural language processing literature on gender bias. Specifically, a prior investigation (*51*) automatically extracted verbs based on whether they were more likely to take words for women (e.g., the verb "giggle") or words for men (e.g., the verb "kill") as syntactic arguments on Wikipedia. This process identified 300 instances of verbs that are relatively more "female-biased" or "male-biased," to use the original authors' terminology. These verbs are suitable for our purposes because they describe things that people (women and men) do and can thus be used as proxies for the concept PEOPLE. Further, the fact that these verbs were already designated as male-biased or female-biased enabled us to test our second prediction about an asymmetry in gender-stereotypical associations reflected in collective concepts.

Note that some verbs appeared more than once on the original authors' (*51*) list because their gender-bias designation depended on two other factors: the verb's valence (i.e., sentiment) and the syntactic position of the gender-biased arguments (subjects vs. objects). Verbs were designated as positive, negative, or neutral in valence, and some verbs had, for instance, positive connotations with arguments of one gender but neutral connotations with arguments of another gender. Verbs also could exhibit bias toward one gender in the subject position but toward another gender in the object position. For instance, the verb "create" was female-biased in the object position with positive connotations but male-biased in the subject position with neutral connotations.

Of the 300 verbs on the initial list, we removed verbs that were *both* male- and female-biased, as long as they also had the same valence in both cases and the bias occurred in same syntactic position. We removed these verbs because our research question requires a list of verbs with distinct gender-stereotypical designations. For verbs that repeated in all respects except that they were found to have multiple valences (e.g., positive and neutral), we removed the non-neutral valence cases to avoid redundancies. Finally, we removed a few items from the initial list

that were not verbs or were otherwise ambiguous (e.g., “brazen” was removed because it is an adjective rather than a verb). This process resulted in a final list of 252 cases of verbs, corresponding to 211 unique verbs. As explained above, this list contained some repetitions based on differing valence or syntactic position of the gender bias (subject vs. object). The list of words for MEN and words for WOMEN was the same from Study 1. Several examples of verbs are provided in Table 1; the full lists are available in the supplementary materials.

Table 1

Summary of Word Lists Across Studies

Word Type	Study	Gender Stereotypicality	Examples	*N*
Words for PEOPLE	1		people, person, somebody, someone, human, humanity	30
Words that describe PEOPLE (traits)	2A	stereotypical of women	accommodating, cheerful, fault-finding, gullible, opinionated, sympathetic	538
		stereotypical of men	abusive, candid, forward, grumpy, outspoken, unaffectionate	
	2B	stereotypical of women	appreciative, complicated, family-oriented, gentle, out-going, suggestive	178
		stereotypical of men	arrogant, controlling, forceful, greedy, rational, witty	
Words that describe PEOPLE (verbs)	3	female-biased	adore, complain, entertain, gossip, kiss, scare	252
		male-biased	appoint, cheat, honor, kill, respect, speak	
Words for WOMEN	1–3		woman, women, female, females, she, ms	38
Words for MEN	1–3		man,[a] men, male, males, he,[a] mr	36

[a]These so-called *masculine generic* terms are sometimes used generically to refer to a person of any gender. Key for our purposes, the present findings are not merely due to these words being in our word list: Similar results are obtained when these words are removed from the analyses (see the supplementary materials).

References

1. Z. S. Harris, Distributional structure. *Word **10***, 146-162 (1954).

2. A. Lenci, Distributional Models of Word Meaning, *Annu. Rev. Linguist. **5***, 151-171 (2018).

3. Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model. *J. Mach. Learn. Res. **3***, 1137–1155 (2003).

4. R. Collobert, J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning" in *Proceedings of the 25th International Conference on Machine Learning*, 160-167 (2008).

5. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev. **104***, 211-240 (1997).

6. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" in *Advances in Neural Information Processing Systems*, 3111-3119 (2013).

7. J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543 (2014).

8. E. Durkheim, *The Elementary Forms of Religious Life* (Oxford University Press, Oxford, UK, 1915)

9. S. Moscovici, Attitudes and opinions. *Annu. Rev. Psychol. **14***, 231-260 (1963).

10. B. K. Payne, H. A. Vuletich, K. B. Lundberg, The bias of crowds: How implicit bias bridges personal and systemic prejudice, *Psychol. Inq*. ***4***, (2017).

11. T. E. Charlesworth, V. Yang, T. C. Mann, B. Kurdi, M. R. Banaji, Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol. Sci.* ***32***, 218-240 (2021).

12. H. Ritchie, M. Roser, "Gender ratio" (Our World in Data, 2019; https://ourworldindata.org/gender-ratio).

13. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, "Advances in pre-raining distributed word representations" in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (2018).

14. B. M. Lake, G. L. Murphy, Word meaning in minds and machines. *Psychol. Rev.* (2021).

15. J. R. Firth, "A synopsis of linguistic theory, 1930-1955" in *Studies in Linguistic Analysis* (Blackwell, Oxford, 1957).

16. S. McDonald, M. Ramscare, "Testing the distributional hypothesis: The influence of context on judgments of semantic similarity" in *Proceedings of the Annual Meeting of the Cognitive Science Society* (2001).

17. L. Gutiérrez, B. Keith, "A systematic literature review on word embeddings" in *International Conference on Software Process Improvement*, 132-141 (2018).

18. A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput.* ***8,*** 842-866 (2020).

19. S. Ruder, I. Vulić, A. Søgaard, A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.* ***65***, 569-631 (2019).

20. D. L. Rohde, L. M. Gonnerman, D. C. Plaut, An improved model of semantic similarity based on lexical co-occurrence. *Commun. ACM* ***8***, 627-633 (2006).

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

22. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* ***115***, E3635-E3644 (2018).

23. M. Lewis, G. Lupyan, Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat. Hum. Behav.* ***4***, 1021-1028 (2020).

24. S. L. Bem, *The Lenses of Gender: Transforming the Debate on Sexual Inequality* (Yale University Press, New Haven, CT, 1993).

25. S. de Beauvoir, *The Second Sex,* C. Borde, S. Malovany- Chevallier, Trans. (Vintage Books, New York, NY, 1949/2010).

26. C. P. Gilman, *The Man-Made World: Or, Our Androcentric Culture* (Charlotte Company, ed. 3, 1911).

27. M. Hellinger, H. Bußmann, Eds., *Gender Across Languages: The Linguistic Representation of Women and Men* (John Benjamins, Philadelphia, PA, 2003), vol. 3.

28. A. H. Eagly, M. E. Kite, Are stereotypes of nationalities applied to both women and men? *J Pers. Soc. Psychol.* ***53***, 451-462 (1987).

29. M. C. Hamilton, Masculine bias in the attribution of personhood: People = male, male = people. *Psychol. Women Q.* ***15***, 393-402 (1991).

30. R. D. Merritt, C. J. Kok, Attribution of gender to a gender-unspecified individual: An evaluation of the people = male hypothesis. *Sex Roles* ***33***, 145-157 (1995).

31. A. H. Bailey, M. Lafrance, Who counts as human? Antecedents to androcentric behavior, *Sex Roles* ***76***, 682-693 (2016).

32. A. H. Bailey, M. LaFrance, J. F. Dovidio, Implicit androcentrism: Men are human, women are gendered. *J. Exp. Soc. Psychol*. ***89***, 103980 (2020).

33. A. H. Bailey, M. LaFrance, J. F. Dovidio, Is man the measure of all things? A social cognitive account of androcentrism. *Pers. Soc. Psychol. Rev.* ***23***, 307-331 (2019).

34. M. Gustafsson Sendén, T. Lindholm, S. Sikström, Biases in news media as reflected by personal pronouns in evaluative contexts, *Soc. Psychol.* ***45***, 103-111 (2014).

35. J. M. Twenge, W. K. Campbell, B. Gentile, Male and female pronoun use in US books reflects women's status, 1900–2008, *Sex Roles* ***67***, 488-493 (2012).

36. C. Criado-Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men* (Abrams Press, New York, 2019).

37. M. Dusenbery, *Doing Harm: The Truth About How Bad Medicine and Lazy Science Leave Women Dismissed, Misdiagnosed, and Sick* (HarperCollins, New York, NY, 2018).

38. P. Hegarty, O. Parslow, Y. G. Ansara, F. Quick, "Androcentrism: Changing the landscape without leveling the playing field" in *The Sage Handbook of Gender and Psychology,* M. K. Ryan, N. R. Branscombe, Eds. (Sage, Thousand Oaks, CA, 2013), pp. 29-44.

39. B. A. Nosek, F. L. Smyth, J. J. Hansen, T. Devos, N. M. Lindner, K. A. Ranganath, C. T. Smith, K. R. Olson, D. Chugh, A. G. Greenwald, M. R. Banaji, Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur. Rev. Soc. Psychol.* ***18***, 36-88 (2007).

40. A. H. Eagly, C. Nater, D. I. Miller, M. Kaufmann, S. Sczesny, Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *Am. Psychol.* ***75***, 301-315 (2020).

41. "May 2017 Crawl Archive Now Available" (Common Crawl, 2017; http://commoncrawl.org/2017/06/)

42. M. A. Mehmood, H. M. Shafiq, A. Waheed, "Understanding regional context of World Wide Web using common crawl corpus" in *Proceedings of the IEEE 13th Malaysia International Conference on Communications (MICC)*, 164-169 (2017).

43. J. Dodge, *et al.,* https://arxiv.org/abs/2104.08758 (2021).

44. J. Cohen, A power primer. *Psychol. Bull.* ***112***, 155-159 (1992).

45. A. Lovakov, E. R. Agadullina, Empirically derived guidelines for effect size interpretation in social psychology. *Eur. J. Soc. Psychol.* ***51***, 485-504 (2021).

46. G. Saucier, K. Iurino, High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *J. Pers. Soc. Psychol.* ***119***, 1188-1219 (2020).

47. E. L. Haines, K. Deaux, N. Lofaro, The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychol. Women Q.* ***40***, 353-363 (2016).

48. D. A. Prentice, E. Carranza, What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychol. Women Q.* ***26***, 269-281 (2002).

49. J. E. Williams, D. L. Best, *Measuring Sex Stereotypes: A Multination Study* (Sage, Thousand Oaks, CA, 1990).

50. S. L. Bem, The measurement of psychological androgyny. *J. Consult. Clin. Psychol.* ***42***, 155-162 (1974).

51. A. Hoyle, *et al.,* https://arxiv.org/abs/1906.04760 (2019).

52. Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH, *Scientific Data* ***6***, 52 (2019).

53. D. DeFranza, H. Mishra, A. Mishra, How language shapes prejudice against women: An examination across 45 world languages. *J Pers. Soc. Psychol.* ***119***, 7–22 (2020).

54. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: The implicit association test. *J Pers. Soc. Psychol.* ***74***, 1464-1480 (1998).

55. "Facts and figures: Leadership and political participation" (UN Women, 2017; https://www.unwomen.org/en/what-we-do/leadership-and-political-participation/facts-and-figures).

56. J. Mu, S. Bhat, P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations" in *Proceedings of the 6th International Conference on Learning Representations* (2018).

57. R. T. Anderson, "Transgender ideology is riddled with contradictions" (The Heritage Foundation, 2018; https://www.heritage.org/gender/commentary/transgender-ideology-riddled-contradictions-here-are-the-big-ones).

58. A. Byrne, Are women adult human females? *Philos.* ***177***, 3783-3803 (2020).

59. R. Dembroff, Beyond binary: Genderqueer as critical gender kind. *Philos. Impr.* ***20***, 1-23 (2020).

60. K. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Rev.* ***43***, 1241-1299 (1990).

61. V. Purdie-Vaughns, R. P. Eibach, Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles* ***59***, 377-391 (2008).

62. S. G. S. Shah, R. Dam, M. J. Milano, L. D. Edmunds, L. R. Henderson, C. R. Hartley, O. Coxall, P. V. Ovseiko, A. M. Buchan, V. Kiparoglou, Gender parity in scientific

authorship in a National Institute for Health Research Biomedical Research Centre: A bibliometric analysis, *BMJ Open* ***11***, (2021).

63. E. H. Wojcik, J. R. Saffran, Toddlers encode similarities among novel words from meaningful sentences, *Cognition* ***138***, 10-20 (2015).

64. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants, *Science* ***274***, 1926-1928 (1996).

65. P. Nayak, "Understanding searches better than ever" (Google, 2019; https://blog.google/products/search/search-language-understanding-bert/)

66. "Embeddings@Twitter" (Twitter, Revenue Platform, 2018; https://blog.twitter.com/engineering/en_us/topics/insights/2018/embeddingsattwitter)

67. A. Renduchintala, *et al.,* https://arxiv.org/abs/2104.07838 (2021).

68. E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, J. Weston, "Queens are powerful too: Mitigating gender bias in dialogue generation" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8173-8188 (2020).

69. C. Metz, "AI is transforming Google search. The rest of the web is next" (WIRED Magazine, 2016; https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/).

70. P. Olson, "The algorithm that helped Google Translate become sexist" (Forbes, 2018; https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=4de9491b7daa).

71. G. Stanovsky, N. A. Smith, L. Zettlemoyer, "Evaluating gender bias in machine translation" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019).

72. A. Renduchintala, D. Diaz, K. Heafield, X. Li, M. Diab, "Gender bias amplification during Speed-Quality optimization in Neural Machine Translation" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joing Conference on Natural Language Processing (Volumne 2: Short Paper)* (2021).

73. H. Gonen, Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019).

74. R. H. Maudslay, H. Gonen, R. Cotterell, S. Teufel, "It's all in the name: Mitigating gender bias with name-based counterfactual data substitution" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (2019).

75. J. W. Pennebaker, R. J. Booth, R. L. Boyd, M. E. Francis, "Linguistic Inquiry and Word Count: LIWC2015" (Pennebaker Conglomerates, Austin, TX, 2015; www.LIWC.net).

76. F. Glen, K. Hurrell, "Technical note: Measuring gender identity" (Equality and Human Rights Commission, Manchester, UK, 2012).

77. S. M. van Anders, N. L. Caverly, M. M. Johns, Newborn bio/logics and US legal requirements for changing gender/sex designations on state identity documents. *Fem. Psychol.* ***24***, 172-192 (2014).

78. C. Felbaum, "About WordNet" [WordNet, Princeton University].

79. T. K. Koo, M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, *J. Chiropr. Med.* ***15***, 155-163 (2016).

80. L. R. Goldberg, From Ace to Zombie: Some explorations in the language of personality, *J. Pers. Assess.* ***1***, 203-234 (1982).

81. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). "lmerTest package: Tests in linear mixed effects models." *Journal of Statistical Software, 82*(13), 1–26.

82. B. de Raad, D. P. Barelds, E. Levert, F. Ostendorf, B. Mlačić, L. D. Blas, M. Herbícková, Z. Szirmák, M. Perugini, A. T. Church, M. S. Katigbak. Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *J. of Pers. Soc. Psychol.* ***98***, 160-173 (2010).

83. L. R. Goldberg, An alternative "description of personality": The big-five factor structure, *J. Pers. Soc. Psychol.* ***59***, 1216-1229 (1990).

84. L. R. Goldberg, The development of markers of the Big-Five factor structure, *Psychol. Assess.* ***4***, 26-42 (1992).

85. W. K. Hofstee, B. De Raad, L. R. Goldberg, Integration of the big five and circumplex approaches to trait structure, *J. of Pers. Soc. Psychol.* ***63***, 146-163 (1992).

86. G. Saucier, L. R. Goldberg, The language of personality: Lexical perspectives on the five-factor model. In *The Five-Factor Model of Personality: Theoretical Perspectives*, J. S Wiggins, Ed. (Guilford Press, 1996), pp. 21-50.

87. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

Acknowledgements: The authors thank three anonymous reviewers, Brenden Lake, Gregory Murphy, Jon Willits, Laurens van der Maaten, Mark Tygert, Mahzarin Banaji, Tessa Charlesworth, Y-Lan Boureau, and members of the New York University Cognitive Development Lab, including Jessica Gladstone, Katharina Block, Mark Bowker, Melis Muradoglu, and Vivian Liu, for their insightful comments on previous versions of this research. The authors also thank Ingrid Friedman for her assistance with manuscript preparation.

Author contributions: April Bailey: Conceptualization, Project Administration, Methodology, Investigation, Formal Analysis, Validation, Data Curation, Visualization, Writing – Original Draft, Writing – Revisions, Writing – Review & Editing. **Adina Williams:** Conceptualization, Methodology, Software, Investigation, Writing – Review & Editing. **Andrei Cimpian:** Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Competing interest: The authors declare that they have no competing interests.

Data and materials availability: All data, analysis scripts, and preregistrations are publicly available at: https://osf.io/3ebqh/?view_only=feeafaf7209a4a0b9f8435273c1a4a4b. All materials are available in the supplementary materials.

Science Advances

AAAS

Supplementary Materials for

Based on billions of words on the internet, PEOPLE = MEN

April H. Bailey, Adina Williams, Andrei Cimpian

Address correspondence to: April H. Bailey, ab9490@nyu.edu

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Data files, analysis scripts, and preregistrations available at:
https://osf.io/3ebqh/?view_only=feeafaf7209a4a0b9f8435273c1a4a4b

Supplementary Materials for
Based on billions of words on the internet, PEOPLE = MEN

Table of Contents

Supplementary Materials for

Based on billions of words on the internet, PEOPLE = MEN

Study 1

Additional Methodological Details of the Findings Reported in the Main Text in Study 1

Our final word lists consisted of 30 words for PEOPLE (Table S1), 38 words for WOMEN (Table S2), and 36 words for MEN (Table S2).

Additional Analytic Details of the Findings Reported in the Main Text in Study 1

As reported in the main text, we found that generic words for PEOPLE were more similar in their usage to words for MEN (M = 0.16, SD = 0.04) than to words for WOMEN (M = 0.14, SD = 0.04), B = 0.02, SE < 0.01, p < .001, d = 0.47. This result was the output of a mixed-effects linear regression with gender (words for MEN vs. words for WOMEN; categorical variable) predicting cosine similarity to words for PEOPLE, with a random intercept for each word for PEOPLE.

All mixed-effects linear regressions were conducted using R and lmerTest; p values were obtained with the Satterthwaite's degrees of freedom method (*81*). Here and throughout, the d values are coefficients from regressions with standardized outcome variables. That is, the d values represent the mean differences between words for MEN and words for WOMEN in standard deviation units.

Table S1

List of Words for PEOPLE in Study 1 With Average Fit Ratings

	Coder rating		Coder rating		Coder rating
beings	-	individual	9.00	somebody	9.00
citizenry	5.17	individuals	9.00	someone	9.00
folk	7.00	masses	8.17	soul	8.17
folks	7.67	mortal	6.50	souls	7.17
group	-	mortals	6.83	their	8.83
human	9.00	multitude	5.67	them	8.83
humanity	9.00	multitudes	6.17	they	8.83
humankind	8.50	people	9.00	tribe	5.50
humanness	6.83	person	9.00	tribes	5.50
humans	9.00	somebodies	7.17	yall	8.00

Note. The words that do not have ratings were added after the rating study was conducted because of suggestions from the coders as described in the Materials and Methods sections of the main text.

Table S2

List of Words for WOMEN and Words for MEN in Studies 1-3 With Average Fit Ratings

Words for WOMEN				Words for MEN			
	Coder rating		Coder rating		Coder rating		Coder rating
female	8.33	lady's	8.67	boy	8.67	lad	6.33
female's	-	lass	6.17	boy's	8.33	laddie	-
females	8.33	lassie	6.00	boyhood	7.83	male	8.83
feminine	8.67	ma'am	8.33	boyish	7.67	male's	8.33
femininity	8.83	maam	7.83	boys	9.00	males	9.00
femme	-	madam	8.33	fella	5.33	man	8.83
gal	6.83	maiden	8.67	gent	6.33	man's	8.67
gals	7.00	missus	8.67	gentleman	9.00	manhood	-
girl	8.83	ms	8.33	gentleman's	-	manly	8.67
girl's	7.00	schoolgirl	6.17	gentlemen	9.00	masculine	8.50
girlhood	7.33	schoolgirls	-	gents	7.17	masculinity	8.67
girlish	7.50	she	7.83	guy	7.33	men	9.00
girls	8.17	shes	-	guys	-	mens	8.67
girly	7.50	woman	9.00	he	9.00	mister	8.33
her	9.00	woman's	8.33	hes	8.83	mr	8.83
hers	9.00	womanhood	9.00	him	8.83	schoolboy	7.50
herself	9.00	womanly	7.50	himself	9.00	schoolboys	6.67
ladies	8.83	women	9.00	his	8.83	sir	8.33
lady	8.83	womens	-				

Note. The words that do not have ratings were added after the rating study was conducted, either because of suggestions from the coders or to parallel an other-gender counterpart already on the list as described in the Materials and Methods sections of the main text.

Study 2A

Additional Methodological Details of the Findings Reported in the Main Text in Study 2A

Our final word list consisted of 538 trait words—145 of which had gender stereotypicality designations (Table S3). These trait words have been widely used to study human personality (*48*, *82-86*). The gender words—i.e., words for WOMEN and words for MEN—were the same as in Study 1 (Table S2).

Additional Analytic Details of the Findings Reported in the Main Text in Study 2A

As reported in the main text regarding our first prediction, we found that trait words were more similar in their usage to words for MEN (M = 0.14, SD = 0.04) than to words for WOMEN (M = 0.13, SD = 0.04), B = 0.01, SE < 0.01, p < .001, d = 0.29. This result was the output of a mixed-effects linear regression with gender (words for MEN, words for WOMEN; categorical variable) predicting cosine similarity to traits, with a random intercept for each trait word.

As reported in the main text regarding our second prediction, we found that the cosine similarity of the 145 trait words (a subset of the 538 trait words) with words for MEN and, separately, with words for WOMEN depended on gender stereotypicality of the traits (i.e., there was an interaction), B = 0.02, SE < 0.01, p < .001. Specifically, there was no statistically significant difference between words for MEN and traits stereotypical of men (M = 0.14, SD = 0.04) and traits stereotypical of women (M = 0.14, SD = 0.05), B < 0.01, SE = 0.01, p = .733, d = 0.06. In contrast, words for WOMEN were more similar to traits designated as stereotypical of women (M = 0.14, SD = 0.05) than to traits stereotypical of men (M = 0.13, SD = 0.04), B = –0.02, SE = 0.01, p = .039, d = –0.34. This finding was the output of a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), trait stereotypicality (stereotypical of men, stereotypical of women; categorical variable), and their interaction predicting cosine similarity to traits, with a random intercept for each trait word. We followed up on the significant interaction within the same model using simple slopes analyses.

Table S3

List of Trait Words in Study 2A With Gender Stereotypicality Designations

Trait	Gender	Trait	Gender		Gender	Trait	Gender
abrupt	-	eager	-	lazy	-	silent	-
absent-minded	-	earnest	-	lenient	-	simple	-
abusive	M[a]	earthy	-	lethargic	-	sincere	W
accommodating	W[b]	easygoing	M	liberal	-	skeptical	-
acquiescent	-	eccentric	-	logical	M	sloppy	-
acquisitive	-	economical	M	lonely	-	slothful	-
active	-	effervescent	-	loyal	-	sluggish	-
adaptable	-	efficient	-	lustful	W	sly	-
adventurous	M	egocentric	M	magnetic	-	smart	-
affectionate	W	egotistical	M	malleable	-	smug	M
aggressive	M	eloquent	-	manipulative	-	snobbish	-
agreeable	W	emotional	W	mannerly	-	sociable	-
aimless	-	empathic	W	masochistic	-	social	W
alert	-	energetic	-	mature	-	soft	W
aloof	-	enterprising	-	meddlesome	-	soft-hearted	-
altruistic	W	enthusiastic	-	meditative	-	solicitous	-
ambitious	M	envious	-	meek	-	somber	-
amiable	-	erratic	-	melancholy	-	sophisticated	-
analytical	-	ethical	-	mercenary	-	spirited	-
angry	-	exacting	-	merry	W	spontaneous	-
animated	-	excitable	W	meticulous	-	steady	-
antagonistic	-	exhibitionistic	-	mischievous	-	stern	-
anxious	W	explosive	M	miserly	-	stingy	-
apathetic	-	expressive	-	modest	W	straightforward	M
argumentative	-	extravagant	-	moody	W	strict	-
articulate	-	extroverted	-	moral	-	strong	M
artistic	W	exuberant	-	moralistic	-	stubborn	-
assertive	M	fair	-	morose	-	subjective	-
assured	-	fastidious	-	naive	W	submissive	W
astute	-	fault-finding	W	narrow-minded	-	suggestive	W
attractive	-	fearful	W	natural	-	superstitious	-
austere	-	fidgety	-	neat	-	surly	-
autocratic	-	finicky	-	negativistic	-	suspicious	-
autonomous	M	firm	M	negligent	-	sympathetic	W
bashful	W	flamboyant	-	nervous	-	systematic	-
belligerent	-	flexible	-	nonchalant	M	tactful	-
benevolent	-	flippant	-	noncommittal	M	tactless	-
bigoted	M	flirtatious	-	nonconforming	-	talkative	W
bitter	-	folksy	-	nonpersistent	-	temperamental	W
bland	-	foolhardy	-	nonreligious	-	tempestuous	-
blase	-	forceful	M	nosey	W	tenacious	M
boastful	M	foresighted	-	objective	-	terse	-
boisterous	-	forgetful	W	obliging	-	theatric	W
bold	M	formal	-	obsessive	-	thorough	-
bossy	-	forward	M	obstinate	-	thoughtful	-
brave	M	frank	M	open-minded	-	thoughtless	-
bright	-	fretful	-	opinionated	W	thrifty	-
brilliant	M	friendly	W	opportunistic	-	timid	W
bullheaded	M	frivolous	-	optimistic	-	tolerant	-
buoyant	-	generous	W	orderly	W	touchy	W
callous	-	genial	-	organized	W	tough	M
candid	M	glib	-	outspoken	M	traditional	M
cantankerous	-	glum	-	particular	-	tranquil	-
carefree	-	gossipy	W	passionate	-	transparent	-
careful	-	greedy	-	passionless	-	trustful	-
careless	M	gregarious	-	passive	-	truthful	-
casual	-	gruff	-	patient	-	unadventurous	-
caustic	-	grumpy	M	patronizing	M	unaffectionate	M
cautious	-	guarded	-	peaceful	-	unaggressive	-
charitable	W	gullible	W	perceptive	-	unambitious	-
cheerful	W	haphazard	-	perfectionistic	W	unassuming	-
circumspect	-	happy	-	persistent	M	unattractive	-
clever	-	happy-go-lucky	-	pessimistic	-	uncharitable	-
coarse	-	hard	-	philosophical	M	uncommunicative	-
cold	-	harsh	-	placid	-	uncompetitive	-
combative	-	hearty	-	playful	-	unconscious	-

communicative	-	helpful	W	pleasant	W	unconventional	-
compassionate	W	helpless	-	poised	W	uncooperative	-
competitive	-	high-strung	-	polite	-	uncouth	-
complex	-	homespun	-	pompous	M	uncreative	-
compliant	W	honest	-	possessive	M	uncritical	-
compulsive	-	humble	-	practical	M	undemanding	-
conceited	-	humorless	W	precise	-	undependable	-
conceitless	-	humorous	M	predictable	-	underhanded	-
conciliatory	-	hypocritical	-	prejudiced	-	understanding	-
concise	-	idealist	W	pretentious	-	unemotional	M
condescending	-	ignorant	-	prideless	-	unenergetic	-
confident	M	ill-tempered	-	principled	-	unenvious	-
conscientious	-	illogical	W	progressive	-	unexcitable	-
conservative	-	imaginative	-	prompt	-	unforgiving	-
considerate	W	imitative	-	proud	M	unfriendly	-
consistent	-	immature	M	provincial	-	ungracious	-
contemplative	-	immodest	-	prudish	W	unimaginable	-
contemptuous	-	impartial	-	punctual	-	uninhibited	-
controlling	-	impatient	-	purposeful	-	uninquisitive	-
conventional	-	imperceptive	-	quarrelsome	-	unintellectual	-
cooperative	-	impersonal	-	quiet	-	unintelligent	-
cordial	-	impertinent	-	rambunctious	M	unkind	-
cosmopolitan	-	imperturbable	-	rash	-	unmoralistic	-
courageous	M	impetuous	-	rational	M	unobservant	-
courteous	-	impolite	-	reasonable	M	unpredictable	-
cowardly	-	impractical	-	rebellious	M	unprejudiced	-
crabby	-	impudent	-	reckless	-	unpretentious	-
crafty	-	impulsive	M	refined	-	unprogressive	-
cranky	-	inarticulate	-	relaxed	-	unreflective	-
creative	-	inconsiderate	M	reliable	-	unreliable	-
critical	-	inconsistent	-	religious	-	unrestrained	-
crude	M	indecisive	W	reserved	-	unruly	-
cruel	-	indefatigable	-	respectful	-	unscrupulous	-
cultured	-	independent	M	responsible	-	unselfconscious	M
cunning	-	indirect	W	restless	-	unselfish	-
curious	-	indiscreet	-	restrained	-	unsociable	-
curt	-	individualistic	-	reverent	-	unsophisticated	-
cynical	-	indulgent	-	rigid	-	unstable	W
daring	M	industrious	-	romantic	W	unsympathetic	M
deceitful	-	inefficient	-	rough	M	unsystematic	-
decisive	-	informal	-	rude	-	untalkative	-
deep	-	informative	-	ruthless	W	unvindictive	-
defensive	-	ingenious	-	sarcastic	-	urbane	-
deliberate	-	inhibited	-	scatter-brained	-	vague	-
demanding	-	inner-directed	-	scornful	-	vain	W
demonstrative	-	innovative	M	scrupulous	-	verbal	-
dependable	-	inquisitive	-	seclusive	-	verbose	-
dependent	W	insecure	W	secretive	-	versatile	-
detached	M	insensitive	M	sedate	M	vibrant	-
devil-may-care	-	insightful	-	self-critical	-	vigilant	-
devious	M	insincere	-	self-disciplined	-	vigorous	M
dignified	-	intellectual	-	self-effacing	W	vindictive	W
diplomatic	-	intelligent	-	self-examining	-	vivacious	W
direct	M	intense	-	self-indulgent	-	volatile	W
disagreeable	-	intolerant	-	self-pity	-	warm	W
discreet	-	introspective	W	self-satisfied	-	wary	-
dishonest	-	introverted	-	self-seeking	-	wasteful	-
disorderly	M	intrusive	-	selfish	-	weak	W
disorganized	-	inventive	M	selfless	W	weariless	-
disrespectful	-	irreverent	-	sensitive	W	wise	M
distrustful	-	irritable	-	sensual	W	wishy-washy	-
docile	W	jaded	-	sentimental	W	withdrawn	-
dogmatic	-	jealous	-	serious	M	witty	-
doleful	-	jovial	-	servile	-	wordy	-
dominant	M	joyless	-	sexy	-	worldly	-
domineering	-	judicious	-	shallow	W	zealous	-
down-to-earth	-	kind	-	short-sighted	-	zestful	-
dramatic	W	knowledgeable	-	shrewd	-		
dull	-	lax	M	shy	W		

Note. Traits from reference *46*. [a] Designated as stereotypical of men. [b] Designated as stereotypical of women.

Study 2B

Additional Methodological Details of the Findings Reported in the Main Text in Study 2B

Our final word list consisted of 178 trait words with gender stereotypicality designations (Table S4). The gender words were the same as in Study 1 (Table S2).

Additional Analytic Details of the Findings Reported in the Main Text in Study 2B

As reported in the main text with respect to our first prediction, we found that, overall, trait words were more similar in their usage to words for MEN (M = 0.15, SD = 0.05) than to words for WOMEN (M = 0.14, SD = 0.05), B = 0.01, SE < 0.01, p < .001, d = 0.19. This result was the output of a mixed-effects linear regression with gender (words for MEN, words for WOMEN; categorical variable) predicting cosine similarity to traits, with a random intercept for each trait word.

As reported in the main text with respect to our second prediction, we found that the cosine similarity of the 178 trait words with words for MEN and, separately, words for WOMEN depended on gender stereotypicality of the traits (i.e., there was an interaction), B = 0.02, SE < 0.01, p < .001. Specifically, the was no statistically significant difference between words for MEN and traits stereotypical of men (M = 0.15, SD = 0.04) compared to traits stereotypical of women (M = 0.14, SD = 0.05), B < 0.01, SE = 0.01, p = .807, d = 0.04. In contrast, words for WOMEN were more similar to traits stereotypical of women (M = 0.14, SD = 0.05) than to traits stereotypical of men (M = 0.13, SD = 0.05), B = –0.01, SE = 0.01, p = .049, d = –0.30. This result was the output of a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), trait stereotypicality (stereotypical of men, stereotypical of women; categorical variable), and their interaction predicting cosine similarity to traits, with a random intercept for each trait. We followed up on the significant interaction within the same model using simple slopes analyses.

Table S4

List of Trait Words in Study 2B With Gender Stereotypicality Designations

Trait	Gender	Trait	Gender	Trait	Gender
active	M[a,c]	forceful	M[g]	rigid	M[c]
adventurous	M[c]	forgiving	W[c]	robust	M[c]
affected	W[b,c]	friendly	W[g]	romantic	W[d]
affectionate	W[d]	frivolous	W[c]	self-confident	M[f]
aggressive	M[d]	fussy	W[c]	self-pitying	W[c]
ambitious	M[d]	gentle	W[f]	self-reliant	M[g]
analytical	M[e]	graceful	W[f]	self-righteous	M[g]
appreciative	W[c]	greedy	M[c]	self-sufficient	M[e]
arrogant	M[d]	gullible	W[g]	selfish	M[d]
assertive	M[d]	hardhearted	M[c]	sensitive	W[d]
athletic	M[d]	hardworking	M[d]	sentimental	W[c]
autocratic	M[c]	helpful	W[f]	serious	M[c]
bossy	M[c]	honest	W[d]	sexy	W[c]
broad-shouldered	M[f]	humorous	M[c]	sharp-witted	M[c]
capable	M[c]	imaginative	W[c]	short	W[f]
cautious	W[c]	impressionable	W[g]	show-off	M[c]
changeable	W[c]	independent	M[d]	shy	W[g]
charming	W[c]	indifferent	M[c]	small-boned	W[f]
cheerful	W[g]	individualistic	M[c]	smart	W[d]
childlike	W[g]	initiative	M[c]	soft	W[f]
clean	W[g]	innovative	M[d]	softhearted	W[c]
coarse	M[c]	intelligent	W[d]	solemn	M[g]
compassionate	W[d]	intense	M[g]	solid	M[f]
competitive	M[f]	interests wide	M[c]	sophisticated	W[c]
complaining	W[c]	inventive	M[c]	spiritual	W[g]
complicated	W[c]	jealous	M[g]	steady	M[c]
conceited	M[c]	kind	W[f]	stern	M[c]
confident	M[d]	lazy	M[c]	stingy	M[c]
confused	W[c]	leader	M[f]	stolid	M[c]
consistent	M[g]	level-headed	M[d]	strong	M[d]
controlling	M[g]	logical	M[d]	stubborn	M[d]
cooperative	W[g]	loud	M[c]	sturdy	M[f]
courageous	M[d]	loyal	W[g]	submissive	W[c]
creative	W[d]	melodramatic	W[g]	suggestive	W[c]
critical	W[d]	mild	W[c]	superstitious	W[g]
cruel	M[c]	modest	W[c]	sympathetic	W[e]
curious	W[c]	muscular	M[f]	talkative	W[c]
cynical	M[c]	naive	W[g]	tall	M[f]
dainty	W[f]	nervous	W[c]	tender	W[e]
decisive	M[d]	obnoxious	M[c]	timid	W[c]
delicate	W[f]	opinionated	M[c]	touchy	W[c]
demanding	M[d]	opportunistic	M[c]	tough	M[c]
dependable	M[g]	organized	W[d]	unambitious	W[c]
dependent	W[c]	outgoing	W[d]	understanding	W[f]
determined	M[c]	patient	W[g]	unfriendly	M[c]
disciplined	M[g]	pleasant	W[c]	unintelligent	W[c]
disorderly	M[c]	pleasure-seeking	M[c]	unscrupulous	M[c]
dominant	M[e]	polite	W[d]	unselfish	W[d]
dreamy	W[c]	possessive	M[d]	unstable	W[c]
emotional	W[d]	precise	M[c]	warm	W[f]
enterprising	M[c]	progressive	M[c]	weak	W[g]
excitable	W[g]	promiscuous	M[g]	well-built	M[f]
family-oriented	W[f]	proud	M[d]	well-dressed	W[f]
fashionable	W[f]	prudish	W[c]	well-mannered	W[d]
fault-finding	W[c]	quick	M[c]	wholesome	W[g]
fearful	W[c]	rational	M[g]	witty	M[c]
fickle	W[c]	realistic	M[c]	worrying	W[c]
flatterable	W[e]	rebellious	M[g]	yielding	W[g]
flirtatious	W[g]	reckless	M[c]		
foolish	W[c]	resourceful	M[c]		

[a] Designated as stereotypical of men. [b] Designated as stereotypical of women. Gender stereotypicality designation was taken from reference [c]*49*, [d]*40*, [e]*50*, [f]*47*, [g]*48*, but note that many traits were repeated across multiple sources.

Study 3

Additional Methodological Details of the Findings Reported in the Main Text in Study 3

The final word list consisted of 252 cases of verbs with male-biased vs. female-biased designations. Note that these 252 cases of verbs corresponded to 211 unique verbs; there were some repetitions based on differing valence or subject vs. object position of the gender bias as explained in the Materials and Methods section of the main text (Table S5). The gender words were the same as in Study 1 (Table S2).

Additional Analytic Details of the Findings Reported in the Main Text in Study 3

Regarding our first prediction, as reported in the main text, we found that verbs were overall more similar in their usage to words for MEN (M = 0.11, SD = 0.04) than to words for WOMEN (M = 0.10, SD = 0.04), B = 0.01, SE < 0.01, p < .001, d = 0.26. This result was the output of a mixed-effects linear regression with gender words (words for MEN, words for WOMEN; categorical variable) predicting cosine similarity to verbs, with a random intercept for each verb.

Regarding our second prediction, as reported in the main text, we also found that the cosine similarity of the 252 verbs with words for MEN and, separately, words for WOMEN depended on gender bias of the verbs (i.e., there was an interaction), B = 0.01, SE < 0.01, p < .001. Specifically, there was no statistically significant difference between words for MEN and verbs that were male-biased (M = 0.11, SD = 0.04) compared to verbs that were female-biased (M = 0.11, SD = 0.04), B = –0.01, SE = 0.01, p = .128, d = –0.20. In contrast, words for WOMEN were more similar to female-biased verbs (M = 0.11, SD = 0.05) than to male-biased verbs (M = 0.09, SD = 0.03), B = –0.02, SE = 0.01, p < .001, d = –0.54. This result was the output of a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), verb syntactic bias (male-biased, female-biased; categorical variable), and their interaction predicting cosine similarity to verbs, with a random intercept for each verb. We followed up on the significant interaction within the same model using simple slopes analysis.

Table S5

List of Verbs in Study 3 with Gender-Bias Designations, Valence, and Position

Verb	Gender bias	Valence	Position of bias	Verb	Gender bias	Valence	Position of bias
adore	W[a]	positive	subject	glorify	M	positive	object
allow	M[b]	positive	subject	go	W	neutral	subject
animate	M	neutral	object	gossip	W	negative	subject
appeal	M	positive	subject	grant	M	positive	subject
appear	W	neutral	subject	greet	M	positive	object
appease	M	positive	object	harm	W	negative	subject
appoint	M	neutral	object	have	W	neutral	object
argue	M	negative	subject	have	W	neutral	subject
ask	W	neutral	object	honor	M	positive	object
assure	W	neutral	object	horrify	M	negative	subject
await	M	neutral	object	hurt	W	negative	subject
be	W	neutral	subject	incarnate	M	neutral	subject
blind	M	negative	subject	inspire	M	positive	object
bore	M	negative	object	insult	W	negative	object
brave	M	positive	object	join	M	positive	object
brave	M	positive	subject	kill	M	negative	object
bribe	M	negative	object	kill	M	negative	subject
bully	M	negative	object	kiss	W	positive	object
burn	W	neutral	object	kiss	W	positive	subject
celebrate	W	positive	subject	lament	W	negative	subject
champion	W	positive	subject	laugh	W	positive	subject
cheat	M	negative	subject	leave	W	neutral	object
clap	W	neutral	subject	like	W	positive	object
clear	M	positive	object	like	W	positive	subject
clear	M	positive	subject	live	W	positive	subject
collect	M	neutral	subject	marry	W	neutral	object
come	W	neutral	subject	marry	W	positive	subject
comfort	M	positive	subject	mature	W	positive	subject
commend	M	positive	object	meet	W	positive	object
compel	M	negative	object	meet	W	positive	subject
complain	W	negative	subject	mock	M	negative	object
concern	M	negative	subject	mourn	W	negative	subject
confess	W	negative	subject	murder	M	negative	object
congratulate	M	positive	object	murder	M	negative	subject
create	W	positive	object	neglect	M	negative	subject
create	M	neutral	subject	obscure	M	negative	subject
cry	W	negative	object	offend	M	negative	object
damn	M	negative	subject	order	M	negative	object
dance	W	positive	subject	overrun	W	negative	subject
deceive	M	negative	object	pay	M	neutral	object
defeat	M	negative	object	pay	M	neutral	subject
denounce	M	negative	object	persecute	W	negative	object
denounce	M	negative	subject	persecute	W	negative	subject
deny	M	negative	object	play	W	positive	object
depose	M	neutral	object	play	W	positive	subject
deprive	M	negative	object	pour	W	neutral	object
deprive	M	negative	subject	praise	M	positive	object
destroy	M	negative	object	praise	M	positive	subject
direct	M	neutral	object	present	W	neutral	object
dispute	M	negative	subject	present	M	neutral	subject
distract	W	negative	object	pretend	M	neutral	subject
drag	W	negative	object	prevent	M	neutral	object
dress	W	neutral	subject	promise	M	positive	subject
drown	W	negative	object	prompt	M	neutral	subject
duplicate	M	neutral	subject	prosper	M	positive	subject
elect	M	neutral	object	prostrate	M	neutral	subject
encourage	M	positive	subject	protect	W	positive	object
enrage	M	negative	object	protect	M	positive	subject
enrich	M	positive	object	protest	M	negative	subject
entertain	W	positive	object	rape	W	negative	object
equal	M	neutral	object	reach	M	neutral	object
escape	M	neutral	object	reach	M	neutral	subject
escape	M	neutral	subject	rescue	M	positive	subject
escort	W	neutral	object	respect	M	positive	object
espouse	W	neutral	object	respect	M	positive	subject

exalt	W	positive	subject	restore	M	positive	object
exalt	M	positive	object	reward	M	positive	object
excel	W	positive	object	reward	M	positive	subject
exchange	W	neutral	object	rush	M	neutral	subject
excite	M	positive	object	saw	W	neutral	object
exclaim	W	neutral	object	scare	W	negative	object
excommunicate	M	neutral	object	scold	W	negative	subject
exempt	M	neutral	object	scold	M	negative	object
expel	M	neutral	object	scream	W	negative	object
expel	M	negative	subject	scream	W	negative	subject
exploit	W	negative	object	see	W	neutral	object
expose	W	neutral	object	set	M	neutral	object
extend	W	neutral	subject	set	M	neutral	subject
extol	W	positive	subject	shame	W	neutral	object
extol	M	positive	object	shock	W	negative	object
eye	W	positive	object	shock	M	negative	subject
facilitate	W	positive	subject	shop	M	neutral	object
fade	W	neutral	object	signal	W	neutral	object
fail	M	negative	object	smile	W	positive	subject
faint	W	neutral	subject	sniff	W	neutral	subject
fall	W	neutral	subject	speak	M	neutral	object
fan	W	positive	subject	spin	W	neutral	subject
fascinate	W	positive	subject	steal	W	negative	object
fatigue	W	negative	subject	strike	M	neutral	subject
favor	M	positive	subject	strut	W	neutral	object
favour	M	positive	subject	succeed	M	positive	object
fear	M	negative	object	succeed	M	positive	subject
fear	M	negative	subject	suffer	W	negative	object
feature	W	neutral	object	summon	M	neutral	object
fee	W	neutral	subject	support	M	positive	subject
feign	W	negative	subject	surpass	W	positive	subject
felicitate	W	positive	subject	take	W	neutral	object
fell	W	neutral	subject	tarry	M	neutral	subject
fertilize	W	neutral	object	tease	W	negative	object
fertilize	W	neutral	subject	temper	M	negative	subject
fight	M	neutral	object	terrify	W	negative	object
fill	W	neutral	subject	thank	M	positive	object
find	W	neutral	subject	threaten	M	negative	subject
fit	M	positive	object	tip	M	neutral	object
fit	M	positive	subject	treat	W	positive	object
flatter	M	positive	object	treat	M	positive	subject
flourish	M	positive	subject	unmake	M	neutral	object
fly	W	neutral	subject	uphold	M	positive	object
follow	M	neutral	object	use	M	neutral	object
fondle	W	positive	object	vanish	W	neutral	subject
forbid	W	negative	object	violate	W	negative	object
forbid	M	negative	subject	visit	W	neutral	object
found	M	neutral	object	wag	M	neutral	subject
found	M	neutral	subject	want	M	neutral	subject
freeze	W	positive	subject	warm	M	positive	subject
freeze	M	neutral	subject	wear	W	neutral	subject
fright	W	negative	object	weep	W	negative	object
fright	M	negative	subject	weep	W	negative	subject
frighten	W	negative	object	welcome	M	positive	object
front	W	neutral	subject	welcome	M	positive	subject
frustrate	M	negative	subject	win	W	positive	object
gasp	W	negative	subject	win	M	positive	subject
gentle	M	positive	object	wish	W	positive	object
get	W	negative	subject	wish	M	positive	subject
giggle	W	positive	subject	woo	W	positive	object
give	W	positive	subject	worry	W	negative	subject

Note. Traits from reference *51*. [a] Designated as female-biased. [b] Designated as male-biased.

Exploratory Analyses in Study 3

The list of 252 verbs was taken from prior work that, in addition to identifying the syntactic gender bias of each verb, indicated the valence (i.e., sentiment) of the verb as positive, negative, or neutral and indicated whether the verb's gender bias occurred with arguments in the subject or object position (*51*). In two sets of exploratory analyses, we tested whether the findings in the present study were further moderated by valence or by the syntactic position in which the gender bias occurred.

Valence of the Verb. To test the potential moderating effect of valence, we first compared a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), valence of the verb (negative, positive, or neutral; categorical variable), and their interaction terms predicting cosine similarity to verbs, with a random intercept for each verb, to an identical model that omitted the interaction terms. There was no evidence that the model with interaction terms explained significantly more variance than the model without the interaction terms, $\chi^2(2) < 0.01$, $p > .999$, indicating that valence did not moderate the difference between words for MEN and words for WOMEN.

Second, we compared a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), verb syntactic gender bias (male-biased, female-biased; categorical variable), verb valence (negative, positive, or neutral; categorical variable), and their interaction terms predicting cosine similarity to verbs, with a random intercept for each verb, to an identical model but without the higher-order valence interaction terms. There was no evidence that the model with the valence interaction terms explained more variance, $\chi^2(6) < 0.01$, $p > .999$. Thus, in both of these analyses, there was no evidence that the valence of the verb moderated either the overall difference between words for MEN and words for WOMEN or the interaction effect between the gender words and the verb syntactic gender bias.

Syntactic Position of the Verb's Gender Bias. To test the potential moderating effect of the syntactic position in which the gender bias occurred for the verbs, we first conducted a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), verb syntactic position of the bias (subject, object; categorical variable), and their interaction term predicting cosine similarity to verbs, with a random intercept for each verb. The interaction between gender and syntactic position was not significant, $B < 0.01$, $SE < 0.01$, $p = .142$, indicating that synaptic position did not moderate the difference between words for MEN and words for WOMEN.

Second, we conducted a mixed-effects linear regression with gender word (words for MEN, words for WOMEN; categorical variable), verb syntactic gender bias (male-biased, female-biased; categorical variable), syntactic position of the bias (subject, object; categorical variable), and their interaction terms. The interaction between gender word, verb syntactic gender bias, and syntactic position was not significant, $B < 0.01$, $SE = 0.01$, $p = .722$. Thus, in both of these analyses, there was no evidence that the syntactic position of the gender bias for each verb moderated either the overall difference between words for MEN and words for WOMEN or the interaction effect between the gender words and the verb syntactic gender bias.

Preregistered Replication Studies

Overview of Replication Studies

We conducted direct, preregistered replications of Studies 1-3. Each replication used identical lists of words and other procedures to Studies 1-3, respectively, with one exception: We used a different set of word embeddings. The goal of these replications was to test whether the present findings are robust to incidental details in the algorithms used to create the word embeddings.

In Studies 1-3 (main text), we used 300-dimensional fastText embeddings extracted from the Common Crawl corpus. For the present replication studies, we used 300-dimensional Global Vectors for Word Representation (GloVe) embeddings (*7*), also trained on the Common Crawl corpus.

For these replications, we preregistered our hypotheses, methods, and analytic approach, including control analyses and robustness checks reported in a subsequent section (see pp. 14-25), prior to retrieving and analyzing the word embeddings (https://osf.io/3ebqh/?view_only=feeafaf7209a4a0b9f8435273c1a4a4b).

Replication of Study 1

We compared words for PEOPLE to words for MEN and to words for WOMEN using the same mixed-effects linear regression described in Study 1. With this different set of word embeddings, we replicated Study 1 and found that words for PEOPLE were more similar in their use to words for MEN (M = 0.19, SD = 0.06) than to words for WOMEN (M = 0.15, SD = 0.04), B = 0.04, SE < 0.01, p < .001, d = 0.67.

Replication of Study 2A

To test our first prediction that, overall, trait words would be more similar in their usage to words for MEN than to words for WOMEN, we used the same multilevel model described in Study 2A. We replicated Study 2A and found that trait words were more similar in their usage to words for MEN (M = 0.14, SD = 0.06) than to words for WOMEN (M = 0.13, SD = 0.06), B = 0.02, SE < 0.01, p < .001, d = 0.26.

To test our second prediction that there would be an asymmetry in gender-stereotypical associations reflected in collective concepts, we conducted the same mixed-effects linear regression described in Study 2A. We again replicated Study 2A and found that the similarity between the words for MEN and WOMEN and the trait words depended on the gender stereotypicality of the traits (i.e., there was an interaction), B = 0.03, SE < 0.01, p < .001. Specifically, words for MEN did not differ significantly in their similarity to traits stereotypical of men (M = 0.16, SD = 0.06) and to traits stereotypical of women (M = 0.16, SD = 0.06), B < 0.01, SE = 0.01, p = .650, d = 0.07. In contrast, words for WOMEN were more similar to traits stereotypical of women (M = 0.15, SD = 0.06) than to traits stereotypical of men (M = 0.13, SD = 0.05), B = –0.02, SE < 0.01, p = .033, d = –0.35.

Replication of Study 2B

We note one departure from the preregistration of this replication study. The preregistration indicates that we will test 180 traits; however, in the present replication study (as in Study 2B), we analyzed 178 traits because we removed the traits “feminine” and “masculine,” which appeared in our list of gender words (Table S2). This was the only departure from the preregistration.

To test our first prediction that, overall, trait words would be more similar in their usage to words for MEN than to words for WOMEN, we used the same mixed-effects linear regression described in Study 2B. We replicated Study 2B and found that trait words were more similar to words for MEN (M = 0.16, SD = 0.06) than to words for WOMEN (M = 0.15, SD = 0.06), B = 0.02, SE < 0.01, p < .001, d = 0.28.

To test our second prediction that there would be an asymmetry in gender-stereotypical associations reflected in collective concepts, we conducted the same mixed-effects linear regression described in Study 2B. We again replicated Study 2B and found that the similarity between the words for MEN and words for WOMEN and the trait words depended on gender stereotypicality of the traits (i.e., there was an interaction), B = 0.02, SE < 0.01, p < .001. Specifically, words for MEN did not differ significantly in their similarity traits stereotypical of men (M = 0.16, SD = 0.06) and to traits stereotypical of women (M = 0.17, SD = 0.06), B = –0.01, SE = 0.01, p = .237, d = –0.17. In contrast, words for WOMEN were more similar to traits stereotypical of women (M = 0.16, SD = 0.06) than to traits stereotypical of men (M = 0.13, SD = 0.05), B = –0.03, SE = 0.01, p < .001, d = –0.55.

Replication of Study 3

To test our first prediction that, overall, verbs would be more similar in their usage to words for MEN than to words for WOMEN, we used the same mixed-effects linear regression described in Study 3. We replicated Study 3 and found that verbs were more similar to words for MEN (M = 0.16, SD = 0.06) than to words for WOMEN (M = 0.14, SD = 0.06), B = 0.02, SE < 0.01, p < .001, d = 0.40.

To test our second prediction that there would be an asymmetry in gender-stereotypical associations reflected in collective concepts, we conducted the same mixed-effects linear regression described in Study 3. We again replicated Study 3 and found that the similarity between the words for MEN and WOMEN and the verbs depended on the gender bias of the verbs (i.e., there was an interaction), B = 0.02, SE < 0.01, p < .001. As in Study 3, words for WOMEN were more similar to female-biased verbs (M = 0.15, SD = 0.06) than to male-biased verbs (M = 0.12, SD = 0.05), B = –0.04, SE = 0.01, p < .001, d = –0.66. Unlike Study 3, we also found that words for MEN were more similar to female-biased verbs (M = 0.17, SD = 0.06) than to male-biased verbs (M = 0.15, SD = 0.05), B = –0.02, SE = 0.01, p = .008, d = –0.35, but note that this effect for words for MEN was significantly weaker than the same effect for words for WOMEN given the significant interaction. This last finding about words for MEN is a minor departure from Study 3, but the overall pattern of results is consistent between the two studies because there was again evidence for an asymmetry in gender-stereotypical associations and specifically for stronger gender-stereotypical associations about WOMEN than about MEN in collective concepts.

Control Analyses and Robustness Checks

Overview of Control Analyses and Robustness Checks

The results of Studies 1-3 were robust to a variety of control analyses and robustness checks. These included the following, each of which is described in greater detail below and was preregistered for the replication studies: (a) in Study 1, adding weights to the analysis such that the words for PEOPLE that were rated as more representative of the concept by coders were weighted more heavily; (b) in Studies 1-3, removing *masculine generic* words and their counterparts and recomputing the analyses; (c) in Studies 1-3, conducting "leave one out" analyses for the key result; (d) in Studies 1-3, conducting a permutation test of the key result; (e) relevant to Studies 1-3, testing for potential differences in word frequencies of the gender words; and (f) in Studies 2A, 2B, and 3, conducting *word-embedding association tests* (WEAT). Finally, we also (g) tested the generalizability of the critical findings in Study 1 to a more specialized domain. We replicated the results of Study 1 in the biomedical domain using word embeddings trained on biomedical and clinical text instead of undifferentiated text on the internet (i.e., the Common Crawl, which was the basis of the studies reported in the main text).

A. Weighted Analysis (Study 1 and Replication)

We conducted a supplementary analysis in which words that were rated by coders as more fitting or representative of the concept PEOPLE were weighed more heavily in the analysis. This was done just in Study 1 because the list of words for PEOPLE was generated for the purposes of this study and was relatively small compared to the list of traits and verbs in Studies 2 and 3, respectively.

As described in detail in the Materials and Methods section of the main text, six coders who were unaware of our hypotheses rated each of the words for PEOPLE on their fit with the underlying concept. We standardized these scores, added a constant (so that they are all positive), and then used these as level-2 (i.e., PEOPLE word-level) weights in the same mixed-effects model described previously. For the two category words added after the coding step ("beings" and "group"), for which we did not have ratings of fit with the concept, we used the average rating of all PEOPLE words because weighted analyses do not permit missing weight values.

In the weighted analysis for Study 1, we again found that words for PEOPLE were more similar in their usage to words for MEN (M = 0.16, SD = 0.04) than to words for WOMEN (M = 0.14, SD = 0.04), B = 0.02, SE < 0.01, p < .001, d = 0.49. In the preregistered replication of Study 1 using this weighted analysis, we also again found that the words for PEOPLE were more similar in their usage to words for MEN (M = 0.19, SD = 0.06) than to words for WOMEN (M = 0.15, SD = 0.04), B = 0.04, SE < 0.01, p < .001, d = 0.72.

B. Masculine Generic Analyses (All Studies)

Some of the words for MEN in our list of gender words (Table S2) are also commonly used to generically refer to people of all genders. For instance, it is common when referring to a person in general to use "he" but not "she" (*27*). These words are known as masculine generics. It was important to rule out the possibility that the results we observed in the present study were merely an artifact of having these masculine generic words in our word lists, which could have artificially inflated the similarity of MEN words and PEOPLE words.

To investigate this alternative explanation, we removed masculine generic words as well as parallel woman-specific ones from our lists (i.e., "he," "hes," "him," "himself," "his," "man," and "man's"; "she," "shes," "her," "herself," "hers," "woman," and "woman's") and re-ran all analyses for Studies 1-3 and their preregistered replications. All results across all studies were robust to removing masculine generic words (for details, see Tables S6 and S7). Thus, the findings reported in the main text are not merely due to the inclusion of masculine generics among the words for MEN in our list of gender words.

Table S6

The Difference Between Gender Words in Studies 1-3 and Replications Without the Masculine Generic Words and in the Original Results

Study	Comparison	Results without masculine generic words			Original results		
		Words for MEN *M* (*SD*)	Words for WOMEN *M* (*SD*)	*d*	Words for MEN *M* (*SD*)	Words for WOMEN *M* (*SD*)	*d*
Study 1	Similarity to PEOPLE words	0.15 (0.04)	0.14 (0.03)	0.43 ***	0.16 (0.04)	0.14 (0.04)	0.47 ***
Study 1 replication		0.17 (0.05)	0.13 (0.04)	0.76 ***	0.19 (0.06)	0.15 (0.04)	0.67 ***
Study 2A	Similarity to traits	0.14 (0.05)	0.13 (0.05)	0.25 ***	0.14 (0.04)	0.13 (0.04)	0.29 ***
Study 2A replication		0.14 (0.06)	0.12 (0.05)	0.31 ***	0.14 (0.06)	0.13 (0.06)	0.26 ***
Study 2B	Similarity to traits	0.14 (0.05)	0.13 (0.05)	0.18 ***	0.15 (0.05)	0.14 (0.05)	0.19 ***
Study 2B replication		0.16 (0.06)	0.14 (0.06)	0.32 ***	0.16 (0.06)	0.15 (0.06)	0.28 ***
Study 3	Similarity to verbs	0.14 (0.05)	0.13 (0.05)	0.21 ***	0.11 (0.04)	0.10 (0.04)	0.26 ***
Study 3 replication		0.14 (0.06)	0.12 (0.06)	0.38 ***	0.16 (0.06)	0.14 (0.06)	0.40 ***

*** $p < .001$

Table S7

The Interactions Between Gender Words and Gender Stereotypicality in Studies 2 and 3 and Replications Without the Masculine Generic Words and in the Original Results

Study	Comparison	Results without masculine generic words				Original results			
		Traits stereotypical of men *M* (*SD*)	Traits stereotypical of women *M* (*SD*)	*d*	Int.	Traits stereotypical of men *M* (*SD*)	Traits stereotypical of women *M* (*SD*)	*d*	Int.
Study 2A	Similarity to words for MEN	0.14 (0.04)	0.14 (0.05)	–0.04	***	0.14 (0.04)	0.14 (0.05)	0.06	***
	Similarity to words for WOMEN	0.13 (0.04)	0.14 (0.06)	–0.37 ***		0.13 (0.05)	0.14 (0.05)	–0.34 *	
Study 2A replication	Similarity to words for MEN	0.15 (0.06)	0.16 (0.06)	–0.04	***	0.16 (0.06)	0.16 (0.06)	0.07	***
	Similarity to words for WOMEN	0.13 (0.05)	0.15 (0.07)	–0.44 ***		0.13 (0.05)	0.15 (0.06)	–0.35 ***	
Study 2B	Similarity to words for MEN	0.14 (0.05)	0.14 (0.05)	–0.03	***	0.15 (0.04)	0.14 (0.05)	0.04	***
	Similarity to words for WOMEN	0.13 (0.05)	0.14 (0.06)	–0.30 ***		0.13 (0.05)	0.14 (0.05)	–0.30 *	
Study 2B replication	Similarity to words for MEN	0.15 (0.05)	0.16 (0.06)	–0.19 ***	***	0.16 (0.06)	0.17 (0.06)	–0.17	***
	Similarity to words for WOMEN	0.12 (0.05)	0.16 (0.06)	–0.54 ***		0.13 (0.05)	0.16 (0.06)	–0.55 ***	
		Male-biased verbs *M* (*SD*)	Female-biased verbs *M* (*SD*)			Male-biased verbs *M* (*SD*)	Female-biased verbs *M* (*SD*)		
Study 3	Similarity to words for MEN	0.10 (0.04)	0.11 (0.04)	–0.24	***	0.11 (0.04)	0.11 (0.04)	–0.20	***
	Similarity to words for WOMEN	0.08 (0.03)	0.10 (0.05)	–0.51 ***		0.09 (0.03)	0.11 (0.05)	–0.54 ***	
Study 3 replication	Similarity to words for MEN	0.13 (0.05)	0.15 (0.06)	–0.39 **	***	0.15 (0.05)	0.17 (0.06)	–0.35 **	***
	Similarity to words for WOMEN	0.10 (0.05)	0.14 (0.06)	–0.66 ***		0.12 (0.05)	0.15 (0.06)	–0.66 ***	

Note. Asterisks in the "Int." (interaction) column indicate that there was an interaction between gender words (words for MEN, words for WOMEN; categorical variable) and trait/verb gender stereotypicality (stereotypical of men, stereotypical of women; categorical variable), which provides evidence for an asymmetry in gender-stereotypical associations in collective concepts.

*** $p < .001$. ** $p < .01$. * $p < .05$

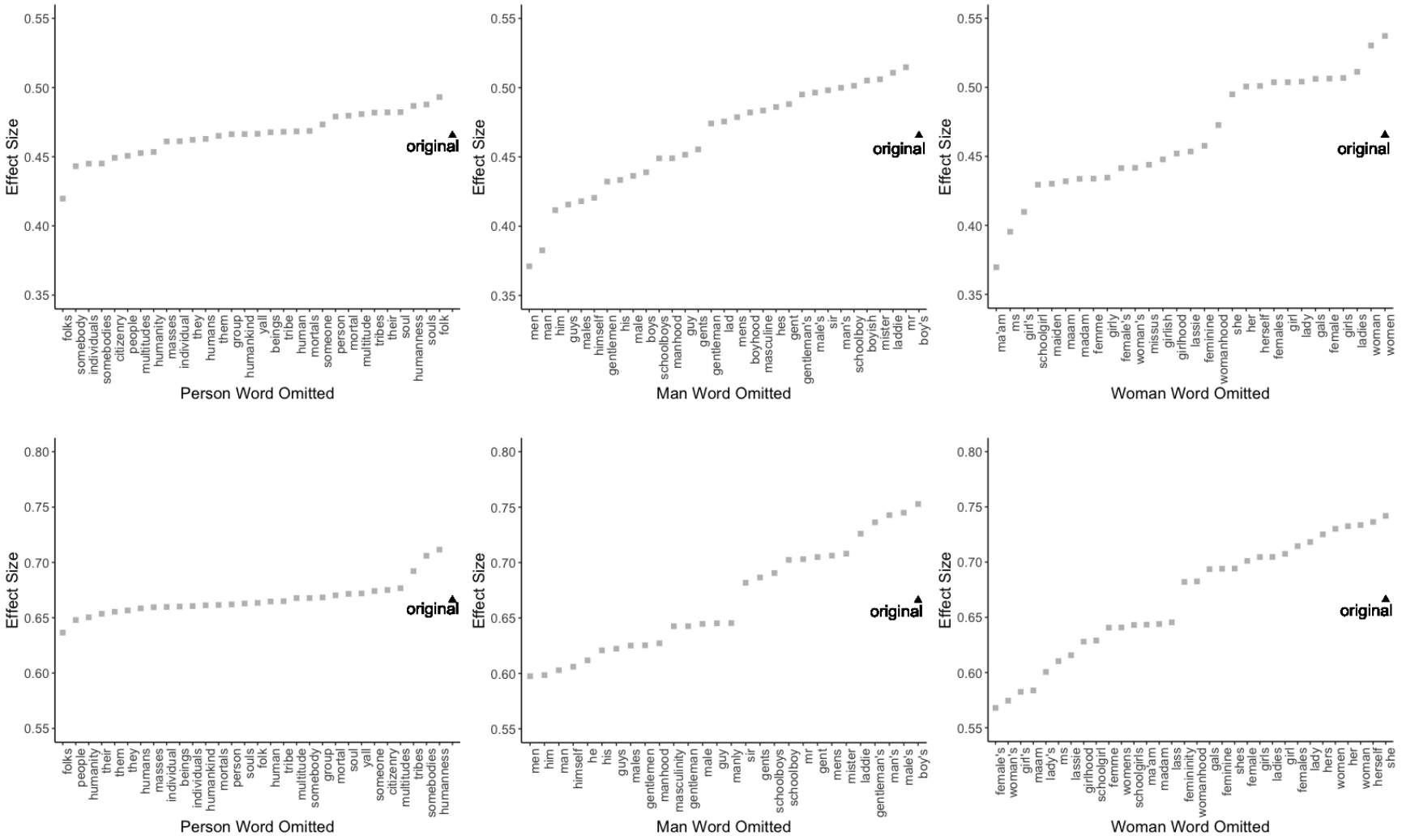
C. “Leave One Out” Analyses (All Studies)

In addition to specifically considering masculine generic words, it was important to rule out the possibility that the results of the present studies were overly reliant on any particular word. To do so, we conducted so-called “leave one out” analyses.

For these analyses, we focused on the difference in similarity between words for MEN vs. words for WOMEN and words for PEOPLE (Study 1), trait words (Studies 2A and 2B), and verbs (Study 3). (That is, we did not examine interactions with gender stereotypicality from Studies 2A, 2B, and 3.) For example, in Study 1 we re-computed the model described above 104 times, each time leaving out a single word for PEOPLE, a single word for WOMEN, or a single word for MEN. The resulting effect sizes for the difference in similarity between words for MEN vs. words for WOMEN with words for PEOPLE for each of these iterations are presented in Fig. S1. For analogous effect sizes for Studies 2A, 2B, and 3, see Figs. S2, S3, and S4, respectively. Visual inspection of these plots suggests that leaving out certain words sometimes resulted in smaller or larger effect sizes, but the effect sizes were generally quite consistent.

Fig. S1

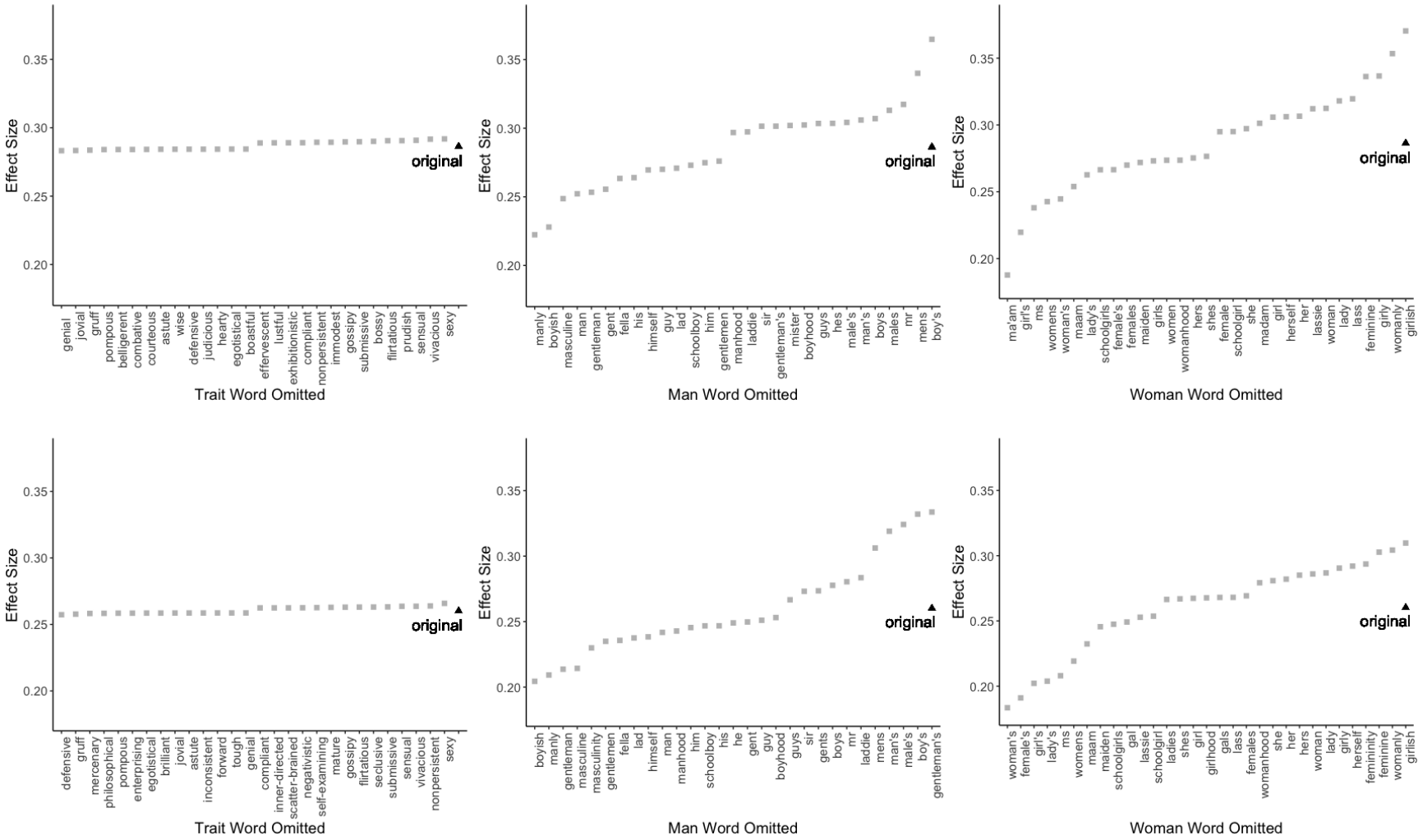
The Difference Between Gender Words When Each Person Word and Each Gender Word is Omitted in Study 1 (Top) and its Replication (Bottom)



Note. "Original" refers to the magnitude of the effect size in the original model when all words were included. For readability, only gender words with the most extreme influence on the original effect size in either direction are depicted.

Fig. S2

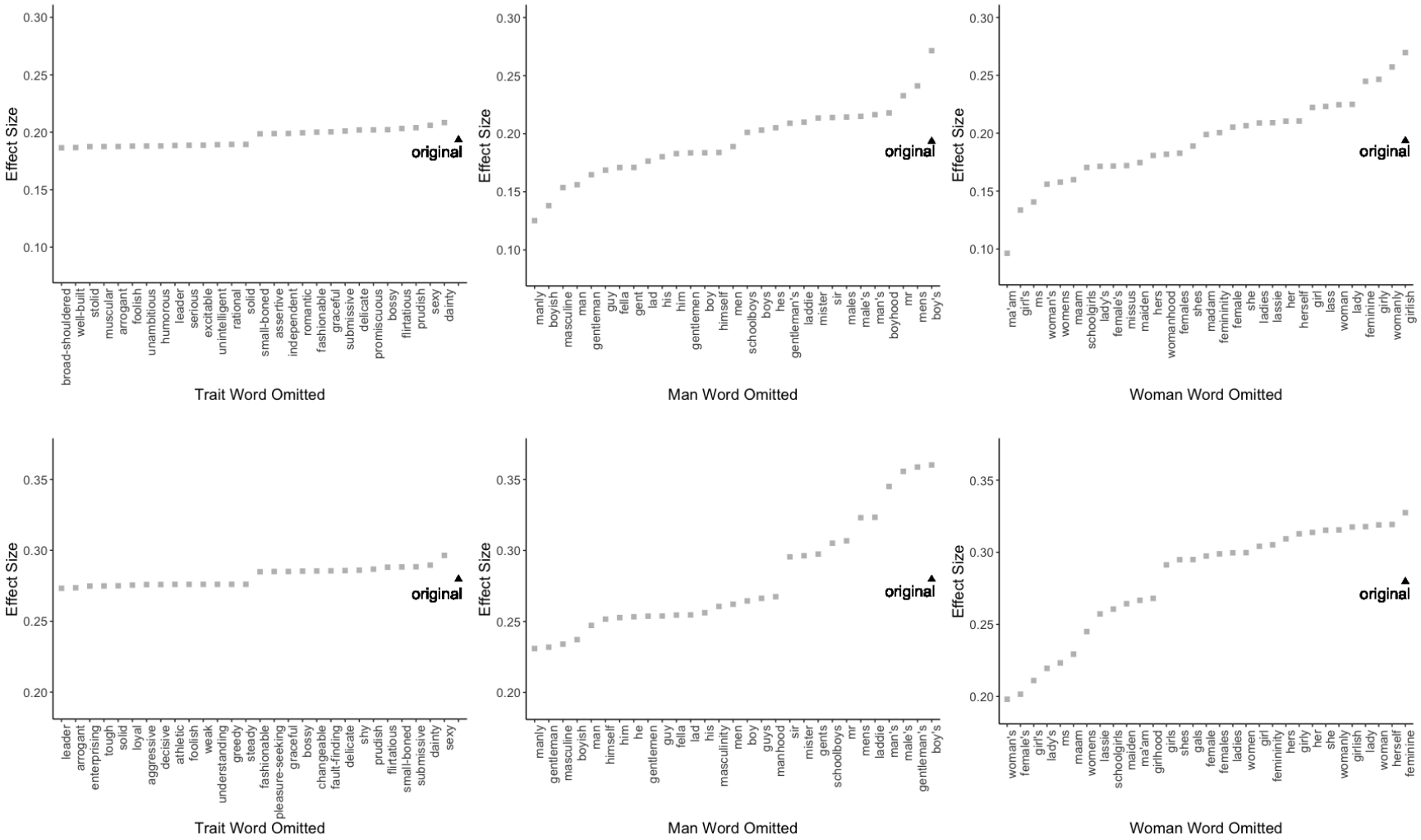
The Difference Between Gender Words When Each Trait and Each Gender Word is Omitted in Study 2A (Top) and its Replication (Bottom)



Note. "Original" refers to the magnitude of the effect size in the original model when all words were included. For readability, only gender words with the most extreme influence on the original effect size in either direction are depicted.

Fig. S3

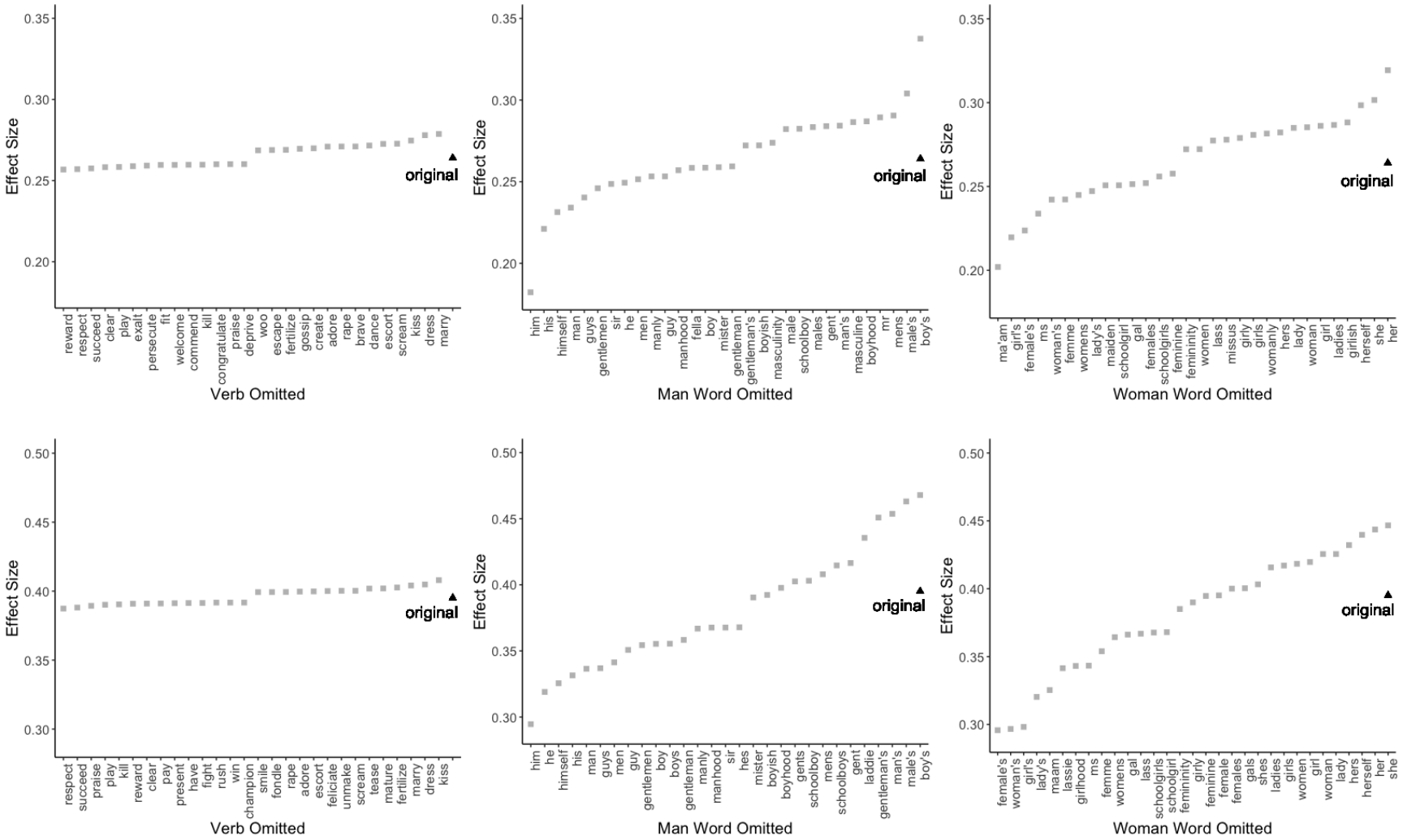
The Difference Between Gender Words When Each Trait and Each Gender Word is Omitted in Study 2B (Top) and its Replication (Bottom)



Note. "Original" refers to the magnitude of the effect size in the original model when all words were included. For readability, only gender words with the most extreme influence on the original effect size in either direction are depicted.

Fig. S4

The Difference Between Gender Words When Each Verb and Each Gender Word is Omitted in Study 3 (Top) and its Replication (Bottom)



Note. "Original" refers to the magnitude of the effect size in the original model when all words were included. For readability, only gender words with the most extreme influence on the original effect size in either direction are depicted.

D. Random Permutation Tests (All Studies)

To again ensure that the results were not overly reliant on particular gender words, we also conducted random permutation tests of the key result. Permutations tests are nonparametric, and do not rely on any particular assumptions about the distribution of the data. For these analyses, we again focused on the difference in similarity between words for MEN vs. words for WOMEN and words for PEOPLE (Study 1), trait words (Studies 2A and 2B), and verbs (Study 3). (That is, we did not examine interactions with gender stereotypicality from Studies 2A, 2B, and 3.)

Taking Study 1 as an example, the permutation test involved recomputing the multilevel model described above 10,000 times, each time randomly shuffling the gender to which each word was assigned (e.g., "he" was randomly designated as a word for WOMEN or as a word for MEN). This procedure was repeated for Studies 2A, 2B, and 3.

This process created data-driven estimates of the null distributions of effect sizes and facilitated a comparison between the null distributions and the observed effects. If any particular gender word or subset of gender words was responsible for the observed effects, then the effect sizes resulting from some of the permutations would be similar to our observed effect sizes. Instead, we consistently found that our observed effect sizes were noticeably larger than the null distributions of effect sizes. Thus, these random permutation tests provide converging evidence that words for PEOPLE (Study 1), trait words (Studies 2A and 2B), and verbs (Study 3) were all more similar to words for MEN than to words for WOMEN (all p's < .001 for both the main studies and replication studies).

E. Frequency Analysis of the Gender Words (All Studies)

We tested potential differences in the frequency of the words for WOMEN and the words for MEN in the training corpus (Common Crawl) used by both fastText (Studies 1-3) and GloVe (replications of Studies 1-3). Although we took care to create lists of words for WOMEN and words for MEN that were parallel in terms of their meaning and syntax, these two sets of gender words may nevertheless differ in terms of frequency. Word embeddings are somewhat sensitive to frequency (*57*), and thus it was important to consider this possibility.

To measure frequency, we used information from fastText, which provides the frequency rank of each word in the Common Crawl corpus. (GloVe does not supply frequency information, but both of these algorithms use the same corpus, so the frequency ranks should be extremely similar.) The most frequent word in the Common Crawl is ranked as 1, the next most frequent word as 2, and so on. Although this frequency information is encoded as ranks (rather than exact frequencies), this metric is relatively precise because of the massive scale of the corpus (i.e., over 630 billion word tokens). This rank data also has the benefit of being based on the same information that the word embeddings themselves were based on.

To test for potential frequency differences between our two sets of gender words, we computed a Mann-Whitney U test, which is appropriate for rank data, but found no evidence for a difference between the frequency ranks of words for MEN (M = 426,964.20, SD = 1,137,915.00, $Median$ = 19,873.00) and words for WOMEN (M = 460,639.70, SD = 1,109,425.00, $Median$ = 26,369.50), U = 760, p = .416, d = –0.03.

F. Word-Embedding Association Tests (Studies 2 and 3 and Replications)

Prior investigations of gender-stereotypical associations in word embeddings conducted a word-embedding association test (WEAT; *21*). This test was designed to be conceptually analogous to a common measure of human stereotypes from the psychology literature: the implicit association test (IAT; *41*). Because both the WEAT and the IAT rely on a double difference score (see details below), they obscure the asymmetry in gender-stereotypical associations we predicted and found in the present research. To compare the present data to previous investigations of gender-stereotypical associations in word embeddings, we conducted a WEAT of gender-stereotypical associations with traits and verbs in Studies 2A, 2B, and 3. We expected to conceptually replicate previous work and find evidence for gender-stereotypical associations in word embeddings.

In Studies 2A and 2B, the WEAT involves first calculating the mean similarity of each trait word to each of the words for WOMEN and, separately, each of the words for MEN and then averaging within gender set. Next, a difference score is calculated between the average similarity of each trait word with words for MEN and words for WOMEN. For traits stereotypical of women, higher difference scores would indicate *less* bias in line with gender-stereotypical associations (i.e., traits stereotypical of women are more similar to words for MEN than to words for WOMEN). For traits stereotypical of men, though, higher

difference scores would indicate *more* bias in line with gender-stereotypical associations (i.e., traits stereotypical of men are more similar to words for MEN than to words for WOMEN). The next step is to sum these difference scores for the traits stereotypical of men and, separately, for the traits stereotypical of women. The final step is then to compute a difference score of these sums. The resulting single number quantifies the extent to which the similarities between trait words and gender words are more in line with gender-stereotypical associations than not. A *p* value can then be obtained by conducting a two-tailed random permutation test.

Formally in the present case, let X represent our set of traits stereotypical of men and Y represent our set of traits stereotypical of women (called *target words* by the original authors; *23*). Let M and W represent our set of words for MEN and words for WOMEN, respectively (called *attribute words* by the original authors; *23*). Let $\cos(\vec{t}, \vec{w})$ represent the cosine of the angle between the word embedding of a given trait word and, in this case, the embedding of a given word for WOMEN. The WEAT test statistic is,

$$s(X, Y, M, W) = \sum_{x \in X} s(x, M, W) - \sum_{y \in Y} s(y, M, W)$$

where for each stereotypical trait (*t*),

$$s(t, M, W) = \text{mean}_{m \in M} \cos\left(\vec{t}, \vec{m}\right) - \text{mean}_{w \in W} \cos\left(\vec{t}, \vec{w}\right)$$

and the effect size (*d*) is,

$$\frac{\text{mean}_{x \in X} s\,(x, M, W) - \text{mean}_{y \in Y} s(y, M, W)}{\text{std\_dev}_{t \in X \cup Y} s(t, M, W)}$$

Applying this test to our data in Studies 2A and 2B, we found greater relative associations between words for MEN and traits stereotypical of men and words for WOMEN and traits stereotypical of women than the inverse (Table S8). We also applied this test to our data in Study 3 and to the replications of Studies 2A, 2B, and 3 and found similar results. Thus, our data are consistent with previous investigations of gender-stereotypical associations in word embeddings.

For instance, reference *21* found that women are associated with the arts and men are associated with the sciences compared to the inverse set of associations (d = 1.24). Similarly, we found that women were associated with certain female-stereotypical traits and verbs (e.g., “compassionate”) and men were associated with certain male-stereotypical traits and verbs (e.g., “brave”) more than the inverse (d range: 0.64-0.89). Crucially, our analyses in the main text show that *gender-stereotypical associations were driven by associations about women, not men*. Because the WEAT relies on two difference scores, it obscures the asymmetry that we predicted and found.

Table S8

WEAT Statistics in Studies 2 and 3 and Replications

Study	Target words	Attribute words	WEAT	*d*
Study 2A	Traits stereotypical of men vs. traits stereotypical of women	Words for MEN vs. words for WOMEN	1.30***	0.67
Study 2A Replication			1.81***	0.89
Study 2B			1.41***	0.57
Study 2B Replication			2.03***	0.75
Study 3	Male-biased verbs vs. female-biased verbs		1.68***	0.64
Study 3 Replication			2.14***	0.73

***$p < .001$

G. Replication of Study 1 in the Biomedical Domain

We conducted another replication of Study 1 using identical lists of words and other procedures to Study 1, with one exception: We used a different set of word embeddings (*53*). The goal of this replication was to test the generalizability of the critical PEOPLE = MEN finding from Study 1 in the biomedical domain.

Similar to the word embeddings analyzed in the main text, these biomedical word embeddings were extracted with the fastText algorithm with 200 dimensions (*6*). But rather than being trained on an undifferentiated corpus of 630+ billion words on the internet (i.e., the Common Crawl corpus), the biomedical embeddings were trained on a smaller corpus of biomedical text: specifically, 4+ billion words from abstracts and titles in the PubMed biomedical and life science research archive and 500+ million words in the MIMIC-III Clinical Database of de-identified hospital clinical notes (vital sign measurements, laboratory test results, procedures, medications, etc.; *87*). The biomedical domain was of particular interest because biomedical research and clinical practice have direct implications for gender (in)equity in health, and it would thus be particularly troubling to find a PEOPLE = MEN bias in this domain.

We compared words for PEOPLE to words for MEN and to words for WOMEN using the same mixed-effects linear regression described in Study 1. Replicating Study 1, we found that words for PEOPLE were more similar in their use to words for MEN (M = 0.08, SD = 0.06) than to words for WOMEN (M = 0.05, SD = 0.04), B = 0.03, SE < 0.01, p < .001, d = 0.49. The effect size in this biomedical domain (i.e., d = 0.49) is similar to that in Study 1 reported in the main text (i.e., d = 0.47), demonstrating the generalizability of the present finding to this different domain based on a different corpus.