Intersectional Race-Gender Stereotypes in Natural Language

Authors: Han-Wu-Shuang Bao[1,2], Peter Gries[2*]

Affiliations:

[1] School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[2] Manchester China Institute, University of Manchester, Manchester, UK

[*] Correspondence to: Peter Gries (peter.gries@manchester.ac.uk)

Citation: Bao, H.-W.-S., & Gries, P. (in press). Intersectional race-gender stereotypes in natural language. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12748

Acknowledgments: The authors thank Cong Peng, Tao Wang, Paton Pak Chun Yam, and Carwyn Morris for their feedback and comments.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability: All data, materials, and analysis code are available at the Open Science Framework (OSF) https://osf.io/rvy8n/.

(Title page updated April 12, 2024 to correct the automatic identification by Google Scholar)

Intersectional Race-Gender Stereotypes in Natural Language

Han-Wu-Shuang Bao, Peter Gries

Abstract

How are Asian and Black men and women stereotyped? Research from the gendered race and stereotype content perspectives has produced mixed empirical findings. Using BERT models pre-trained on English language books, news articles, Wikipedia, Reddit, and Twitter, with a new method for measuring propositions in natural language (the Fill-Mask Association Test, FMAT), we explored the gender (masculinity-femininity), physical strength, warmth, and competence contents of stereotypes about Asian and Black men and women. We find that Asian men (but not women) are stereotyped as less masculine and less moral/trustworthy than Black men. Compared to Black men and Black women, respectively, both Asian men and Asian women are stereotyped as less muscular/athletic and less assertive/dominant, but more sociable/friendly and more capable/intelligent. These findings suggest that Asian and Black stereotypes in natural language have multifaceted contents and gender nuances, requiring a balanced view integrating the gender schema and the stereotype content model. Exploring their semantic representations as propositions in large language models, this research reveals how intersectional race-gender stereotypes are naturally expressed in real life.

Keywords: stereotypes, intersectionality, race, gender, natural language processing, large language models

Statement of Contribution:

What is already known on this subject?

- Asian and Black racial stereotypes intersect with gender stereotypes.
- Asian stereotypes contain both feminine and masculine content.

What does this study add?

- Natural language models can reveal the complexities of intersectional race-gender stereotypes in natural language.
- The gender schema and the stereotype content model can jointly explain intersectional stereotypes of Asian and Black men and women.

Introduction

In 2012, Trayvon Martin, a 17-year-old Black teenager, was shot by a neighbor in Florida. In Summer 2020, tens of millions of Americans demonstrated in #BlackLivesMatter protests following the police killing of another Black man, George Floyd, in Minneapolis. In May 2021, six Asian women were killed in a mass shooting in Atlanta. The #StopAsianHate movement then arose to combat the rise of anti-Asian COVID-related prejudice—much of it involving the racial harassment of Asian men.

Might racial stereotypes underly this pattern of racial aggression? Social psychologists have long studied how racial stereotypes shape dominant group attitudes and behaviors towards Asian and Black people. Two predominant theoretical perspectives have produced mixed empirical findings, however. First, a gendered race view suggests that Asian and Black stereotypes intersect with gender stereotypes. This "race is gendered" line of scholarship has demonstrated that Asian people are stereotyped as relatively more feminine, while Black people are stereotyped as relatively more masculine (Galinsky et al., 2013; Johnson et al., 2012). Second, scholarship using the Stereotype Content Model (SCM) as the theoretical framework finds that Asian people are stereotyped as lower on warmth but higher on competence, while Black people are stereotyped as both less warm and less competent (Fiske et al., 2002, 2007; see also Berdahl & Min, 2012; Lin et al., 2005).

The *Dual Perspective Model* (Abele & Wojciszke, 2007, 2014) can help clarify the contradictions between these two sets of findings: (1) the *communal* content involves the "feminine" traits attributed to the female gender role (e.g., warm, sympathetic; Bem, 1974) and the "warm" traits associated with stereotyped groups (e.g., friendly, trustworthy; Fiske, 2018); (2) the *agentic* content, by contrast, involves the "masculine" traits attributed to the male gender role (e.g., competitive, dominant; Bem, 1974) and the "competent" traits associated with stereotyped groups (e.g., capable, assertive; Fiske, 2018). There are overlaps in traits describing gender roles and stereotype contents, which should be disentangled. It is thus ambiguous to use only "feminine" and "masculine" to oversimplify Asian and Black racial stereotypes as gendered.

We seek to disentangle this complexity by inductively exploring the contents of Asian and Black racial stereotypes and their possible gender nuances (i.e., intersectionality between race and gender). To test how people naturally express such *intersectional* race-gender stereotypes of Asian and Black men and women in the real world, we employ Natural Language Processing (NLP) with Large Language Models (LLMs) to detect stereotypes from massive text corpora. By using language models to summarize people's natural expressions (without explicitly asking about their stereotypes), we can obtain more ecologically valid evidence and a more general perspective on human stereotyping processes (Argyle et al., 2023; Dillion et al., 2023; Grossmann et al., 2023).

The Gendered Race View of Asian and Black Stereotypes

Social psychologists have long investigated the nature and consequences of racial and gender stereotypes. Recent work has demonstrated an "Asian = feminine, Black = masculine" overlap between these two types of stereotypes, suggesting that race is gendered. This effect is robust, with cognitive basis, developmental changes, real-world implications, and subgroup nuances. Specifically, Asian female (vs. male) and Black male (vs. female) faces were more accurately and quickly categorized, remembered, and recognized (Goff et al., 2008; Johnson et al., 2012; Schug et al., 2015; Sesko & Biernat, 2010). This cognitive pattern exists not only in adults but also in children, who categorized prototypically gendered race exemplars more efficiently as they aged, with an onset at around five years old (Lei et al., 2020, 2022).

Gendered racial stereotypes have real-world consequences. Since Black men are perceived as more masculine than Asian men, they are preferred by White women in heterosexual romantic relationships, are more likely to be selected for masculine leadership positions, and are better represented in masculine sports (Galinsky et al., 2013; Hall et al., 2015). White men, by contrast, prefer (feminine) Asian (vs. Black) women in heterosexual relationships (Galinsky et al., 2013). Asian men and Black women, as less prototypical groups, were underrepresented in popular magazines (Schug et al., 2017), but Black women coaches were better represented in men's athletic teams (Cunningham et al., 2021). Furthermore, in same-sex sexual behavior, gay men tended to stereotype Asian men as feminine "bottoms" and Black men as masculine "tops" (Lick & Johnson, 2015).

There are also nuances in gendered racial stereotypes concerning the perceiver and the target. One recent study found a *perceiver difference* that Asian men and Black women (i.e., counter-stereotypical groups) held weaker race–gender associations, while Chinese (vs. American) participants held stronger race–gender associations (Axt et al., 2023). Another recent study identified a *target difference* that East Asian men (but not women) were viewed as less masculine and more feminine, compared to their South Asian, White, and Black counterparts (Goh & Trofimchuk, 2022).

A closer look reveals that gendered racial stereotypes overlap with *physical* attributes. In particular, Black (vs. Asian) people are stereotyped as more physically strong (muscular) and athletic. Basically, gender stereotypes often incorporate noticeable physical and athletic attributes, as reflected in both gender roles (Bem, 1974) and stereotype contents (Alexander et al., 2005; Rosette et al., 2016). Previous research on gendered racial stereotypes has indeed predominantly focused on the *physical*, rather than *behavioral*, attributes of masculinity and femininity. Most studies used an important feature of physical appearance—faces—as stimuli to examine the gendered race effect (Axt et al., 2023; Goff et al., 2008; Goh & Trofimchuk, 2022; Johnson et al., 2012; Lei et al., 2020, 2022; Schug et al., 2015; Sesko & Biernat, 2010). Even an experiment without facial stimuli relied upon words that reflect physically, rather

than behaviorally, gendered attributes: *strong*, *muscular*, *vigorous*, and *burly* for masculinity; and *delicate*, *gentle*, *graceful*, and *beautiful* for femininity (Galinsky et al., 2013, Study 2). Furthermore, physical size (Wilson et al., 2017) and physical masculinity-femininity (Wilkins et al., 2011) play crucial roles in racial stereotypes. Physicality (physical strength), therefore, deserves closer examination as an underlying attribute in Asian and Black stereotypes.

The gendered race effect is consistent with a theoretical perspective that posits gender (masculinity and femininity) as the primary schema through which individuals perceive their social environment, accounting for many "Big Two" dimensions across domains (Martin & Slepian, 2021). However, recent work suggests that gender is not necessarily the source of all other social categories: gender is relative and context-dependent (Martin, 2023); gender roles are culturally constructed (Diekman & Schmader, in press); and gender stereotypes change over time (Bhatia & Bhatia, 2021; Charlesworth et al., 2022; Garg et al., 2018; Zhu & Chang, 2019). Thus, a static gendered view of Asian and Black stereotypes, even with physicality (physical strength) as its core underlying facet, appears insufficient.

The Stereotype Content View of Asian and Black Stereotypes

Asian and Black stereotypes may be more complex than only gender and physicality based, and may involve multiple attributes that warrant attention. Such a multifaceted view of stereotypes can be grounded in the Stereotype Content Model, which posits *Warmth* and *Competence* as two fundamental dimensions of stereotype content (Fiske et al., 2002, 2007). The latest theoretical framework elaborates the SCM by subdividing Warmth into the Sociability/Friendliness and Morality/Trustworthiness facets, and Competence into the Capability/Intelligence and Assertiveness/Dominance facets (Fiske, 2018; Nicolas et al., 2021). The SCM can be used as a structural framework to understand racial stereotypes.

Early studies found that Asian people were stereotyped as lower on warmth but higher on competence, while Black people were stereotyped as lower on both warmth and competence (e.g., Fiske et al., 2002, 2007). More recent inductive and data-driven approaches have revealed more nuanced contents of Asian and Black stereotypes (Charlesworth et al., 2022; Ghavami & Peplau, 2013; Lick & Johnson, 2015). One qualitative study that collected freeform responses found that Asian people were most commonly stereotyped as *intelligent*, *quiet*, *shy*, *studious*, *nerdy*, *skinny*, *good at math*, and *lacking in social skills*, whereas Black people were most commonly stereotyped as *athletic*, *loud*, *confident*, *assertive*, *violent*, *unintelligent*, *good at basketball*, and *having an attitude* (Ghavami & Peplau, 2013). This study also found that both *intelligent* and *physically strong* were among the most commonly mentioned stereotypes of men (Ghavami & Peplau, 2013). Another study replicated these findings, with the traits *intelligent* and *athletic* being the most frequently selected to describe Asian and Black men, respectively (Lick & Johnson, 2015). Word embeddings, a basic NLP technique that represents word meanings as numeric vectors, has also been used to uncover

more intricate Asian and Black stereotype contents and their changes over time (Charlesworth et al., 2022).

Overall, these studies provide a multifaceted view of Asian and Black stereotypes, and suggest that racial stereotypes do *not* fully overlap with the gender content. A more balanced perspective is needed to integrate the gendered race and stereotype content views to better understand the intersectional stereotypes of Asian and Black men and women.

The Present Study

Balancing the gender schema (Martin & Slepian, 2021) and the SCM (Fiske, 2018; Fiske et al., 2002, 2007), our study explores Asian and Black stereotype contents across six facets. Based on the extant literature, we hypothesized that in English speakers' natural expressions in the real world, Asian (vs. Black) people are stereotyped as (1) less masculine, (2) less physically strong (muscular), (3) less assertive/dominant, but (4) more capable/intelligent. For the two facets of warmth, however, there are diverging predictions. Through the lens of gender (i.e., "female = warm"), Asian (vs. Black) people might be stereotyped as more sociable/friendly and more moral/trustworthy. By contrast, previous qualitative studies suggested that Asian (vs. Black) people might be viewed as less sociable but equally moral. Thus, we refrained from making specific hypotheses about these two facets, instead seeking insights inductively from the data.

More importantly, beyond the overall contrast between Asian and Black people, we also sought to explore, without specific hypotheses, the gender nuances in Asian and Black stereotypes. Specifically, we explored stereotypes of the four intersectional race-gender groups "Asian men", "Asian women", "Black men", and "Black women"; see Lei et al., 2023 for a sociohistorical model of intersectional social category prototypes). In doing so, we tested (1) race-focused contrasts (Asian vs. Black) separately for men and women and (2) gender-focused contrasts (men vs. women) separately for Asian and Black groups. These pairs of contrasts would convergently delineate the intersectional race-gender stereotypes.

To examine stereotypes of Asian and Black men and women in natural language, we employed a new method: the *Fill-Mask Association Test* (FMAT; Bao, in press). The FMAT uses BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2018) to compute *semantic probabilities* of option words filling in the masked blank of a designed query (i.e., a cloze-like contextualized sentence), measuring *propositional* semantic representations in natural language. The initial validation work, consisting of 15 studies, has established the reliability and validity of the FMAT in predicting factual associations (e.g., the name-gender association), measuring attitudes and biases (e.g., the European–African race bias), capturing social stereotypes (e.g., the gender-career stereotype), and retrospectively delineating lay perceptions of sociocultural changes over time (e.g., the change in cultural looseness–tightness). The FMAT replicated seminal findings previously

obtained with human participants (e.g., the Implicit Association Test) and other big-data text-analytic methods (e.g., the Word Embedding Association Test), demonstrating good internal consistency reliability (among queries) and inter-rater agreement (among BERT models) as well as good criterion, convergent, discriminant, and incremental validity (Bao, in press).

By using BERT as contextualized language models, the FMAT has several advantages over traditional methods. Compared to interviews and surveys, the FMAT allows for testing how the people who produced the English language training texts naturally expressed their stereotypes and biases. Compared to other NLP methods such as word embeddings, the FMAT with contextualized BERT models can provide more nuanced insights into social cognition by using phrases and sentences (rather than only single words) to measure propositions in natural language. While word embeddings can test semantic associations between words, static word analysis cannot disambiguate word meanings between contexts (e.g., the different meanings of "warm" in "she is *warm*," "she is not a *warm* person," and "she lives in a *warm* city"). By using phrases that specify relational and contextual information, the FMAT enables the natural language analysis of stereotypes about intersectional social groups (e.g., Asian and Black men and women). The FMAT is thus particularly appropriate and useful for the current research inductively exploring multifaceted stereotypes of Asian and Black men and women in natural language.

Method

We used the FMAT to examine intersectional Asian and Black stereotypes in natural language, exploring possible gender nuances within different facets of these stereotypes.

Sample of Language Models

The FMAT requires the use of BERT models, a family of large pre-trained language models that can fill in the masked blank in a designed query (Devlin et al., 2018). Hugging Face provides a collection of variants of BERT models (see https://huggingface.co/models). We sampled the 12 most representative BERT models spanning from the original "bert-base-uncased" model to more advanced ones such as "roberta-base" (see Table 1 for details of these models). In doing so, we aimed to obtain more robust and generalizable findings than relying solely upon a single model.

As shown in Table 1, the 12 BERT models have been pre-trained on large English language text corpora including Wikipedia, BookCorpus (11,038 unpublished books), CommonCrawl (63 million English news articles), OpenWebText (8 million documents from Reddit), and Twitter (850 million English language Tweets, from 2012 to 2020). As with other text-analytic methods, the gender, race, nationality, and other demographic characteristics of the content producers is unknown. However, it is reasonable to assume that the producers were largely English speakers from the West, amongst whom White people were likely the largest proportion (Dillion et al., 2023). Thus, we assume that this sample

provides general estimates of the Asian and Black stereotypes held by many White people, but we caution readers about the unknown demographic characteristics of the text producers.

Table 1. Summary of the 12 BERT Language Models Sampled in the Present Study

Model name	Case-sensitive	Vocabulary	Dimensions	Layers	Pre-training corpora
bert-base-uncased	No	30,522	768	12	wiki, book
bert-base-cased	Yes	28,996	768	12	wiki, book
bert-large-uncased	No	30,522	1,024	24	wiki, book
bert-large-cased	Yes	28,996	1,024	24	wiki, book
distilbert-base-uncased	No	30,522	768	6	wiki, book
distilbert-base-cased	Yes	28,996	768	6	wiki, book
albert-base-v1	No	30,000	768	12	wiki, book
albert-base-v2	No	30,000	768	12	wiki, book
roberta-base	Yes	50,265	768	12	wiki, book, cc, open
distilroberta-base	Yes	50,265	768	6	open
vinai/bertweet-base	Yes	64,001	768	12	twitter
vinai/bertweet-large	Yes	50,265	1,024	24	twitter

Note. wiki = Wikipedia; book = BookCorpus (11,038 unpublished books scraped from the Internet); cc = CommonCrawl (63 million English news articles); open = OpenWebText (8 million documents from Reddit); twitter = Tweets (850 million English Tweets, from 2012 to 2020).

Query Design

In the FMAT, a query (i.e., a sentence with a masked word) is input for BERT models to understand the linguistic context and then estimate how likely a certain target word could replace the mask. In this study, we used a direct, simple, and unambiguous query design consisting of the following three versions of wording to increase the robustness of results:

1. "[MASK] people/men/women {ATTRIB}." [*Asian* vs. *Black*]
2. "The [MASK] people/men/women {ATTRIB}." [*Asian* vs. *Black*]
3. "Most [MASK] people/men/women {ATTRIB}." [*Asian* vs. *Black*]

In practical implementation, each query template contained a subgroup word (i.e., *people*, *men*, or *women*), and the {ATTRIB} was replaced by one of the phrases describing a facet of the stereotype content (e.g., "Most [MASK] people are feminine."). The [MASK] token was left blank for BERT models to estimate the semantic probability of the two racial group words (i.e., *Asian* vs. *Black*). Note that the words used to fill in the [MASK] blank must be a single word in a BERT model's vocabulary. Thus, infrequent words that are out of the model vocabulary could not be the target words for [MASK]. In our study, "Asian" and "Black" ("asian" and "black" for case-insensitive models) were in each model's vocabulary.

Facets of Stereotype Contents

The {ATTRIB} label in the query, as mentioned above, was substituted (before the fill-mask task) by a phrase that can precisely describe the stereotype content. We examined

six facets of stereotype content: (1) gender (masculine vs. feminine), (2) physical strength (muscular/athletic), (3) warmth: sociable/friendly, (4) warmth: moral/trustworthy, (5) competence: capable/intelligent, and (6) competence: assertive/dominant.

For each of these six facets, we designed 12 sets of phrase pairs suitable for pairwise contrasts. Table 2 presents the complete list of words and phrases for readers to evaluate their appropriateness. To generate these phrases, we relied upon both the definition of each facet and an established dictionary of stereotype content (Nicolas et al., 2021). To disentangle masculinity and femininity from the other facets, we restricted the gendered facet to phrases that are exclusively related to gender, such as "are masculine – are feminine" and "are like men – are like women" (Table 2). For the physical strength facet, we designed phrases that specifically describe *physical* strength, such as "are physically strong – are not physically strong" and "have a tough body – have a fragile body" (Table 2).

For the four facets of warmth and competence, we selected the most representative words from Nicolas et al.'s (2021) dictionary, ensuring that each word was paired with an antonym (e.g., "warm – cold") and that both words were included only in the corresponding facet and not other facets. We did this for four reasons. First, using the most representative words can ensure high content validity by excluding less relevant words. Second, ensuring the unique classification of the words with no overlap with the other facets minimizes the confounding of different attributes. Third, all the chosen words had appropriate antonyms, while other words from the dictionary did not necessarily have frequently used antonyms. Finally, by having an equal number of phrases (12 pairs) for each facet, we sought to maintain proportionality across facets, both avoiding biases due to overrepresented facets with disproportionately more items, and making the coefficient estimates more comparable across facets.

Analytic Strategy

We used the R package "FMAT" (Bao, 2023) to complete the fill-mask workflow for each query sentence. The estimated raw probability of a target word is not its actual frequency, but the likelihood of this word appearing in the query context based on a BERT language model's *understanding* of it. That said, its value is likely to be influenced by word frequencies in the pre-training text corpora. Therefore, it is necessary to have a reference point to measure the *relative* association. To this end, we computed the *log probability ratio* (LPR) of a word *w* between the two phrases (*A* vs. *B*) in each pair of attributes:

$$\text{LPR}(w) = \log\frac{P(w|attrib_A)}{P(w|attrib_B)} = \log P(w|attrib_A) - \log P(w|attrib_B)$$

This LPR index is normally distributed, less affected by word frequencies, and thus more appropriate than raw probabilities for linear modeling. This computation produced N = 15,552 observations of LPR, consisting of 2 [MASK] target words × 6 facets of stereotype

attributes × 12 pairs of attribute phrases × 3 query templates × 3 gender subgroups × 12 BERT models. A sensitivity power analysis (two-tailed $\alpha = .05$) showed that this sample size could achieve 80% power to detect the minimum effect size Cohen's $d = 0.11$ for each facet.

To account for the nested structure of the data, we used the R package "lmerTest" (Kuznetsova et al., 2017) to fit a linear mixed model (LMM). It specified the cross-classified random intercepts of the 12 language models and the 72 pairs of attribute phrases, with LPR (divided by its standard deviation [*SD*]) as the outcome variable while all the other variables and their interactions as predictors.[1] The LMM formula in R was:

LPR ~ MASK * facet * subgroup * query + (1 | model) + (1 | attribute)

As the LPR has been scaled to have $SD = 1.00$, the main results of the MASK contrast between *Asian* and *Black* can be directly interpreted as an effect size equivalent to Cohen's *d*. Data were analyzed using R (version 4.3.0; R Core Team, 2023). All data, analysis code, and full results are available at the Open Science Framework (OSF) https://osf.io/rvy8n/.

[1] To check the robustness of results, we also fit an LMM with necessary *random slopes* that allowed the model to both converge (without singular fit) and account for potential variances of key predictors across BERT models and attribute phrases (see Bates et al., 2015; Matuschek et al., 2017 for guidance on how to balance false discovery rate and statistical power in LMMs). The results showed identical point estimates, with more conservative significance tests, of all effects reported in the main text; and the main conclusions remained unchanged (see the "Supplemental Analysis" section in online supplemental materials).

Table 2. Query Design for the FMAT

Facet of attribute	Pairwise contrast (12 pairs of phrases for each facet)	
	High	Low
Gendered (masculinity)	are like a man	are like a woman
	are like men	are like women
	are manful	are womanful
	are manlike	are womanlike
	are manly	are womanly
	are masculine	are feminine
	are typical of a man	are typical of a woman
	are typical of men	are typical of women
	have a masculine personality	have a feminine personality
	have a masculine trait	have a feminine trait
	have masculine characteristics	have feminine characteristics
	have masculine traits	have feminine traits
Physical strength (muscularity)	are athletic	are not athletic
	are athletic	are unathletic
	are more like athletes	are less like athletes
	are muscular	are not muscular
	are physically able	are not physically able
	are physically strong	are not physically strong
	have a burly body	have a slender body
	have a solid body	have a soft body
	have a strong body	have a weak body
	have a tough body	have a fragile body
	have an athletic body	have an unathletic body
	have lots of muscles	have very few muscles
Warmth: sociability/friendliness	are affectionate	are unaffectionate
	are agreeable	are disagreeable
	are amiable	are unamiable
	are amicable	are unamicable
	are friendly	are unfriendly
	are likable	are unlikable
	are outgoing	are shy
	are responsive	are unresponsive
	are sociable	are unsociable
	are social	are unsocial
	are supportive	are unsupportive
	are warm	are cold

Facet of attribute	Pairwise contrast (12 pairs of phrases for each facet)	
	High	Low
Warmth: morality/trustworthiness	are authentic	are inauthentic
	are ethical	are unethical
	are faithful	are unfaithful
	are honest	are dishonest
	are loyal	are disloyal
	are moral	are immoral
	are reliable	are unreliable
	are responsible	are irresponsible
	are sincere	are insincere
	are tolerant	are intolerant
	are trustworthy	are untrustworthy
	are truthful	are untruthful
Competence: capability/intelligence	are able	are unable
	are capable	are incapable
	are clever	are foolish
	are competent	are incompetent
	are competitive	are uncompetitive
	are intelligent	are unintelligent
	are knowledgeable	are unknowledgeable
	are skilled	are unskilled
	are skillful	are unskillful
	are smart	are stupid
	are sophisticated	are unsophisticated
	are wise	are unwise
Competence: assertiveness/dominance	are active	are inactive
	are adventurous	are unadventurous
	are ambitious	are unambitious
	are assertive	are unassertive
	are brave	are afraid
	are certain	are uncertain
	are confident	are unconfident
	are decisive	are indecisive
	are determined	are undetermined
	are dominant	are submissive
	are independent	are dependent
	are proactive	are passive

Note. FMAT = Fill-Mask Association Test. Query 1: "[MASK] people/men/women {ATTRIB}." Query 2: "The [MASK] people/men/women {ATTRIB}." Query 3: "Most [MASK] people/men/women {ATTRIB}." [MASK] = racial group words (*Asian* vs. *Black*). {ATTRIB} = pairwise phrases of stereotype contents.

Results

Reliability Analysis

Before testing the main results, we analyzed the reliability of the FMAT in two ways. First, to assess the inter-rater agreement among the 12 BERT models (treated as "raters") in understanding the queries and estimating the probabilities (log-transformed), we computed the average-score intraclass correlation coefficient ($ICC_{average}$), with both BERT models and individual query sentences considered as random effects. For all query sentences, $ICC_{average}$ = .90; for each of the three query templates by the three gender subgroups (i.e., nine queries), the ICCs were also high, ranging from .70 to .93 (Figure 1).[2]

Second, we assessed the internal consistency of LPRs both among the query templates (k = 9) and among the attribute phrase pairs (k = 12), separately for each facet of stereotypes. As shown in Table 3, the results revealed good internal consistency among the queries (α_{query} = .86~.91) and acceptable internal consistency among the phrase pairs (α_{pair} = .56~.82).

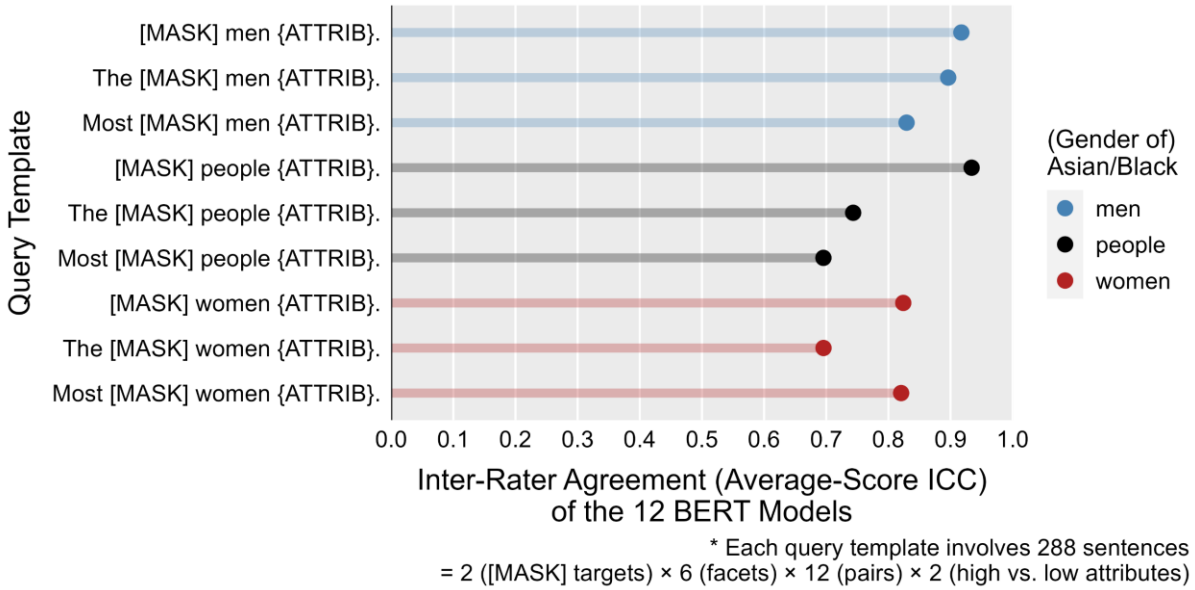

Figure 1. Inter-Rater Agreement (Intraclass Correlation Coefficient, $ICC_{average}$) of the 12 BERT Models

[2] The ICC reliability analysis treated (1) log probabilities of words filling in the mask of a query as "rating scores"; (2) n uniquely filled sentences as "rating items" (rows); and (3) k BERT models as the "raters" (columns). ICCs were estimated using the "icc()" function from the R package "irr" (Gamer et al., 2019). The formula of average-score ICC is shown below, indicating the reliability of average results across all sampled BERT models (McGraw & Wong, 1996; Shrout & Fleiss, 1979). See online supplemental materials for details about the R code.

$$ICC_{average} = \frac{MS_{row} - MS_{error}}{MS_{row} + \frac{1}{n}(MS_{column} - MS_{error})}$$

Table 3. Internal Consistency (Cronbach's α for Queries and Item Pairs) and Effect Sizes (*d*) of the FMAT

Facet (each with 12 pairs of phrases)	α_{query} ($k = 9$)	α_{pair} ($k = 12$)	Asian vs. Black contrast			
			Overall	Subgroup		
				Men	People	Women
Gender (masculinity)	.86	.82	–0.24***	–0.46***	–0.27***	0.00
Physical strength (muscularity)	.90	.75	–0.20***	–0.22***	–0.14*	–0.24***
Warmth: sociability/friendliness	.91	.79	0.44***	0.40***	0.55***	0.37***
Warmth: morality/trustworthiness	.90	.73	–0.09**	–0.27***	0.05	–0.05
Competence: capability/intelligence	.91	.80	0.36***	0.27***	0.42***	0.40***
Competence: assertiveness/dominance	.91	.56	–0.23***	–0.25***	–0.22***	–0.23***

Note. FMAT = Fill-Mask Association Test. Effects were estimated based on a linear mixed model (LMM) where both the 12 BERT models (Table 1) and the 72 pairs of phrases in query (Table 2) were specified as random intercepts. The overall results were marginal mean effects across all subgroups. Subgroup contrasts: (1) Asian men vs. Black men; (2) Asian people vs. Black people; and (3) Asian women vs. Black women. See online supplemental materials for parallel results of an LMM with necessary random slopes included, which showed identical point estimates of the effect sizes as reported in this table, with more conservative significance tests, and did not change the main conclusions.

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

Race-Focused Analysis: Asian vs. Black Stereotypes for Gender Subgroups

All factors and their interactions in the LMM explained $R^2_{marginal} = 7.0\%$ of the total variance. All fixed effects and random effects together accounted for $R^2_{conditional} = 24.7\%$ of the total variance. As expected, the main effect of the MASK contrast between *Asian* and *Black* across all conditions was not significant, $F(1, 15367) = 0.17$, $p = .68$. We only anticipated differences between facets; and as predicted, the MASK contrast interacted with facets, $F(5, 15367) = 82.08$, $p < .001$, $\eta^2_p = .026$. This interaction, furthermore, had subgroup nuances, as indicated by a small three-way interaction between MASK, facets, and subgroups, $F(10, 15367) = 3.24$, $p < .001$, $\eta^2_p = .002$. The MASK × facet interaction did not interact with query templates, $F(10, 15367) = 0.68$, $p = .75$, suggesting that the multifaceted effects were consistent across query designs. Additionally, the most complicated four-way interaction was not significant, $F(20, 15367) = 0.17$, $p > .99$.

Given the three-way interaction between MASK, facets, and subgroups, we examined the simple effect of Asian vs. Black stereotypes separately for the *people*, *men*, and *women* subgroups within each facet. We interpreted the results in the order of the six facets. Table 3 and Figure 2 summarize the resulting effect sizes (*d*), with 95% confidence intervals (CI) reported in the main text. More detailed results are available in the supplemental materials.

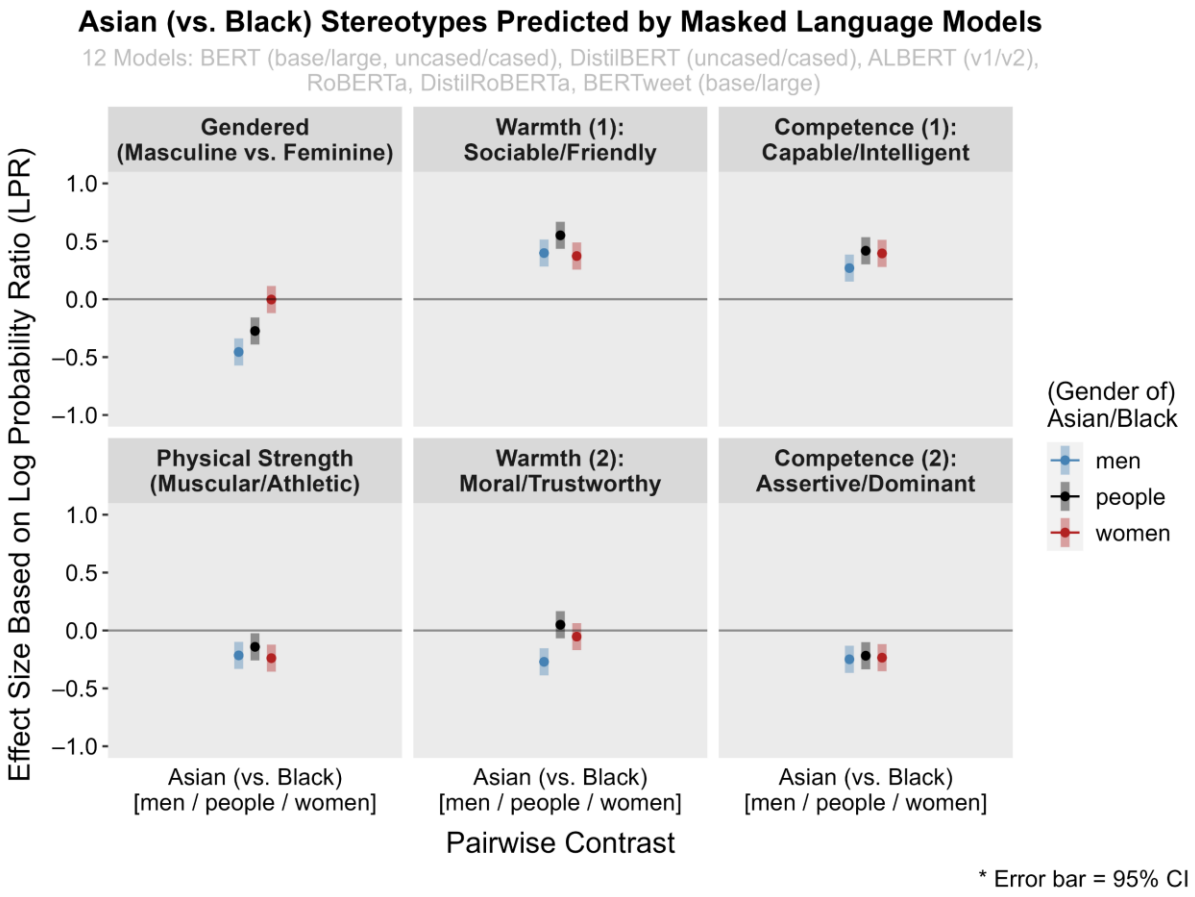

Figure 2. Effect Sizes (*d*) of the FMAT for Gender Subgroups (Men, People, Women) of Asians and Blacks *Note*. Each point represents the standardized difference between Asian and Black (e.g., a red point shows the standardized difference between Asian *women* and Black *women* for the facet).

First, we replicated the gendered race effect (Figure 2, top-left). Compared to Black people and men, respectively, Asian people ($d = –0.27$, $p < .001$, 95% CI [–0.39, –0.16]) and Asian men ($d = –0.46$, $p < .001$, 95% CI [–0.57, –0.34]) were stereotyped as less masculine. However, Asian women and Black women did not differ in gender content ($d = 0.00$, $p = .96$, 95% CI [–0.12, 0.11]). These findings suggest that the gendered race effect may be salient for Asian (vs. Black) men but not women, at least in natural language.

Second, we found that Asian people ($d = –0.14$, $p = .018$, 95% CI [–0.26, –0.02]), Asian men ($d = –0.22$, $p < .001$, 95% CI [–0.33, –0.10]), and Asian women ($d = –0.24$, $p < .001$, 95% CI [–0.36, –0.12]) were all stereotyped as less physically strong (less muscular and less athletic) than their Black counterparts (Figure 2, bottom-left). These findings show both consistency with and nuances from the result pattern identified in the gender attribute. Specifically, Asian (vs. Black) people and men stereotypes were both characterized as more feminine and physically weaker, with the physical-strength effect roughly half in size of the gendered effect. Nonetheless, while Asian women were still stereotyped as physically weaker than Black women, the two groups did not show significant differences in gender content. This pattern suggests a dissociation between the gender facet and the physical-strength facet,

as well as a meaningful nuance between men and women in Asian and Black stereotypes.

Next, we found differential effects for the two facets of warmth. Asian (vs. Black) targets were more associated with stereotypes of being more sociable/friendly (Figure 2, top-middle), which was consistent across "people" ($d = 0.55$, $p < .001$, 95% CI [0.43, 0.67]), "men" ($d = 0.40$, $p < .001$, 95% CI [0.28, 0.52]), and "women" ($d = 0.37$, $p < .001$, 95% CI [0.26, 0.49]). By contrast, Asian (vs. Black) men were perceived as less moral/trustworthy ($d = –0.27$, $p < .001$, 95% CI [–0.39, –0.15]), but this was not significant for Asian women ($d = –0.05$, $p = .37$, 95% CI [–0.17, 0.06]) or for Asian people as a whole ($d = 0.05$, $p = .41$, 95% CI [–0.07, 0.17]) (Figure 2, bottom-middle). Again, these results suggest meaningful gender nuances in Asian and Black stereotypes.

Finally, the results for the two facets of competence also supported our hypotheses. While Asian (vs. Black) targets were stereotyped as more capable/intelligent ("people": $d = 0.42$, $p < .001$, 95% CI [0.30, 0.54]; "men": $d = 0.27$, $p < .001$, 95% CI [0.15, 0.39]; "women": $d = 0.40$, $p < .001$, 95% CI [0.28, 0.51]) (Figure 2, top-right), they were seen as less assertive/dominant ("people": $d = –0.22$, $p < .001$, 95% CI [–0.33, –0.10]; "men": $d = –0.25$, $p < .001$, 95% CI [–0.37, –0.13]; "women": $d = –0.23$, $p < .001$, 95% CI [–0.35, –0.12]) (Figure 2, bottom-right). The consistency of the results between men and women suggests that the competence facet of Asian (vs. Black) stereotypes did not have significant gender subgroup nuances.

Gender-Focused Analysis: Men vs. Women Contrasts for Racial Stereotypes

In addition to the Asian vs. Black contrast within their respective gender subgroups, we also conducted an exploratory analysis of the men vs. women contrast within each race. Such a gender-focused analysis can complement the above race-focused analysis to provide further insights into the intersectional stereotypes of Asian and Black men and women. The gender analysis used the identical LMM but changed the main contrast variable from MASK to subgroup in the simple effect specification. We report the full results in the supplemental materials, summarizing the key gender differences below. All p values were adjusted ($p_{adj.}$) for multiple comparisons with the Bonferroni correction method.

We found no significant gender differences for Asian and Black men and women on the two facets of competence (ps > .10) or on the physical-strength facet (ps > .99). For the gender content, Black men were viewed as more masculine than Black women ($d = 0.14$, $p_{adj.} = .046$, 95% CI [0.002, 0.29]), while Asian men were perceived as even *less* masculine than Asian women ($d = –0.31$, $p_{adj.} < .001$, 95% CI [–0.45, –0.17]), a finding convergent with the lower prototypicality and lower psychological visibility of Asian men (Lei et al., 2023; Schug et al., 2015). For the two facets of warmth, accordingly, we detected meaningful gender differences only for the Asian stereotype in morality/trustworthiness, with Asian men viewed as *less* moral/trustworthy than Asian women ($d = –0.23$, $p_{adj.} < .001$, 95% CI [–0.37, –0.09]).

Discussion

The present study used large language models and the new FMAT method to explore the multifaceted contents of intersectional race-gender stereotypes of Asian and Black men and women in natural language. By prompting BERT models to perform a "fill-in-the-blank" (cloze) test, we uncovered differences in probability estimates that reflect stereotypes held by English speakers who produced the large corpora of texts used for training the BERT models sampled. First, the BERT models estimated "Asian" people to be more likely than "Black" people to fill in the masked blank for queries describing lower levels of masculinity, physical strength, and assertiveness/dominance, but higher levels of sociability/friendliness and capability/intelligence. These findings from language models replicated and extended previous findings from human participants that identified a "competent but cold and nondominant" descriptive stereotype of East Asian people (Berdahl & Min, 2012; Fiske et al., 2002, 2007; Lin et al., 2005), and a "masculine and athletic" descriptive stereotype of Black people (Galinsky et al., 2013; Ghavami & Peplau, 2013; Lick & Johnson, 2015). Second, the "race is gendered" effect was replicated only in the contrast between Asian men and Black men, but not between Asian women and Black women. This gender nuance in racial stereotypes is consistent with recent findings (Goh & Trofimchuk, 2022) and may help explain why Asian men (but not women) are stereotyped as less moral/trustworthy, which is likely related to their lower prototypicality (Schug et al., 2015). Overall, this research reveals how English speakers naturally expressed intersectional stereotypes about Asian and Black men and women in the real world.

Implications

This research has significant methodological, theoretical, and practical implications. Methodologically, we used the FMAT with BERT language models to specify stereotypes as concrete propositions (e.g., "Most Black men are physically strong.") rather than ambiguous associations between words, providing a more accurate measure than word embeddings. The FMAT is particularly useful for measuring complex concepts that cannot be captured by single words (e.g., physical strength) and for studying intersectional social categories that have multi-word labels (e.g., Black women, Asian men). Therefore, our work addresses important questions that are difficult to answer using previous text-analytic methods such as word counting and word embedding. Our findings also confirm past findings on Asian and Black stereotypes from human participants, suggesting the usefulness and incremental contributions of both language models and the FMAT method in social psychology research, especially for the study of social cognition. The FMAT method should then travel well to broader research questions.

Our research also has theoretical implications, addressing racial (Asian vs. Black) stereotypes beyond gender. The gendered view of racial stereotypes is more parsimonious

than the multifaceted stereotype content view (Martin & Slepian, 2021). However, our findings challenge an overly parsimonious view of racial stereotypes, showing that the gender schema (Bem, 1981; Martin & Slepian, 2021) and the stereotype content model (Fiske, 2018; Fiske et al., 2002, 2007) best *jointly* account for Asian and Black stereotypes. Consistent with the gender schema, Asian people were stereotyped as less masculine and less muscular, as well as less assertive/dominant and more sociable/friendly. Beyond the lens of gender, however, the Asian stereotypes also included competence content, in particular higher capability/intelligence, which aligns with findings in the SCM literature (Fiske, 2018) and previous qualitative analyses (Ghavami & Peplau, 2013; Lick & Johnson, 2015). Importantly, for the Asian (vs. Black) stereotypes, although the overall difference in effect sizes for competence versus warmth ($d = –0.22$) was comparable to the parsimonious gendered effect ($d = –0.24$), we uncovered meaningful nuances in the two facets of competence and warmth (see Table 3 and Figure 2), which should not be neglected when understanding Asian and Black stereotypes. Racial stereotypes are complex—more than gendered—requiring a more comprehensive understanding of their multiple facets and subgroup nuances. This is also in line with recent work using NLP methods suggesting significant nuances for both single groups and intersectional groups (Nicolas et al., 2022; Nicolas & Fiske, 2023).

Finally, the multifaceted contents and gender nuances of Asian stereotypes have practical implications for understanding other phenomena. For instance, the finding that Asian men were stereotyped as less trustworthy may account for a bias against male Chinese scholars (Acciai et al., 2023). In their experiment, authors of scientific papers were less responsive and less willing to share data with male (but not female) Chinese requestors, which, as they suggested, may be due to stereotypes about lower trustworthiness and deservingness of Chinese men (Acciai et al., 2023). Our findings add important natural language evidence consistent with this explanation, showing that Asian men (but not women) were stereotyped as less masculine and less moral/trustworthy. These findings may also provide practical guidance for practitioners and policy makers to mitigate the detrimental effects of racial stereotypes on targets' well-being and real-life outcomes.

Limitations and Future Research

Our research has contributed new insights into how Asian and Black stereotypes are naturally expressed in the real world. A few limitations and open questions warrant further investigation. First, our use of the FMAT to explore subgroup nuances between men and women could be extended to explore stereotypes of other ethnic and racial groups and subgroups, such as Hispanics, the Irish, and Native Americans (cf. Charlesworth et al., 2022). Also, there are likely substantial differences between Asian subgroups, including East Asian, South Asian, Southeast Asian, and West Asian people (Goh & McCue, 2021). The FMAT is particularly useful for this differentiation so as to address the "bamboo ceiling" issue faced by

East Asian people due to their lower level of assertiveness compared to South Asian people (Lu et al., 2020, 2022).

Second, we leveraged large language models and NLP techniques to identify racial stereotypes in people's online expressions, but our conclusions are limited to the population of the content producers specific to the English language models we used. Since this population was likely dominated by English-speaking White people (Dillion et al., 2023), our findings may not generalize to other language-speaking populations. This is a limitation of the BERT models, not the FMAT method. Future studies could employ FMAT using language models trained on different English text corpora and/or non-English texts.

Third, our findings are based on the contemporary texts on which the BERT models were trained, but a sociohistorical perspective on intersectional social stereotypes is also possible. Recent work proposes a sociohistorical model of intersectional social category prototypes, highlighting that both race and gender are socially and historically constructed, impacting the people being racialized and gendered (Lei et al., 2023). Although some of the historical stereotypes of Asian and Black men and women may persist in the contemporary text corpora (e.g., the period of text production in the present study was around the 2010s), it is crucial to explore the sociohistorical roots of these stereotypes and to examine how intersectional race-gender stereotypes have changed over time.

Finally, future research should explore how stereotypes of Asian and Black men and women shape discrimination against Asian and Black people in real-life situations. For instance, might emasculating stereotypes about Asian men (as low in physical strength) encourage anger and opportunistic aggression against them, as seen in the racial abuse directed at Asian men following the outbreak of COVID-19 in Spring 2020? And might hypermasculine stereotypes about black men (as physically formidable) inflate fear, increasing disproportionate (e.g., lethal) uses of force against them, as seen in the 2021 George Floyd case? If these stereotypes of Asian and Black men are shaping patterns of aggression against them, can targeted interventions be designed to counter them?

Conclusion

Stereotypes of Asian and Black people in natural language have multiple facets and meaningful gender nuances. Compared to Black counterparts, Asian men (but not women) were stereotyped as less masculine and less moral/trustworthy, while both Asian men and Asian women were stereotyped as less muscular/athletic and less assertive/dominant—but more sociable/friendly and more capable/intelligent. These intersectional findings suggest how multiple facets of Asian and Black stereotypes are distinctively manifested in people's natural language in the real world, highlighting the need to design interventions to mitigate racial prejudice and discrimination against Asian and Black men and women.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763.

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A Dual Perspective Model. *Advances in Experimental Social Psychology, 50*, 195–255.

Acciai, C., Schneider, J. W., & Nielsen, M, W. (2023). Estimating social bias in data sharing behaviours: An open science experiment. *Scientific Data, 10*, Article 233.

Alexander, M. G., Brewer, M. B., & Livingston, R. W. (2005). Putting stereotype content in context: Image theory and interethnic stereotypes. *Personality and Social Psychology Bulletin, 31*(6), 781–794.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

Axt, J. R., Atwood, S., Talhelm, T., & Hehman, E. (2023). Asian men and Black women hold weaker race–gender associations: Evidence from the United States and China. *Social Psychological and Personality Science, 14*(6), 686–697.

Bao, H.-W.-S. (2023). *FMAT: The Fill-Mask Association Test* (Version 2023.8) [Computer software]. https://CRAN.R-project.org/package=FMAT

Bao, H.-W.-S. (in press). The Fill-Mask Association Test (FMAT): Measuring propositions in natural language. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000396

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint. https://doi.org/10.48550/arXiv.1506.04967

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155–162.

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review, 88*(4), 354–364.

Berdahl, J. L., & Min, J.-A. (2012). Prescriptive stereotypes and workplace consequences for East Asians in North America. *Cultural Diversity and Ethnic Minority Psychology, 18*(2), 141–152.

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly, 45*(1), 106–125.

Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences, 119*(28), Article e2121798119.

Cunningham, G. B., Wicker, P., & Kutsko, K. (2021). Gendered racial stereotypes and

coaching intercollegiate athletic teams: The representation of Black and Asian women coaches on U.S. Women's and Men's Teams. *Sex Roles, 84*(9–10), 574–583.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. https://doi.org/10.48550/arXiv.1810.04805

Diekman, A. B., & Schmader, T. (in press). Gender as embedded social cognition. In D. Carlston, K. Johnson, & K. Hugenberg (Eds.), *The Oxford Handbook of Social Cognition* (2nd ed.). Retrieved from https://doi.org/10.31234/osf.io/wvx2s

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597–600.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Galinsky, A. D., Hall, E. V., & Cuddy, A. J. C. (2013). Gendered races: Implications for interracial marriage, leadership selection, and athletic participation. *Psychological Science, 24*(4), 498–506.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.

Ghavami, N., & Peplau, L. A. (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly, 37*(1), 113–127.

Goff, P. A., Thomas, M. A., & Jackson, M. C. (2008). "Ain't I a woman?": Towards an intersectional approach to person perception and group-based harms. *Sex Roles, 59*(5–6), 392–403.

Goh, J. X., & McCue, J. (2021). Perceived prototypicality of Asian subgroups in the United States and the United Kingdom. *Journal of Experimental Social Psychology, 97*, Article 104201.

Goh, J. X., & Trofimchuk, V. (2022). *Gendered perceptions of East and South Asian faces*. PsyArXiv preprint. https://doi.org/10.31234/osf.io/es89d

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham,

W. A. (2023). AI and the transformation of social science research. *Science, 380*(6650), 1108–1109.

Hall, E. V., Galinsky, A. D., & Phillips, K. W. (2015). Gender profiling: A gendered race perspective on person–position fit. *Personality and Social Psychology Bulletin, 41*(6), 853–868.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology, 102*(1), 116–131.

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

Lei, R. F., Foster-Hanson, E., & Goh, J. X. (2023). A sociohistorical model of intersectional social category prototypes. *Nature Reviews Psychology, 2*(5), 297–308.

Lei, R. F., Leshin, R. A., Moty, K., Foster-Hanson, E., & Rhodes, M. (2022). How race and gender shape the development of social prototypes in the United States. *Journal of Experimental Psychology: General, 151*(8), 1956–1971.

Lei, R. F., Leshin, R. A., & Rhodes, M. (2020). The development of intersectional social prototypes. *Psychological Science, 31*(8), 911–926.

Lick, D. J., & Johnson, K. L. (2015). Intersecting race and gender cues are associated with perceptions of gay men's preferred sexual roles. *Archives of Sexual Behavior, 44*(5), 1471–1481.

Lin, M. H., Kwan, V. S. Y., Cheung, A., & Fiske, S. T. (2005). Stereotype content model explains prejudice for an envied outgroup: Scale of anti-Asian American stereotypes. *Personality and Social Psychology Bulletin, 31*(1), 34–47.

Lu, J. G., Nisbett, R. E., & Morris, M. W. (2020). Why East Asians but not South Asians are underrepresented in leadership positions in the United States. *Proceedings of the National Academy of Sciences, 117*(9), 4590–4600.

Lu, J. G., Nisbett, R. E., & Morris, M. W. (2022). The surprising underperformance of East Asians in US law and business schools: The liability of low assertiveness and the ameliorative potential of online classrooms. *Proceedings of the National Academy of Sciences, 119*(13), Article e2118244119.

Martin, A. E. (2023). Gender relativism: How context shapes what is seen as male and female. *Journal of Experimental Psychology: General, 152*(2), 322–345.

Martin, A. E., & Slepian, M. L. (2021). The primacy of gender: Gendered cognition underlies the Big Two dimensions of social cognition. *Perspectives on Psychological Science, 16*(6), 1143–1158.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–

315.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology, 51*(1), 178−196.

Nicolas, G., Bai, X., & Fiske, S. T. (2022). A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology, 123*(6), 1243–1263.

Nicolas, G., & Fiske, S. T. (2023). Valence biases and emergence in the stereotype content of intersecting social categories. *Journal of Experimental Psychology: General, 152*(9), 2520–2543.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosette, A. S., Koval, C. Z., Ma, A., & Livingston, R. (2016). Race matters for women leaders: Intersectional effects on agentic deficiencies and penalties. *The Leadership Quarterly, 27*(3), 429–445.

Schug, J., Alt, N. P., & Klauer, K. C. (2015). Gendered race prototypes: Evidence for the non-prototypicality of Asian men and Black women. *Journal of Experimental Social Psychology, 56*, 121–125.

Schug, J., Alt, N. P., Lu, P. S., Gosin, M., & Fay, J. L. (2017). Gendered race in mass media: Invisibility of Asian men and Black women in popular magazines. *Psychology of Popular Media Culture, 6*(3), 222–236.

Sesko, A. K., & Biernat, M. (2010). Prototypes of race and gender: The invisibility of Black women. *Journal of Experimental Social Psychology, 46*(2), 356–360.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Wilkins, C. L., Chan, J. F., & Kaiser, C. R. (2011). Racial stereotypes and interracial attraction: Phenotypic prototypicality and perceived attractiveness of Asians. *Cultural Diversity and Ethnic Minority Psychology, 17*(4), 427–431.

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology, 113*(1), 59–80.

Zhu, N., & Chang, L. (2019). Evolved but not fixed: A life history account of gender roles and gender inequality. *Frontiers in Psychology, 10*, Article 1709.