PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Historical representations of social groups across 200 years of word embeddings from Google Books

Tessa E. S. Charlesworth[a], Aylin Caliskan[b], and Mahzarin R. Banaji[a,1]

Contributed by Mahzarin R. Banaji; received December 3, 2021; accepted May 4, 2022; reviewed by Gary Lupyan and Keith Payne

Using word embeddings from 850 billion words in English-language Google Books, we provide an extensive analysis of historical change and stability in social group representations (stereotypes) across a long timeframe (from 1800 to 1999), for a large number of social group targets (Black, White, Asian, Irish, Hispanic, Native American, Man, Woman, Old, Young, Fat, Thin, Rich, Poor), and their emergent, bottom-up associations with 14,000 words and a subset of 600 traits. The results provide a nuanced picture of change and persistence in stereotypes across 200 y. Change was observed in the top-associated words and traits: Whether analyzing the top 10 or 50 associates, at least 50% of top associates changed across successive decades. Despite this changing content of top-associated words, the average valence (positivity/negativity) of these top stereotypes was generally persistent. Ultimately, through advances in the availability of historical word embeddings, this study offers a comprehensive characterization of both change and persistence in social group representations as revealed through books of the English-speaking world from 1800 to 1999.

attitude change | natural language processing | stereotype change | word embeddings

Significance

How did societies of the past represent the various social groups of their world? Here, we address this question using word embeddings from 850 billion words of English-language books (from 1800 to 1999) to uncover principles of change and stability in stereotype content and valence (positivity/negativity) toward 14 social groups. Results revealed nuanced patterns of both change and stability in group representations across 200 y. Change emerged in top stereotype content (which words were most uniquely associated with a group over time), especially for racial/ethnic/nationality groups; yet the valence of group representations generally remained stable. This work highlights empirical and theoretical discoveries that ensue from applying word embeddings to long-standing questions of social change over centuries.

For as long as humans have been writing, they have been writing about their beliefs and attitudes toward the various social groups in society. Analyzing the contents of such historical written texts can therefore provide a window into the social representations of societies that have long since passed (1, 2). We can now ask: How did societies in the 1800s, 1900s, or 1990s represent group concepts as various as age, gender, social class, body weight, ethnicity, or race? What were the traits and words that were most strongly associated with each group? And most importantly, how have these group representations changed across history?

Advances in natural language processing (NLP), coupled with the availability of massive archives of historical text, have newly made it possible to study social group representations at unprecedented scales. The current paper uses word embeddings from 200 y of English-language Google Books text to provide an extensive quantitative and qualitative study of social group representations using a) a long timeframe (20 decades from 1800 to 1999), b) a large number of social groups (14 groups including gender, race, nationality, age, social class, and body weight), and c) an empirical, bottom-up approach to identify stereotypes with d) extensive lists of over 14,000 words and a subset of over 600 traits.

Studying Stereotype Change through Surveys: Insights from the Princeton Quartet

The current research has its roots in one of the most famous demonstrations of social stereotypes—the Princeton quartet. Starting with Katz and Braly in 1933 (3), followed up by Gilbert in 1951 (4), Karlins and colleagues in 1969 (5), and most recently updated by Bergsieker and colleagues for 2000 to 2007 (6), these four studies surveyed four generations of college students at Princeton who explicitly selected the top five traits that they most associated with a set of 10 racial and ethnic groups. The primary conclusion reported from these studies is the "fading" of negative representations of racial and ethnic outgroups. For example, in 1933, the stereotypes of Black Americans as "lazy" and "superstitious" were endorsed by 75% and 84% of respondents, respectively; but by 2000 to 2007, only 11% and 3% of respondents reported holding these same stereotypes (6). Additionally, the authors of later studies noted how participants were reticent to report any stereotypes at all and were especially opposed to endorsing negative stereotypes (6). It remains to be seen whether this fading negativity holds 1) for nonracial groups (e.g., age, gender, class) that may be less contentious in public discourse and 2) over much longer time spans of 200 y.

Author affiliations: [a]Department of Psychology, Harvard University, Cambridge, MA 02138; and [b]Information School, University of Washington, Seattle, WA 98195

Author contributions: T.E.S.C., A.C., and M.R.B. designed research; T.E.S.C. performed research; T.E.S.C. analyzed data; and T.E.S.C., A.C., and M.R.B. wrote the paper.

Reviewers: G.L., University of Wisconsin–Madison; and K.P., University of North Carolina at Chapel Hill.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: mahzarin_banaji@harvard.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2121798119/-/DCSupplemental.

Published July 5, 2022.

Beyond the changing valence (positivity/negativity) in group representations, the Princeton quartet also revealed how the semantic representations changed over time. On the one hand, the top-associated traits changed over time, generally with a positive word emerging while a negative word fell away. On the other hand, there also appeared to be persistence in deeper, underlying semantics of racial and ethnic stereotypes (6, 7). For instance, between 1933 and 1967, the trait "stolid" was replaced by "efficient" to describe Germans, and "conventional" was replaced with "conservative" to describe the English (5). Thus, while the trait itself changed, the broader semantic representation remained relatively stable. However, past studies could artificially show stability in stereotype content because of the relatively small number of traits (84 traits) for participants to draw from. The current methods offer a more compelling test of stability or change in group representations by using a larger sample space of traits.

Studying Stereotype Change through Texts: Insights from Word Embeddings

This project also draws on a second research tradition: applying NLP to the study of psychological phenomena (8, 9). Recently, there has been an explosion of research transforming text data into word embeddings—vector representations of word meaning computed from billions of word co-occurrences (10)—and then analyzing changes in the associations between word embeddings. Today, researchers across the social and computer sciences have applied historical (also called "diachronic embeddings") word embeddings to study topics including antisemitism in French newspapers from the late 1700s (11), the evolution of scientific concepts in science journals since 1655 (12), gendered stereotypes in historical books since 1910 (13). Most relevant to the current work, Garg and colleagues (14) used embeddings from book and newspaper text between 1910 and 2005 and identified the top traits and occupations associated with gender and racial/ethnic groups.

Today, using word embeddings, researchers can expand beyond past work in the social sciences, which has largely relied on explicit, self-report measures to study stereotype change.

The Current Manuscript: Empirical Discovery of Stereotype Content across History

Their many contributions notwithstanding, both the Princeton quartet and the adoption of NLP methods face a limitation for inference: They require experimenters to decide, a priori, which traits, words, or domains are most likely to characterize the stereotypes of certain groups. For instance, in a typical word embeddings study of social group biases using the Word Embeddings Association Test from Caliskan and colleagues (15), the researcher selects a set of words to represent a concept (e.g., science vs. arts) and its association with groups (e.g., female vs. male). This design approximates the typical studies of social scientists, in which a small set of traits is provided to the participant to use in their descriptions of social groups (3–6). In short, the dominant approach in both NLP and survey research does not yet enable an empirical discovery of social biases bottom-up, without prespecified dimensions.

Here, we bridge research on stereotype change and NLP methods to offer a bottom-up, empirical portrait of social representation change in text. We examine the associations between 14 social groups (spanning gender, race, nationality, age, body weight, and class) and lists of nearly 14,000 words (16) and a subset of 600 traits (17)—to our knowledge, these are the most extensive lists available that also provide ratings of the words' and traits' valence (positivity/negativity).* Using pretrained embeddings from 20 decades of English-language Google Books text between 1800 and 1999 (18), we then analyze cosine similarities between word embeddings in each decade to identify which of the 600 traits (e.g., "competent") and 14,000 words (e.g., "leader") emerge empirically, bottom-up, as most associated with a given social group concept (e.g., Male). Crucially, we do not prespecify which domains of words we are interested in a priori, allowing for both the validation of the method (when the emergent words converge with well-known stereotypes such as Man–leader or Woman–maid) as well as the potential for novel discoveries on which domains of stereotype emerge. Finally, using the ratings of word and trait valence provided by previous researchers (19), we also calculate whether the stereotypes have shifted in their degree of average positivity/negativity over time.

Overview of the Methodological Approach

Data Source. The starting point for our project is the English-All (Eng-all) Google Books n-grams dataset (second version) (20), a corpus of approximately 850 billion words of all English books archived over 20 decades from 1800 to 1999. The Google Books dataset is estimated to represent 4% to 6% of all books published, thus capturing dominant cultural conversations across history (21). From the Eng-all texts, we used pretrained word embeddings provided by Hamilton and colleagues (11) (*SI Appendix* provides further details on how the authors trained the embeddings). Of course, the Eng-all dataset has limitations (22, 23). For instance, there are biases in who is featured as authors, in the English- and Western-centric focus of texts, and in the imbalance between nonfiction and fiction (23). To address these concerns, we replicate all results in a smaller but genre-balanced corpus, the *Corpus of Historical American English* (COHA; ref. 24; see *SI Appendix*), and find similar results.

Selection and Representation of Social Groups in Group Label Lists. To characterize patterns of change in social group representations, we chose social groups that a) were sufficiently diverse to arguably represent a broad sample space of societal stereotypes, b) could be well represented by a sufficient number of group label synonyms, c) had a clear comparison group, and d) had group labels that were reasonably stable in meaning across time. In the end, we chose to investigate 14 groups (see Table 1), with each group represented through a list of group labels derived from online searches and dictionaries (e.g., *Oxford Historical Thesaurus*).

As in any study of social group stereotypes, the choice of labels to represent the group can affect the final stereotype representation (e.g., Black vs. African American conjures different psychological representations; ref. 25). In *SI Appendix*, we show that the top word associates to the group label lists convey the intended group-specific meanings of the list. Additionally, we outline potential concerns in the choice of group labels, including, for instance, whether the inclusion of slur-words (e.g., the "N-word"), or the inclusion of a polysemous term (e.g., "Black" that could also refer to color), substantially alters results. In

*In a supplemental analysis, we also investigate the associations to all 100,000 possible words in the word embedding vocabulary. The qualitative content revealed from the 14,000 list and 100,000 list of words are similar. However, the latter has the disadvantage of including words less relevant to the understanding of social stereotypes (e.g., first names, prepositions, conjunctives), and we therefore focus on the other lists.

general, we find that the inclusion (or exclusion) of a single term in the longer list of labels does little to alter the overall results.

These 14 social groups provide an opportunity to test historical traces of contemporary differences in attitude and stereotype change. That is, contemporary data from 2007 to 2020 with implicit and explicit tests have shown that race attitudes (Black/White), and gender stereotypes (e.g., Man–Science/Woman–Arts) have decreased in bias over the past decade (26, 27). In contrast, attitudes about age (Young, Old), weight (Fat, Thin), and representations of social class (Rich, Poor) have remained high in bias and largely stable across time (26, 28). These findings separate groups that have been evolutionarily relevant to status and health (age, class, gender) from so-called "arbitrary" groups (racial/nationality groups) that have been relatively recently culturally constructed and that may be more open to change (29). We can now quantitatively test the long historical roots of such empirical and theoretical differences in change across groups.

Computing Social Group Representations from Word Embeddings. We follow Manzini and colleagues (30) in computing the mean average cosine similarity (or MAC) between a single target group and lists of nearly 14,000 words (19) and a subset of over 600 traits (17). To compute the MAC, we first calculate all pairwise cosine similarities between a given target word (e.g., strong) and each available group label within a decade (e.g., the cosine similarities of strong–Europeans, strong–European, strong–Caucasians, and so on for all labels representing White). We then average those pairwise cosine similarities to give the average cosine similarity of the target word to the social group concept (e.g., strong–White) within each decade. A positive (negative) MAC indicates that the target word (e.g., strong) is positively (negatively) associated with the given social group. While the MAC on its own has been used elsewhere (30), we perform additional transformations (e.g., by looking at the relative MAC between groups, calculating valence, computing cross-group correlations) that provide more focused insights into the unique content of social group representations. Each new transformation is described before the relevant results below.

Results

Overview. We organize the results to answer four questions: First, collapsing across time, which of the 600 traits and 14,000 words[†] are most strongly and uniquely associated with a target group? Second, bringing time into the equation, which traits and words are the top 10 and 50 associates with a target group for each of the 20 decades from 1800 to 1999? What changes can be seen with respect to semantic content as well as the valence of these top associates? Third, moving beyond just the top associates for each group, how does the full list of over 600 traits and 14,000 words change or remain stable for a given target group over time? That is, to what extent are the cosine similarities between a social group and all words or traits in 1800 correlated with the cosine similarities for all words or traits in 1810, 1820, and so on? Fourth, moving beyond analyses of a single group, how has the overlap in representations between two groups (e.g., Old–Young, Rich–Poor) increased or decreased over 200 y?

[†]Throughout the remainder of the manuscript, we will refer to the results in terms of "words" and "traits," referencing the list of 14,000 words versus the list of a subset of 600 traits. Additionally, given space constraints, we only report the detailed results for the trait analyses, leaving the details of the word analyses to *SI Appendix*.

Analysis 1: Top Relative Word and Trait Associates, Collapsing across 200 y. Over all time, we identify the top 10 (and 50) word and trait associates that are most strongly, relatively associated with one group (and least associated with a comparison group) by 1) computing the MAC of all words and traits to each individual target group in each decade (e.g., the MAC of all traits to Irish in 1800), as well as to a comparison group (e.g., the MAC of all traits to White in 1800); 2) calculating the difference of the MAC scores; and 3) averaging these MAC differences across all 20 decades and ranking the words or traits according to their difference scores (Table 1).

Qualitative content of top unique trait associates, over all time. Given the qualitative nature of the data, we recognize that there may be multiple interpretations of the stereotype content. Below, we focus our discussion on selected broad themes; we remain open to other researchers interpreting any detailed trends in the data generated through this paper.

First, the most negatively stereotyped groups were Poor, Black, Hispanic, and Irish (in that order). Although all four of these groups are represented with predominantly negative traits, we see differences in the source of that negativity. Black, Hispanic, and Irish (all racial groups) are negative largely with reference to negative character (e.g., cruel, deceitful). In contrast, Poor stands out as having the most associations to negative competence (e.g., helpless, weak); indeed, whenever the dimension of competence emerged for Poor, it always highlighted a lack of competence. Similarly, for the relatively neutral representations of both Women and Young, the few negative words referred to the absence of competence (helpless). In sum, it appears that, in historical texts, negativity for racial groups centers on character-related traits, while negativity for nonracial groups centers specifically on competence-related traits.

What are the most positively viewed groups? Fat, Men, and Rich (in that order). The positive content of Rich and Men representations are similar in that both are characterized by high proportions referring to positive competence (even more so than positive character), in line with their historical and contemporary dominance in society. Additionally, as discussed below, the positive representation of Fat is driven by the early decades when Fat connoted ideas of prosperity (e.g., generous), and a jovial character (e.g., jolly, merry).

Qualitative content of top unique word associates, over all time. Next, we briefly discuss the representations using the more comprehensive list of 14,000 words but leave additional details to *SI Appendix*. Here, we again find that the groups Poor, Black, and Hispanic have negative representations, with each portraying a slightly different source of that negativity: For instance, negativity appears centered on notions of dejection and ill health for Poor; disgust and enslavement for Black; and uprisings and unrest for Hispanic. Thus, the more comprehensive word associations reveal additional semantic nuance in stereotype content that did not emerge in the smaller subset of trait associations. In general, however, the emergent word-based representations converge with the known historical record of social group stereotypes.

Convergence with past research does not mean the current word embedding method is redundant with past approaches. Rather, convergence suggests that group representations are sufficiently collective and widespread to be detected by diverse methods. Additionally, the word embedding method can provide novel, quantitative insights into which of the well-known stereotypes of social groups are most dominant across time in our historical language. For instance, the finding that Hispanic appears focused on disobedience or uprisings rather than other features (e.g., culinary or music associations, stereotypes of

Table 1. Top 10 traits emerging as most strongly (relatively) associated with group A (and least with group B), collapsing across all time

Group A (vs. Group B)	Top 10 traits: Relative association	Valence
White (vs. Black)	Critical, polite, hostile, decisive, friendly, diplomatic, understanding, philosophical, able, belligerent	0.65
Black (vs. White)	Earthy, lonely, cruel, sensual, lifeless, deceitful, helpless, rebellious, meek, lazy	−1.18
Asian (vs. White)	Pompous, theatrical, verbal, superstitious, curious, traditional, melancholy, solemn, artificial, sensual	0.08
Irish (vs. White)	Passionate, pompous, melancholy, fanatical, headstrong, sly, grim, sarcastic, solemn, romantic	−0.45
Hispanic (vs. White)	Verbal, pompous, formal, solemn, abrupt, diplomatic, impetuous, traditional, evasive, lifeless	−0.63
Native American (vs. White)	Superstitious, rude, earthy, wholesome, spontaneous, artificial, kind, sensual, lonely, dependent	0.30
Men (vs. Women)	Able, competent, enterprising, honest, independent, brave, efficient, confident, practical, decisive	1.75
Women (vs. Men)	Charming, feminine, soft, romantic, modest, fair, lonely, gentle, tender, helpless	1.15
Old (vs. Young)	Traditional, pompous, solemn, humble, dignified, strict, grim, detached, diplomatic, possessive	−0.42
Young (vs. Old)	Vigorous, bright, hopeful, alert, fair, helpless, intellectual, thoughtless, patient, tender	0.97
Fat (vs. Thin)	Jolly, brave, honest, merry, generous, cheerful, wholesome, intelligent, compassionate, angry	2.06
Thin (vs. Fat)	Logical, theatrical, flexible, original, rigid, brilliant, superficial, detached, precise, artificial	0.43
Rich (vs. Poor)	Dominant, brilliant, dignified, conservative, decisive, respectable, diplomatic, independent, intellectual, artistic	1.45
Poor (vs. Rich)	Helpless, lonely, weak, lazy, dull, stupid, worried, ignorant, cold, careless	−1.85

dirtiness and poverty) suggests that the dominant historical Hispanic stereotypes may be centered on their potential disruption of hierarchies—a stereotype that persists today in the fear of Latinx immigrants taking "American" jobs (31). In this way, perhaps understanding which features of social group stereotypes are emphasized (or omitted) in the top associates of historical texts can guide an understanding of the stereotype dimensions that will shape group relations today as well.

Analysis 2: Top Relative Word and Trait Associates, Separating across Time. In the next analysis, we take advantage of the 200 y of historical data to examine the top trait and word associates separated by each decade. We use decade-specific MAC difference scores and rank words and traits within each decade (Table 2). We also summarize change in the top 10 (and top 50) trait and word associates over decades by counting the number of traits/words that overlapped in successive decades (e.g., 1800 to 1810, 1810 to 1820, etc.; see *SI Appendix*). Finally, we compute average decade-wise valence of the top trait and word associates by averaging ratings of the top words' positivity/negativity (obtained from previous researchers; ref. 19) in each of the 20 decades.[‡]

Qualitative content of top trait associates, separating across time. We look first at the top traits emerging for the six racial/nationality groups (Table 2). For White, the traits conveyed stereotype content of diplomacy and competence, perhaps capturing the persistent historical associations of White with high competence but also with colonialism and dominance. In contrast, the top traits for Black were evidently negative (e.g., lonely, rebellious, untidy) and generally with reference to negative character (rather than to negative competence). Similarly negative character-focused content was observed for Asian, Irish, and Hispanic group concepts, with these three groups also showing some overlapping content (e.g., traits such as pompous appeared frequently in all three groups' top lists). Perhaps such overlap may indicate that non-White racial groups were written about in similar, disparaging ways (similar to an "outgroup homogeneity" effect). Finally, Native American was associated with generally negative traits, although with group-specific, and sometimes positive, content emerging (e.g., spiritual and wholesome) that captured the complexities of the "noble savage" myth (32).

Turning next to nonracial groups: For gender, the traits in all decades consistently differentiated men as competent and able versus women as charming, feminine, and warm, aligning with known gender stereotype dimensions (33). For age groups, the emergent dimension dividing Old and Young appeared to emphasize restraint (e.g., objective, traditional for Old) versus impulsivity (e.g., emotional, impulsive for Young). It is notable that restraint/impulsivity was more obvious than dimensions such as health or incompetence/competence that appear more prevalent in contemporary stereotypes (34). Next, Fat was associated with positive traits in early decades (e.g., merry, generous) but, by the mid-1900s, acquired negativity through traits associated to disease (e.g., "patient" likely to a patient with obesity) and a lack of control (e.g., negligent). Finally, for social class, Rich consistently emphasized status using both positive, earned status (e.g., brilliant) and more negative, forceful status (e.g., dominant, obnoxious). In contrast, Poor always emphasized the negative stereotype content of incompetence and helplessness (indeed, helpless was the top trait in every single decade).

Qualitative content of top word associates across time. Among the thousands of possible words that could be associated with a target group, it seems particularly impressive that the final outputs were generally interpretable through existing theories of

[‡]Assessing valence or favorability is difficult when examined across historical time since words that are rated as positive/negative by contemporary raters may have changed somewhat in their valence over 200 years. Nevertheless, following Karlins and colleagues (1969) (5), we note that contemporary ratings of word valence is our best approximation and that ratings of favorability are often highly correlated across time (rank-order correlation of $r = 0.88$ reported by Karlins).

Table 2. Top 10 traits emerging as most strongly (relatively) associated with group A (and least with group B), separated by decade

Group A (vs. Group B)	1800	1850	1900	1950	1990
White (vs. Black)	Confident, able, complex, consistent, understanding, agreeable, inconsistent, critical, decisive, cautious	Critical, conservative, diplomatic, impartial, discriminating, cordial, kind, philosophical, consistent, polite	Enterprising, reliable, intelligent, critical, thoughtful, conservative, able, philosophical, cultured, friendly	Belligerent, diplomatic, philosophical, dominant, profound, conventional, decisive, friendly, outspoken, concise	Friendly, profound, intellectual, diplomatic, analytical, accurate, skeptical, understanding, hostile, cooperative
Black (vs. White)	Lonely, helpless, lifeless, rebellious, deceitful, tranquil, forgiving, cruel, charitable, peaceful	Rebellious, helpless, headstrong, rash, wasteful, cruel, sensual, worried, soft, irritable	Untidy, foolhardy, egotistical, listless, wasteful, spiteful, lonely, deceitful, despondent, lifeless	Envious, untidy, sluggish, unruly, earthy, bashful, devious, bright, angry, listless	Suave, grouchy, greedy, egotistical, lustful, earthy, playful, inconsiderate, opinionated, indecisive
Asian (vs. White)	Pompous, theatrical, superstitious, subtle, profound, crude, indiscreet, sensual, concise, crafty	Pompous, superstitious, traditional, theatrical, deceptive, extravagant, clumsy, cynical, curious, wasteful	Wasteful, pompous, untidy, lifeless, pretentious, listless, theatrical, nervous, heartless, spiteful	Verbal, pompous, fickle, punctual, unconventional, envious, melancholy, devious, abrupt, peaceful	Egotistical, gullible, suave, inconsiderate, fussy, listless, glum, indecisive, spiteful, evasive
Irish (vs. White	Theatrical, pompous, crude, expressive, articulate, superstitious, dignified, passionate, subtle, abrupt	Headstrong, worried, stubborn, talkative, pompous, spiteful, jolly, fanatical, noisy, stupid	Passionate, bashful, egotistical, deceitful, melancholy, argumentative, sarcastic, persuasive, gentle, pompous	Melancholy, romantic, rude, forgiving, expressive, feminine, scornful, unconventional, humorous, grim	Glum, spiteful, solemn, unruly, sensual, grouchy, conceited, opinionated, lively, pompous
Hispanic (vs. White)	Theatrical, pompous, superstitious, subtle, profound, lifeless, abrupt, flattering, persuasive, arrogant	Pompous, formal, precise, verbal, subtle, traditional, conventional, indecisive, impatient, diplomatic	Wasteful, pompous, abrupt, ruthless, indecisive, verbal, formal, reckless, detached, assertive	Indiscreet, verbal, devious, possessive, constructive, formal, critical, progressive, impetuous, diplomatic	Verbal, informal, abusive, emotional, unethical, complex, antisocial, social, persistent, reflective
Native American (vs. White)	Cold, dissatisfied, wholesome, superstitious, obnoxious, disagreeable, friendly, gloomy, warm, pleasant	Spontaneous, sensual, natural, curious, simple, emotional, ignorant, dependent, deceptive, rebellious	Wasteful, nervous, superstitious, crude, listless, lonely, lifeless, earthy, inconsiderate, rude	Wholesome, sensual, artificial, coarse, soft, rude, spiritual, kind, natural, dependent	Inconsiderate, glum, listless, fussy, lustful, egotistical, gullible, sluggish, easygoing, suave
Men (vs. Women)	Independent, practical, decisive, active, responsible, efficient, enterprising, rational, vigorous, brave	Efficient, active, confident, able, systematic, enterprising, philosophical, decisive, independent, responsible	Direct, practical, able, honest, bold, enterprising, confident, strong, efficient, aggressive	Able, honest, strong, satisfied, tough, confident, stupid, courageous, curious, competent	Honest, clever, ignorant, suspicious, tough, skeptical, cautious, stupid, shrewd, able
Women (vs. Men)	Charming, lonely, helpless, fair, cordial, soft, merry, silent, modest, tender	Charming, feminine, bashful, lonely, soft, romantic, modest, fair, gentle, silent	Feminine, charming, untidy, modest, refined, immature, fair, bright, despondent, gentle	Feminine, untidy, lifeless, playful, devious, soft, sluggish, romantic, heartless, scornful	Feminine, artificial, abusive, romantic, observant, insightful, dependent, compulsive, materialistic, social
Old (vs. Young)	Objective, blunt, exacting, upright, independent, pompous, dependent, verbal, stable, spiritual	Diplomatic, traditional, pompous, grim, solemn, peaceful, informal, humble, possessive, coarse	Strict, respectable, stern, traditional, obnoxious, grim, disrespectful, pompous, conventional, informal	Haphazard, dignified, polite, unkind, rude, discreet, pompous, strict, detached, disrespectful	Steadfast, conventional, bossy, gracious, deceitful, loyal, untidy, strict, humble, fickle
Young (vs. Old)	Vigorous, tender, strong, romantic, intense, soft, enthusiastic, happy, helpless, warm	Impulsive, thoughtless, spirited, clever, sensitive, vigorous, bashful, witty, hopeful, bright	Emotional, patient, intellectual, bright, imaginative, sensual, hopeful, rational, impulsive, sensitive	Emotional, intellectual, vigorous, creative, optimistic, sophisticated, immature, patient, intelligent, social	Aggressive, antisocial, reflective, inventive, impulsive, vigorous, severe, imaginative, irritable, manipulative
Fat (vs. Thin)	Loyal, forgiving, brave, merry, obedient, angry, rebellious, cheerful, friendly, spirited	Jolly, merry, honest, brave, cheerful, intelligent, obedient, talkative, gracious, active	Jolly, generous, stingy, moderate, merry, enthusiastic, sophisticated, rash, stable, loyal	Methodical, nervous, cultured, patient, immature, immoral, unstable, sociable, mischievous, negligent	Antisocial, dependent, inept, immature, excitable, cultured, wholesome, depressed, patient, sensitive
Thin (vs. Fat)	Logical, theatrical, superficial, sensual, progressive, relaxed, casual, frivolous, intellectual, lifeless	Superficial, gloomy, frivolous, theatrical, artificial, detached, rigid, deceptive, unfair, conventional	Stern, flexible, rigid, detached, insecure, deceptive, clumsy, lifeless, curious, neat	Blunt, brilliant, stern, articulate, steady, defensive, bold, expressive, curious, neat	Articulate, brilliant, discreet, solemn, grim, neat, melancholy, artistic, gloomy, impatient
Rich (vs. Poor)	Brilliant, decisive, efficient, dignified, original, respectable, independent, mature, obnoxious, belligerent	Dominant, conservative, dignified, uncompromising, respectable, refined, talented, systematic, original, brilliant	Social, ethical, dominant, intellectual, brilliant, consistent, loyal, dignified, diplomatic, conservative	Materialistic, philosophical, progressive, conservative, intellectual, brilliant, dominant, reserved, traditional, artistic	Dominant, diplomatic, intellectual, artistic, religious, idealistic, traditional, loyal, romantic, masculine
Poor (vs. Rich)	Helpless, lazy, stupid, lifeless, tough, stable, dull, merry, lonely, vulgar	Helpless, lonely, lazy, ignorant, worried, weak, careless, shy, lifeless, devious	Helpless, lonely, worried, weak, melancholy, cold, dull, stupid, bashful, negligent	Helpless, weak, angry, untidy, indifferent, worried, lonely, cold, heartless, patient	Helpless, irritable, cheerful, careless, satisfied, impatient, easygoing, sober, negligent, severe

stereotype content (see *SI Appendix* for a full discussion of the content). For instance, for racial groups, results again showed that the group concept White contained words referring to diplomacy (e.g., foreign, century). Black, in contrast, was consistently associated with negative, role-related words (e.g., slave), as well as bodily words (e.g., fart), perhaps emphasizing the historical preoccupations with the Black body. Gender groups showed Men consistently associated with competence words (e.g., effective, ability) and Women associated with family (e.g., relationship roles like aunt, spouse) or sexuality (virgin, fertility).

Quantitative changes in the overlap of top trait and word associates across time. How many traits (or words) appeared in the top 10 or top 50 lists for both successive decades (e.g., 1800 and 1810, 1810 and 1820)? Whether we took the top 10 or 50 associates, we found that, on average, 34% (top 10) and 48% (top 50) of traits and 26% (top 10) and 27% (top 50) of words overlapped for two successive decades. The generally lower numbers of overlap observed for words (versus traits) is understandable given the substantially larger "pool" of words to be drawn from. Taken at the highest level across groups and metrics, at least 50% (but sometimes as much as 70% to 80%) of the top stereotype associates was changing over each successive decade. See *SI Appendix* for detailed results.

Quantitative changes in valence of top trait and word associates across time. How much did the average valence of the top trait (and word) associates change over 200 y? As can be seen in Fig. 1, the dominant pattern in the valence timeseries was stability. For traits, 9 (out of 14) groups revealed no significant slopes of positivity/negativity over time. Of the remaining five changing groups, four showed decreasing positivity over time: Fat, Native, Asian, and Young (b range = [−0.11, −0.06], all $P < 0.02$). The trait representation of Poor was alone in showing increasing positivity ($b = +0.05$, $P < 0.001$). Similarly, for word representations, 9 (out of 14) groups revealed no significant slopes: Fat, Asian, Young, and Rich decreased in positivity over time (b range = [−0.09, −0.05], all $P < 0.01$), while Men became slightly more positive ($b = +0.05$, $P = 0.01$). Thus, at least for the top trait and word associates, the positivity or negativity of most (albeit not all) social group representations in text appears to be remarkably stable over 200 y.

Valence stability can exist alongside aforementioned changes in the content of top stereotypes over time. For instance, the exact words that are negative can change over time (e.g., lonely is a top associate in 1800, while untidy is a top associate in 1950 for the Black stereotype), changing the content while preserving the valence. This dominant pattern—of valence stability despite changes in top content—departs from the Princeton quartet studies of racial/ethnic stereotypes, where the authors reported more evidence of valence change (at least for the most negatively stereotyped groups of African Americans and Turks). The difference may lie in measurement: Self-reported stereotypes assessed in the Princeton quartet are likely more sensitive to concerns of social desirability and norms against expressing any negative group stereotypes (6). In contrast, the more indirect assessment of co-occurring words in historical texts may be more able to pick up on the persistence of negativity in group representations across time.

Analysis 3: Consistency of All Word and Trait Associates across Time. We next turn to the full trait and word space: the associations between a target group and all words (of which 12,236 were available in the pretrained embeddings) and the subset of traits (of which 400 were available). We test the correlation between the trait-to-group cosine similarities computed in 1800, for example, with those from 1810, 1820, and so on for all pairs of decade-wise correlations. For illustration, to investigate the stability of words associated with the group concept Black, we compute the Pearson's correlation between the MAC effect sizes for 12,236 words in 1800 and the MAC effect sizes for the same 12,236 words in 1810, 1820, and so on resulting in a 20-by-20 correlation matrix across time (Fig. 2; see *SI Appendix* for words).

General patterns of consistency across groups. Beginning with the furthest possible correlation—1800 correlated with 1990 or a 200-y span of historical text—and averaging across all groups, the average correlation was $r = 0.52$ (traits) and $r = 0.55$ (words), with a range of $r = 0.26$ to $r = 0.73$, depending on the group (Table 3). Such consistently positive correlations across 200 y suggest that the general space of words used to describe these 14 social groups retains at least some content even over an extensive time span.

Although this correlation across 200 y is already an impressive starting point, the correlations generally increased in magnitude as the years became closer in time. Taking the averages of columns in Table 4 shows that the average correlations across a 150-, 100-, 50-, or 10-y period were $r = 0.60$, 0.69, 0.76, and 0.82 (traits) or $r = 0.64$, 0.72, 0.77, and 0.83 (words), respectively.[§] Crucially, this general increase in correlations would not have occurred if the full representations were perfectly consistent across time, since perfect consistency in trait and word associates would have resulted in equivalently high correlations regardless of the chosen pair of decades. The results demonstrate that 1) social group stereotypes maintain a portion of similar content even across 200 y (reflected by the moderate correlations over 200 y), and yet 2) there are gradual shifts in the semantic content of stereotypes over time, with decades close in time having more similar words than decades far in time.

Differences in patterns of consistency across groups. Although all groups exhibited roughly similar patterns over time (starting with a moderate-to-high correlation and increasing gradually as decades become closer), there were nevertheless some groups that revealed higher or lower consistency than others. Specifically, the representations of Women and Men were the most consistent across all time (highest mean correlations in Table 4), perhaps because gender is used as a consistent organizing social dimension of every society (29). In contrast, representations of Hispanic and Asian showed the lowest consistency over time (lowest correlations in Table 4). More generally, all nonracial groups showed greater consistency across time than all racial/nationality groups. Such a finding is interpreted as quantitative evidence for the theory that racial/nationality groups are arbitrary and thus less stable (29). In contrast, nonracial groups (particularly gender and age) are more evolutionarily relevant to maintaining social hierarchies and thus may have more consistent representations over historical texts.

Analysis 4: Overlap in Word and Trait Associates across Two Groups over Time. In the fourth and final analysis, we consider how the word and trait representations of two connected but oppositional groups (e.g., Rich–Poor, Old–Young) have converged (overlapped) or diverged (separated) across historical English text. That is, we ask: To what extent are the same words and traits associated with two contrasting groups (e.g., Male and Female) and how have such associations between pairs of groups changed over time? The answer has initial hints in the Princeton quartet: When stereotype change was observed in those studies, it often

[§]The pattern of increasing correlations is consistent regardless of whether we look at early years (e.g., the average correlation across the first 100 years, 1800 to 1900, $r = 0.69$ for traits and $r = 0.72$ for all words) or later years (e.g., the average correlation across the last 100 years, 1900 to 1990, $r = 0.69$ for traits and $r = 0.72$ for all words).

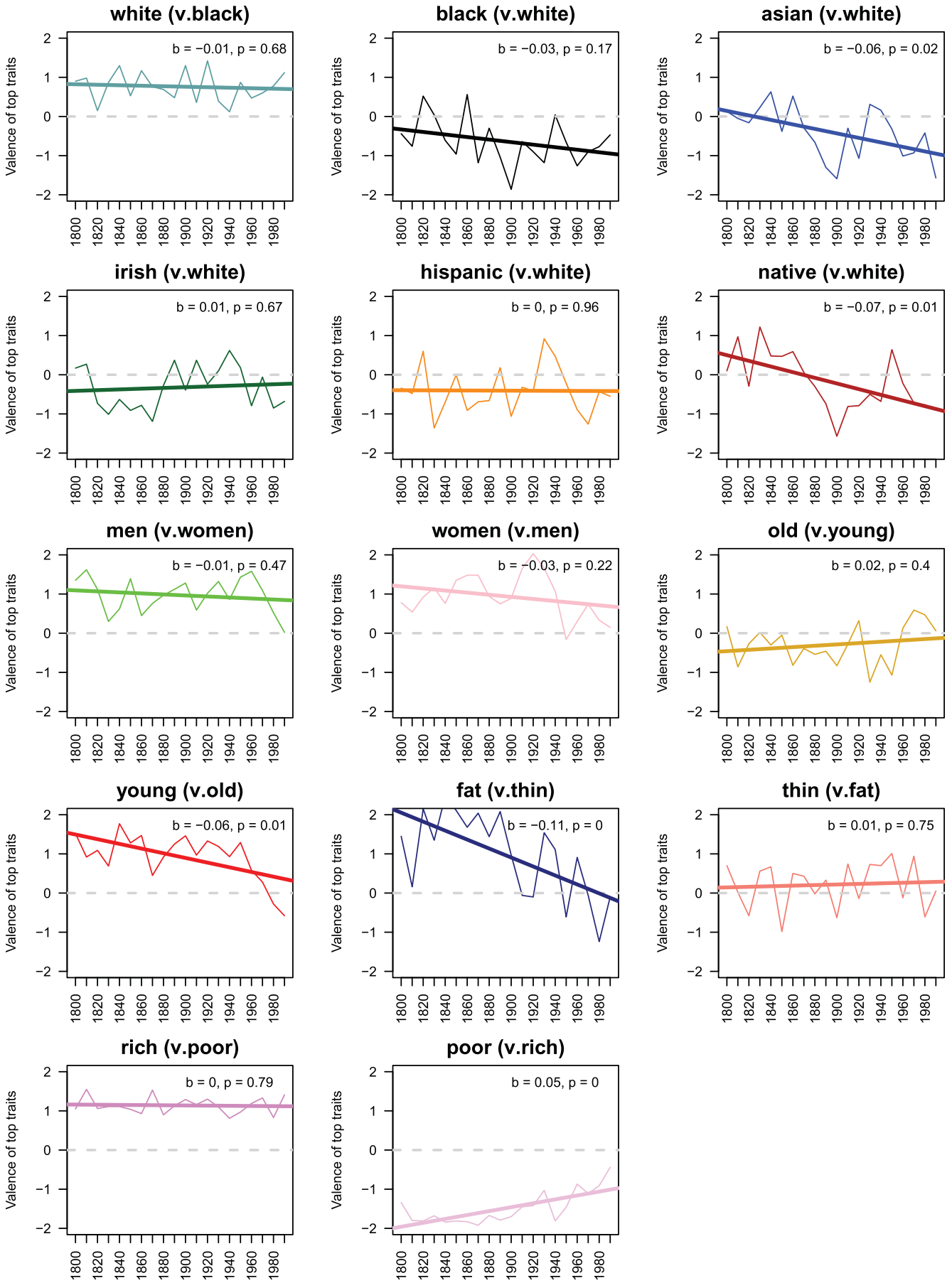

Fig. 1. Valence timeseries of top 10 traits. Valence computed from the average valence rating (from ref. 19) for the top 10 traits in each decade (on a scale from −4 [very negative] to +4 [very positive]). For plots of valence timeseries for words, see *SI Appendix*.

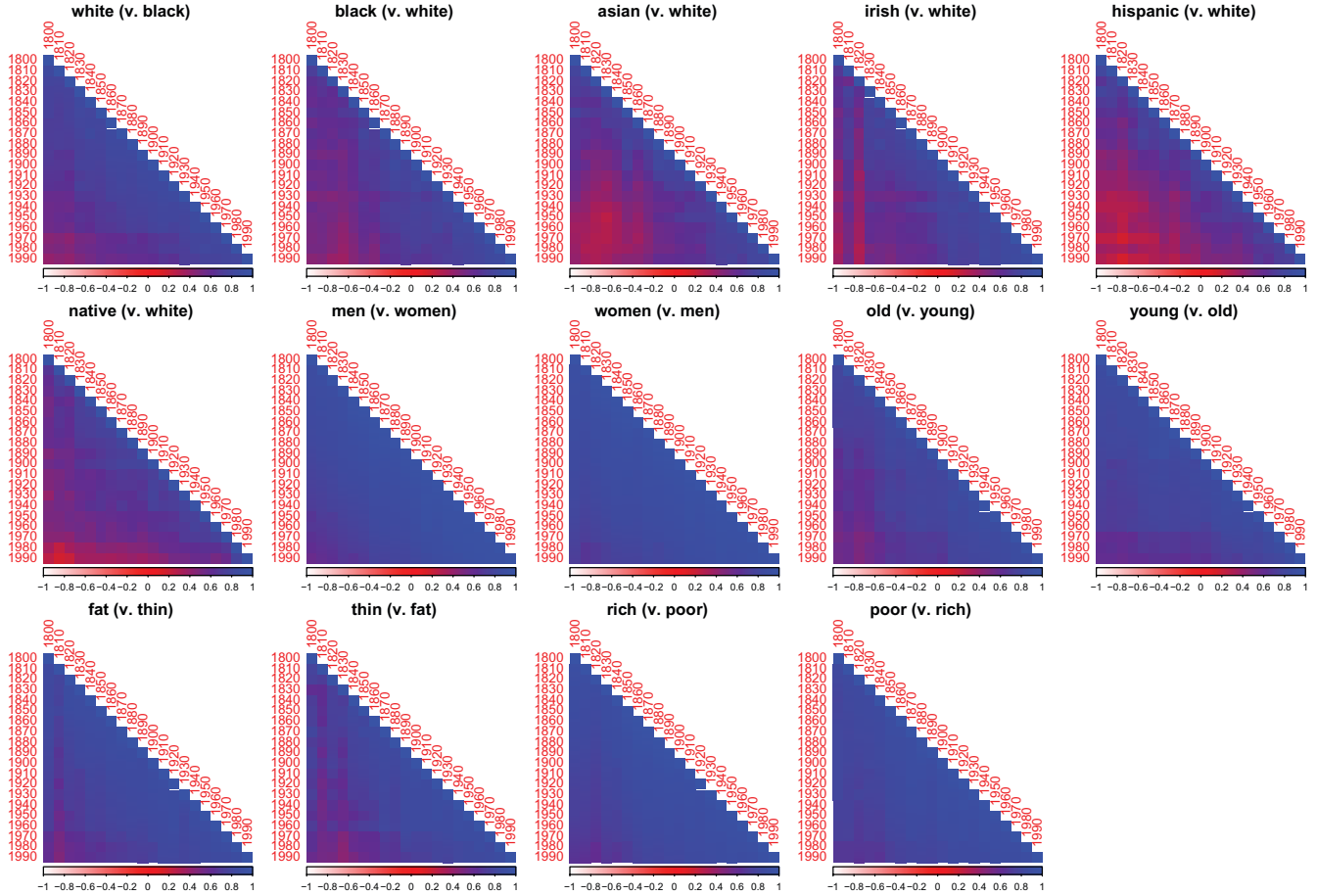

Fig. 2. Within-group correlations of trait representations across 20 decades of historical text. Red colors indicate lower correlations, and blue colors indicate higher correlations. Correlations are computed between the vectors of all trait-group cosine similarities for each pair of decades. The first column of the triangle indicates the correlations between representations in 1800 with all other decades (1810 to 1990); the last row of the triangle indicates the correlations in 1990 with all other decades (1800 to 1980); and the bottom corner indicates the furthest pair of decades (1800 to 1990) while increasing toward the diagonal indicates closer pairs of decades. For correlation plots of word representations, see *SI Appendix*.

moved in the direction of increasing the similarity between group representations. That is, groups described with different traits in Katz and Braly's (3) original study in 1933 were described with more overlapping traits in later iterations of the studies (e.g., Irish and Italians were both "very religious") (5). Until now, this finding of group convergence has rarely been discussed, let alone examined empirically.

To study overlapping representations between two relevant comparison groups over time, we compute the Pearson's correlation between the MAC effect sizes for group A (e.g., White) and the MAC effect sizes for group B (e.g., Black) in each decade. In the end, we have a 20-decade timeseries of Pearson correlation coefficients, with high positive correlations indicating more overlap between the representations of group A and B and low correlations indicating divergent representations of groups A and B (Table 4 and Fig. 3).[¶]

General patterns of overlap across groups. In general, there was moderate-to-high overlap between the representations of any two groups in any given decade, with an average correlation across groups and across time of $r = 0.65$. Indeed, even the minimum correlation ($r_{min} = 0.42$ between White and Black in 1850) was still of moderate magnitude. This means that, in any given decade and for any two contrasting and even polarized groups (e.g., Rich/Poor, Black/White), most traits (or words) are associated to a similar degree with both groups.

Of course, there are traits in the language that are uniquely associated with one group over another, as reported in Analyses 1, 2, and 3. For example, the words able and competent have cosine similarities of 0.04 and 0.02, respectively, to men (but −0.02 and −0.04 to women). There are other words, however, that are similar in their degree of association: The words kind, nervous, and religious have nearly identical cosine similarities of 0.01, −0.02, and −0.02, to both men and women. These latter words, of which there are many, account for the high correlations in group overlap. Why such words are associated to similar degrees with two groups likely reflects a combination of factors. For instace, the groups are often written about in pairs (e.g., "grandma and grandpa are so generous"), and the use of negation (e.g., "men are not kind," "women are not religious") can lead to a group–word association even if the writer meant the opposite [although negation is relatively uncommon, especially in written text (35)].

Differences in overlap across groups. Beyond the general finding of moderate overlap, we also find that the degree of overlap varied by group comparison. Taken across all time, the largest overlap was seen between Men–Women ($r = 0.82$ for traits and 0.83 for words), perhaps because speakers and writers are particularly likely to use gendered pronouns and role labels in pairs (e.g., "one for him and one for her"). In contrast, the lowest overlaps were observed for White and Black ($r = 0.54$ for traits and 0.53 for words) and for White and Irish ($r = 0.54$

[¶]All cross-race comparisons (e.g., Black–Asian, Black–Native, Asian–Native, and so on) are reported in *SI Appendix*.

Table 3. Within-group stability in trait representations across decades

Group A (vs. Group B)	10 y (1800–1810)	50 y (1800–1850)	100 y (1800–1900)	150 y (1800–1950)	200 y (1800–1990)	Mean	Min	Max
White (vs. Black)	0.84	0.77	0.69	0.57	0.41	0.72	0.41	0.86
Black (vs. White)	0.74	0.67	0.61	0.53	0.46	0.65	0.37	0.86
Asian (vs. White)	0.80	0.77	0.66	0.40	0.41	0.59	0.27	0.82
Irish (vs. White)	0.69	0.54	0.43	0.37	0.38	0.66	0.30	0.88
Hispanic (vs. White)	0.83	0.65	0.47	0.36	0.31	0.54	0.17	0.83
Native American (vs. White)	0.67	0.55	0.55	0.51	0.26	0.60	0.18	0.79
Men (vs. Women)	0.89	0.87	0.79	0.67	0.58	0.85	0.58	0.97
Women (vs. Men)	0.93	0.91	0.86	0.81	0.73	0.87	0.67	0.96
Old (vs. Young)	0.81	0.76	0.66	0.57	0.55	0.74	0.50	0.89
Young (vs. Old)	0.87	0.84	0.76	0.75	0.62	0.78	0.58	0.92
Fat (vs. Thin)	0.85	0.85	0.80	0.70	0.61	0.79	0.49	0.92
Thin (vs. Fat)	0.78	0.75	0.76	0.66	0.58	0.74	0.45	0.90
Rich (vs. Poor)	0.92	0.87	0.82	0.74	0.66	0.84	0.61	0.94
Poor (vs. Rich)	0.87	0.83	0.83	0.76	0.68	0.83	0.65	0.93

Correlations in the first five columns represent Pearson's *r* correlations between the mean average cosine similarity (MAC) scores for trait (or all word) associated with a given group (e.g., White) in decade 1 (e.g., 1800) and decade 2 (e.g., 1810). The final three columns represent the mean, minimum, and maximum across all pairwise Pearson's *r* correlations for MAC scores across decades (i.e., across all 190 pairs of decades, or all correlations visualized in Fig. 2) for a given group (e.g., White). For the table reporting results from word analyses, see *SI Appendix*; results closely replicate for trait and word analyses.

for traits and 0.58 for words). This last result of low overlap between White and Irish is perhaps surprising given that these two groups share an overlapping group membership of "Whiteness," at least in contemporary notions. However, such constructions of Whiteness have only emerged over time (36), a finding reflected in the increasing correlation (see below).

Changes in overlap across groups. Finally, although most cross-group overlaps began at a medium-to-large magnitude, the overlaps most often increased in magnitude across time (except for Old/Young, and White/Native American on trait representations). Increasing overlap may reflect a change from historical texts that emphasized group differences to contemporary texts that emphasized group similarities (e.g., as in the "colorblindness" ideology) (37). This contemporary emphasis on similarity may be especially prominent for groups that become "assimilated" following waves of immigration (indeed, the largest linear increases for racial groups were between White/Asian and White/Hispanic, two of the largest immigrant groups in the White, English-speaking world).

Discussion

Bridging the long history of psychological studies on stereotype change (3–6) with the new frontiers from NLP methods in historical word embeddings (18), the current manuscript examines patterns of change and stability in social group representations across 200 y of English-language book text. Using pretrained embeddings from over 850 billion words, we empirically identified the top words and traits (from lists of over 14,000 words and a subset of 600 traits) that were most associated with 14 diverse social groups across historical writings.

Convergence with Past Work on the Content of Social Group Representations. In our first analysis, we found that the emergent content of social group representations, whether collapsing across all time or separating by decade, often converged with the expected content from past research on social group stereotypes. For instance, groups characterized by high competence in contemporary studies (e.g., Men, White) were also associated with positive competent-related words in historical text. In contrast, groups typically characterized by low competence (e.g., Women, Poor people; ref. 33) were also associated with low-competence content in language. Convergence across contemporary data and the historical data of the current analysis has two implications for the nature of social group representations and the methods we use to uncover them.

First, convergence across methods can lend greater confidence in the remaining discoveries of the manuscript. For instance, the discoveries regarding which dimensions were emphasized in language remains worthy of future research. Why, for example, did the representation of Hispanic appear to focus on uprisings and unrest over an alternative representation (e.g., cultural connotations)? Ultimately, the convergence suggests we are on solid ground when interpreting these more novel discoveries about the qualitative focus of word representations.

Second, convergence across methods suggests that the representations being assessed, on the one hand, by homogeneous samples of Princeton students using relatively small samples of selected traits and, on the other hand, by massive studies of English-language books across 200 y using entirely bottom-up approaches, are tapping into the same widespread "collective representations" (38). Such representations of groups (e.g., that Men are competent or that Poor individuals are helpless) appear to have infiltrated

Table 4. Overlapping trait representations across two groups over time

Group A–Group B	Overall	1800	1850	1900	1950	1990
White–Black	0.54	0.60	0.50	0.42	0.57	0.73
White–Asian	0.60	0.45	0.51	0.62	0.70	0.74
White–Irish	0.57	0.61	0.51	0.54	0.63	0.74
White–Hispanic	0.62	0.55	0.56	0.64	0.69	0.72
White–Native American	0.54	0.66	0.58	0.52	0.48	0.64
Men–Women	0.82	0.74	0.74	0.84	0.90	0.94
Old–Young	0.70	0.69	0.75	0.59	0.71	0.72
Fat–Thin	0.77	0.72	0.67	0.84	0.84	0.81
Rich–Poor	0.61	0.56	0.50	0.63	0.72	0.71

Correlations represent Pearson's *r* correlations between the MAC scores for trait (or all word) associated with group A (e.g., White) and group B (e.g., Black) in sample decades (1800, 1850, 1900, 1950, and 1990) or collapsing across all time (column 1). For the table reporting results from word analyses, see *SI Appendix*; results closely replicate for trait and word analyses.

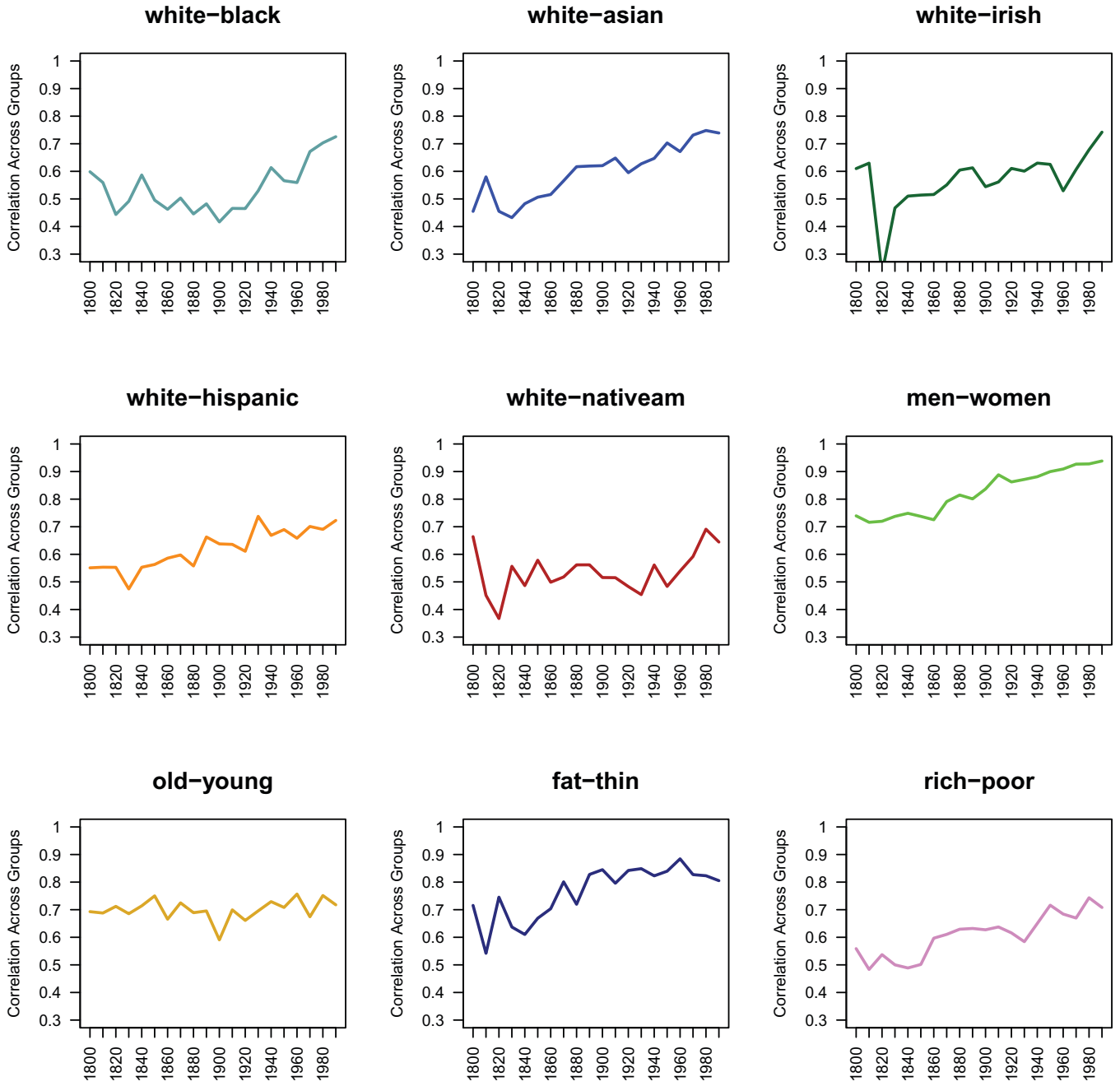

Fig. 3. Cross-group correlations of trait representations for two groups over 20 decades (1800 to 1990) of historical English-language text. Higher correlations indicate greater overlap across the two comparison groups listed above each correlation.

society to such a degree that every perceiver or author in society is exposed to and expressing the content of such stereotypes (39).

Discoveries about Change and Stability in Social Group Representations. Taken together, the current data paint a nuanced picture of how historical social group representations have both changed and persisted over 200 y of historical book text. From a birds-eye view, we highlight five major conclusions, each elaborated with their implications below. First, the top 10 (or 50) word and trait associates with a given social group were generally found to change in semantic content (i.e., at least 50% of traits or words shifted in content across decades). Second, the valence of these top word and trait associates remained generally stable for most groups. Third, the larger space of all trait and word associates also remained significantly correlated across time for most groups. Fourth, when there was evidence of change in the larger space of associates, it was greater for racial and nationality groups than nonracial groups. Fifth and finally, the overlap in representations across groups (e.g., Men/Women, Old/Young) generally increased toward more similar representations over time.

Top Words and Traits Change in Content, But Valence Often Remains Stable. Whether social group representations can be said to change or remain stable falls along two distinctions: 1) change in valence versus semantic content and 2) change in the top associates versus in the full space of associates. With respect to the first distinction, we find that, although the top words and traits themselves have shifted over time, the average valence of these representations has been generally stable (9 out of 14 groups showed no significant slopes of valence change). Again, we note that stability in valence can occur alongside shifts in the content of top word associates when new but similarly

negatively valenced words take the places of the old words. The valence stability is a departure from the fading negativity of racial/ethnic stereotypes reported in the Princeton quartet; in the current manuscript, only one group (Poor) moved from strongly negative to weakly negative, and even over 200 y, no group showed a reversal in negativity. Future work is poised to examine whether the distinction between the Princeton quartet and the current research is due to differences in the directness of the methods (i.e., word embeddings are indirect), the diversity of social group topics studied (i.e., racial groups may change more than nonracial groups), or the scope of the timespan (200 y may "smooth" over relatively sporadic decade-long changes).

The results also demonstrate the second distinction of change and stability. Despite changes in semantic content among the top 10 or 50 trait and word associates, the longer lists of 14,000 words and 600 traits reveal relatively more consistency in the underlying representations (see the high correlations in Analysis 3). Thus, top words were shifting over time, but the degree of associations with the collection of most other words was relatively more similar across 200 y. This means that, when representing a social group in text, the group will most often be associated with a wealth of words and traits that consistently characterizes its image in our society. The impact of history, however, can still be seen in altering which of these 14,000 words or 600 traits are pulled out as the strongest and most unique features of that group in a particular moment.

Group Differences in Change and Stability: Racial versus Nonracial Group Stereotypes. The current breadth of analysis—examining representations across 14 social groups—enabled a test of whether there were certain groups that revealed more or less change than others. Indeed, in the full space of trait and word associates, racial and nationality groups showed relatively greater change than nonracial groups. Such results align with contemporary data from human participants (26) and theoretical perspectives that highlight a contrast between so-called arbitrary racial groups and evolutionarily relevant groups of age, class, or gender (24). The current paper provides new empirical evidence for these theoretical distinctions traced back across an extensive scope of history.

Increases in the Overlap across Social Groups Representations. Finally, the current methods offered an opportunity to expand beyond the analysis of a single group representation and consider how the representations of two opposing groups (e.g., Old–Young, Rich–Poor) have transformed together or apart over time. In Analysis 4, we found that, in any given decade, two opposing groups often overlapped in at least a portion of their representations, with moderate correlations at baseline. Furthermore, these overlaps changed over time, most often in the direction of increasing the similarity between two groups' representations (however, this finding was less consistent across our replication in the COHA dataset). As speculated by Karlins (5), the current data suggest early evidence of "traces of assimilation" (p. 8) in historical written stereotypes of social groups. Future research is equipped to explain why these increasing overlaps occur, including a general push toward "colorblindness" in text (37) and/or assimilation following immigration waves (14).

Limitations and Future Work.

English-language, dominant group authors. The current work is limited by the scope of the text we study. Although Eng-all is the largest available historical corpus of continuous text, it suffers from representing only a small and possibly biased slice of society (23). The text is only from English-speaking authors and Western countries, and the authors are likely from a select class of society. The findings cannot be interpreted to depict the full complexity of social representations from all cross-sections of historical societies; rather, they should be interpreted as the stereotypes dominant in cultural conversations and shared through written book text.

Book text. Examining the pace of change in social representations through book text has unique limitations including, among others, 1) the lag time it takes for a manuscript to go from idea to print, 2) the possible (if rare) inclusion of reprinted book texts (artificially reproducing a social representation in a later decade), and 3) the description (or prediction) of events in the past (or future) rather than a depiction of the current culture (e.g., a text in 1980 that describes how people in 1800 used to view women). It is therefore possible that the current analyses could underestimate the amount of change in real opinions, beliefs, and attitudes held by people of a given society, because those changes in the minds of people take time to trickle into changes in texts. Future work, using contemporary data (where both book text and survey data are available at sufficient and overlapping temporal granularity), will be helpful in testing how social representations uncovered through NLP align or lag behind attitudes measured through typical psychological tests.

Semantic drift. Tracking change in societal group representations through words is inevitably tied up in the simultaneous changes of word meaning itself (i.e., "semantic drift," such as "gay" moving from meaning "joyful" to "homosexual") (18). In *SI Appendix*, we show that the semantic drift of a trait (or word) does not fully explain the degree to which that trait or word becomes more or less associated with a group. Additionally, by selecting social groups that have been relatively consistent in their dictionary meaning, we argue that we are more likely examining change in social representations than in semantic drift. Nevertheless, there is likely a complex and bidirectional interaction between social change and semantic drift. After all, groups described with different labels (e.g., Black vs. African American) conjure different representations in mind (25); conversely, changes in attitudes toward a group (e.g., decreases in negativity toward Black Americans) can prompt certain pejorative labels (e.g., the N-word) to fall out of favor or gain new meanings (40). Future work would benefit from deeper examinations of the relationship between social representation change and semantic drift.

Word valence ratings. The current methods are limited in using contemporary and decontextualized ratings of word valence to infer the average valence of group stereotypes from across history. Although others have shown that word favorability (i.e., valence) is generally stable across time, with rank-order correlations around 0.88 over 70 y (5, 6), it is possible that some words may have changed subtly in valence across the 200-y span we use here. Unfortunately, it is impossible to ask people from 200 y ago to provide contextualized historical ratings of words; our best approach is to use the largest set of contemporary norms of valence and offer the caution that all valence analyses come from contemporary ratings.

Single word embeddings. Finally, the work is limited in using pretrained single word embeddings; a polysemic word such as "black" has only one embedding that captures all its meanings (collapsing across black referring to a group and a color). Single word embeddings offer multiple conveniences: They have been extensively studied and validated in previous work on social biases, they can be merged valence norms, and they are relatively small in computational size and thus can be analyzed by researchers with differing levels of computational resource access. Yet single word embeddings also limit the types of social groups that can be studied. It is difficult to represent

intersectional categories or groups with two-word labels (e.g., dark skinned, Black woman, etc.), as well as to ensure that we are accurately capturing the social group in question (e.g., ensuring we are representing Black as a social group rather than a color; although see *SI Appendix* for additional tests). We are encouraged by ongoing work in computer science to generate large databases of historical contextualized embeddings that will be necessary to test the accuracy of the current conclusions (41) and advance research on social representation change.

Materials and Methods

All data and analysis scripts are provided through the Open Science Framework (https://osf.io/th89x/). Additionally, *SI Appendix* provides details on a) the logic of word embeddings, b) the data from HistWords, c) the selection of group labels, d) tables and visuals for the word-based analyses, e) replication of results with additional lists of group labels, f) replication with the full list of 100,000 word tokens, and g) replication of results with COHA.

In this project, we use pretrained embeddings made available through HistWords (18), with decade-wise word embeddings for all archived English-language books available from 1800 to 1990. Using pretrained embeddings is beneficial because they bypass the need for computationally expensive training of novel word embeddings, have been validated in previous work (42), and address methodological concerns such as realignment of vector spaces across time (18) (see *SI Appendix* for details). To originally train the word embeddings by decade, Hamilton and colleagues (18) used Google Books 5-grams (i.e., five-word sequences) and preprocessed the text by converting all words to lowercase and removing punctuation. Next, they trained word embedding models for the top 100,000 most frequent word tokens, discarding any words with less than 500 observations per decade. Separate word2vec models were then trained for each of the 20 decades from 1800 to 1990.

Data Availability. All data have been deposited in a publicly accessible database on the Open Science Framework (https://osf.io/th89x).

ACKNOWLEDGMENTS. This research was supported by the Harvard Mind Brain Behavior Inter-Faculty Initiative, the Foundations of Human Behavior, and the Hao Family Inequality in America Support Fund awarded to M.R.B. and T.E.S.C. We are grateful to Wil Cunningham, Dan Hoyer, and Yoav Rabinovich for feedback on earlier versions of this manuscript.

1. J. C. Jackson *et al.*, From text to thought: How analyzing language can advance psychological science. *Perspect. Psychol. Sci.* **17**, 805–826 (2022).
2. T. E. S. Charlesworth, M. R. Banaji, "Word embeddings reveal social group attitudes and stereotypes in large language corpora" in *Handbook of Language Analysis in Psychology*, M. Dehghani, R. L. Boyd, Eds. (Guilford Press, 2022), pp. 494–508.
3. D. Katz, K. Braly, Racial stereotypes of one hundred college students. *J. Abnorm. Soc. Psychol.* **28**, 280–290 (1933).
4. G. M. Gilbert, Stereotype persistence and change among college students. *J. Abnorm. Psychol.* **46**, 245–254 (1951).
5. M. Karlins, T. L. Coffman, G. Walters, On the fading of social stereotypes: Studies in three generations of college students. *J. Pers. Soc. Psychol.* **13**, 1–16 (1969).
6. H. B. Bergsieker, L. M. Leslie, V. S. Constantine, S. T. Fiske, Stereotyping by omission: Eliminate the negative, accentuate the positive. *J. Pers. Soc. Psychol.* **102**, 1214–1238 (2012).
7. P. G. Devine, A. J. Elliot, Are racial stereotypes really fading? The Princeton trilogy revisited. *Pers. Soc. Psychol. Bull.* **21**, 1139–1150 (1995).
8. J. W. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us* (Bloomsbury Publishing) 2013).
9. M. Dehghani, R. L. Boyd, *Handbook of Language Analysis in Psychology* (Guilford Press, 2022).
10. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. arXiv [Preprint] (2013). https://arxiv.org/abs/1310.4546 (Accessed 21 June 2019).
11. R. Tripodi, M. Warglien, S. Levis Sullam, D. Paci, "Tracing antisemitic language through diachronic embedding projections: France 1789-1914" in *Proceedings of the First International Workshop on Computational Approaches to Historical Language Change* (Association for Computational Linguistics, 2019), pp. 115–125.
12. Y. Bizzoni, S. Degaetano-Ortlieb, P. Fankhauser, E. Teich, Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Front. Artif. Intell.* **3**, 73 (2020).
13. J. J. Jones, M. R. Amin, J. Kim, S. Skiena, Stereotypical gender associations in language have decreased over time. *Sociol. Sci.* **7**, 1–35 (2020).
14. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
15. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora necessarily contain human biases. *Science* **356**, 183–186 (2016).
16. C. E. Osgood, G. J. Suci, P. H. Tannenbaum, *The Measurement of Meaning* (University of Illinois Press, 1967).
17. D. Peabody, Selecting representative trait adjectives. *J Pers Soc Psychol*, **52**, 59–71 (1987).
18. W. L. Hamilton, J. Leskovec, D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change" in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2016), vol. 1, pp. 1489–1501.
19. A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013).
20. Y. Lin *et al.*, *Syntactic Annotations for the Google Books Ngram Corpus* (Association for Computational Linguistics, 2012).
21. J.-B. Michel *et al.*, Quantitative analysis of culture using millions of digitized books. *Science.* **331**, 176–182 (2011).
22. N. Younes, U. D. Reips, Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. *PLoS One* **14**, e0213554 (2019).
23. E. A. Pechenick, C. M. Danforth, P. S. Dodds, Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One* **10**, e0137041 (2015).
24. M. Davies, *Corpus of Historical American English* (Brigham Young University, 2010).
25. E. V. Hall, S. S. M. Townsend, J. T. Carter, What's in a name? The hidden historical ideologies embedded in the Black and African American racial labels. *Psychol. Sci.* **32**, 1720–1730 (2021).
26. T. E. S. Charlesworth, M. R. Banaji, Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).
27. T. E. S. Charlesworth, M. R. Banaji, Patterns of implicit and explicit stereotypes III: Long-term change in gender stereotypes. *Soc. Psychol. Personal Sci.* **13**, 14–26 (2022).
28. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
29. J. Sidanius, F. Pratto, *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression* (Cambridge University Press, 1999).
30. T. Manzini, Y. C. Lim, Y. Tsvetkov, A. W. Black, "Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019) vol. 1, pp. 615–621.
31. A. Bellovary, A. D. Armenta, C. Reyna, "Stereotypes of immigrants and immigration in the United States" in *Stereotypes: The Incidence and Impacts of Bias*, J. T. Nadler, E. C. Voyles, Eds. (ABC-CLIO LLC, 2020), pp. 146–164.
32. T. Ellingson, *The Myth of the Noble Savage* (University of California Press, 2001).
33. S. T. Fiske, A. J. C. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* **82**, 878–902 (2002).
34. T. D. Nelson, *Ageism: Stereotyping and Prejudice against Older Persons* (MIT Press, 2002).
35. G. Tottie, *Negation in English Speech and Writing: A Study in Variation (Quantitative Analyses of Linguistic Structure)* (Academic Press, 1991).
36. K. A. Appiah, *The Lies That Bind: Rethinking Identity* (Liveright, 2019).
37. E. Bonilla-Silva, *Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America* (ProQuest Ebook Central, 2017).
38. E. Durkheim, *Sociologie et Philosophie* (Felix Alcan, 1924).
39. T. E. S. 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).
40. J. Rahman, The N word: Its history and use in the African American community. *J. Eng. Linguist.* **40**, 137–171 (2012).
41. V. Hofmann, J. Pierrehumbert, H. Schütze, "Dynamic contextualized word embeddings" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Association for Computational Linguistics, 2021), pp. 6970–6984.
42. A. Toney, A. Caliskan, ValNorm: A new word embedding intrinsic evaluation method reveals valence biases are consistent across languages and over decades. arXiv [Preprint] (2020). https://arxiv.org/abs/2006.03950 (Accessed 12 June 2020).