Identifying and predicting stereotype change in large language corpora: 72 groups, 115 years (1900-2015), and 4 text sources

Tessa E.S. Charlesworth[1], Nishanth Sanjeev[2], Mark L. Hatzenbuehler[1], Mahzarin R. Banaji[1]

[1]Harvard University, Cambridge, Massachusetts
[2]New York University, New York, NY

Author Note

Tessa E.S. Charlesworth, Mark L. Hatzenbuehler, and Mahzarin R. Banaji, Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138; Nishanth Sanjeev, Department of Computer Science, New York University, 251 Mercer St, New York, NY 10012.
The R code and data included in this article are available at the Open Science Framework: https://osf.io/gzuy4/.
This research was supported by the Hao Family Inequality in America Support Grant, the Foundations of Human Behavior Initiative, and the Mind Brain Behavior Interfaculty Initiative awarded to Mahzarin R. Banaji and Tessa Charlesworth.
Disclaimer: this manuscript contains references to offensive slurs used to refer to groups in the past and present.
Correspondence concerning this article should be addressed to Tessa Charlesworth, Department of Psychology, Harvard University, Cambridge, MA 02138, email: tessa_charlesworth@fas.harvard.edu.

Abstract

The social world is carved into a complex variety of groups each associated with unique stereotypes that persist and shift over time. Innovations in natural language processing (word embeddings) enabled this comprehensive study on variability and correlates of change/stability in both manifest and latent stereotypes for 72 diverse groups tracked across 115 years of 4 English-language text corpora. Results showed, first, that group stereotypes changed by a moderate-to-large degree in manifest content (i.e., top traits associated with groups) but remained relatively more stable in latent structure (i.e., average cosine similarity of top traits' embeddings and vectors of valence, warmth, or competence). This dissociation suggests new insights into how stereotypes and their consequences may endure despite documented changes in other aspects of group representations. Second, results showed substantial variability of change/stability across the 72 groups, with some groups revealing large shifts in manifest and latent content, but others showing near-stability. Third, groups also varied in how consistently they were stereotyped across texts, with some groups showing divergent content, but others showing near-identical representations. Fourth, this variability in change/stability across groups was predicted from a combination of linguistic (e.g., frequency of mentioning the group; consistency of group stereotypes across texts) and social (e.g., the type of group) correlates. Groups that were more frequently mentioned in text changed more than those rarely mentioned; sociodemographic groups changed more than other group types (e.g., body-related stigmas, mental illnesses, occupations), providing the first quantitative evidence of specific group features that may support historical stereotype change.

Keywords: social groups, stereotypes, historical change, text analysis, word embeddings

Significance Statement

How have group stereotypes changed or persisted over history? How does change/stability differ between the *manifest* (traits associated with groups) versus *latent* structure of stereotypes (average valence, warmth, competence)? How does change/stability vary across a more diverse sample space of groups than ever previously investigated? Leveraging natural language processing applied to 4 English-language text corpora we track stereotype change/stability towards 72 groups across 115 years. Results showed that: (1) although manifest stereotype content shifted over time, latent representations had more enduring stability; (2) groups varied widely in change/stability and in their consistency of representation across texts; and (3) such variation in change/stability was generally predictable through linguistic and social features of groups. The increasing availability of massive text, coupled with new methods to interrogate them, will advance understanding of whether and when (along which metrics, and for which groups) change may be possible in group stereotypes across history.

Introduction

Language provides a record of how humans think and feel about the various social groups that make up their worlds. How a group is stereotyped in language, even in a single moment and from a single societal discourse (e.g., Internet text), can reveal both qualitative and quantitative insights into the stereotypes that are shared and communicated in the real world (for reviews, see Charlesworth & Banaji, 2022; Jackson et al., 2022). Yet perhaps what is most unique about investigating group stereotypes through language is the way it can expand horizons in at least four directions. First, across dozens of groups that vary in meaningful and socially-relevant ways, from sociodemographic groups (e.g., *White, Black, Religious, Atheist*), to body-related and physical groups (e.g., *Abled, Disabled, Short, Tall*), to mental health-related groups (e.g., *Schizophrenic, Autistic, Depressed*), to occupational groups (*Manager, Server, Employed, Unemployed*) and more. Second, across decades or even centuries of human history, reflected through archives of historical text sources. Third, across multiple societal discourses, that range from more controlled and edited (books) to more spontaneous (Internet) media and communications. And fourth, across multiple metrics of stereotypic representations, including both the manifest structure of stereotypes (i.e., the actual traits associated with groups) and the latent structure of stereotype meaning (e.g., the average ratings of stereotypes along latent positivity/negativity, warmth/coldness, or competence/incompetence). It is this unique expansion of understanding stereotypes across distinct groups, time, texts, and metrics that we pursue in the current research. The result is the most comprehensive portrait to-date of group stereotype change, ultimately yielding new insights into whether, how, and to what extent, stereotypes unfold in naturalistic historical language.

Natural language processing as a tool to study stereotype change

The starting point for this investigation is the methodological innovation of word embeddings (for more detailed, formal explanations see Mikolov et al., 2018; Mikolov, Chen, et al., 2013; Pennington et al., 2014). At a high level, word embeddings can be understood by first assuming that all word meaning is represented in a large "cloud" of meaning, with each word in the language embedded in this cloud using a long vector of coordinates. To create such vectors (i.e., the word embeddings), we: (1) take a massive text as input (e.g., thousands to billions of conversations, books, or words from the Internet); (2) compute all word co-occurrences across contexts in the text (e.g., the number of co-occurrences of words *young* and *healthy, young* and *strong, young* and *decrepit* and so on); and (3) reduce the dimensionality of these co-occurrences to obtain a single compressed vector that positions each word in relation to all other words in the text input.

For instance, a final set of word embeddings will represent the meaning of *young* in a single long vector (generally about 300 numbers long, positioning that vector in a 300-dimensional coordinate space), as well as the meaning of all other words such as *old, healthy, unhealthy, strong, weak,* and so on, each with their own 300-length vectors. If the word embeddings successfully represent semantic concepts, we would expect that, stereotypically, the vector for *young* would be closer to vectors for words such as *healthy* or *strong,* whereas the vector for *old* would be closer to vectors for words such as *unhealthy* or *weak* (Caliskan et al., 2016; Charlesworth et al., 2022a). In this way, one can use word embeddings to study social group representations by comparing the relative closeness between words (formally, the relative differences in cosine similarities between word vectors; Caliskan et al., 2016) referring to different groups (e.g., *old, young*) and attributes (e.g., *healthy, unhealthy*).

While early research using word embeddings to study group stereotypes focused on *static* representations drawn from text at one time point (for a review see Charlesworth & Banaji, 2022), interest has recently turned to using *diachronic* embeddings, or embeddings trained from text corpora across successive time steps, to investigate changes in stereotypes. For instance, research has tracked changes in gender stereotypes (and, specifically, the manifest content of stereotypes) across book texts from the 1900s (Bhatia & Bhatia, 2021; J. J. Jones et al., 2020). Despite its utility, such research focusing exclusively on the stereotypes of gender groups, or any other group in isolation, cannot capture the vast diversity of group stereotypes in the real world. This is especially relevant because, as we discuss below, groups vary in conceptually meaningful ways that may affect their likelihood or rate of change.

Indeed, recognizing the limitations of focusing on groups in isolation, two previous studies sought to expand the study of stereotype change in language across multiple group targets. First, Garg and colleagues (2018) examined the manifest stereotypes for 8 ethnic and gender groups tracked through news media and books across 1900-2000. The authors found that changes in the manifest trait content associated with women aligned with the height of the women's movement in the 1960s and 70s; similarly, changes in the traits associated with Asians aligned with waves of immigration (Garg et al., 2018). Such results were pivotal for validating the use of diachronic embeddings in uncovering changes in multiple group stereotypes.

Garg and colleagues' findings inspired a subsequent investigation of group stereotypes for a larger set of 14 sociodemographic groups (expanding beyond gender, race, and ethnicity, to also capture nationality, age, class, and body weight) tracked across a longer time span of 200 years of English-language book text (Charlesworth et al., 2022a). The results showed nuanced patterns of both change and stability that varied across the 14 groups, with some groups (e.g.,

gender) reflecting relatively greater stability than others (e.g., race, ethnicity, and nationality). Additionally, results hinted at differences in change between relatively manifest stereotype content versus deeper-level latent stereotype valence (i.e., the average positivity/negativity of traits associated with groups). Suggestions of such variability across the handful of 14 groups and two metrics demanded a more comprehensive study of more groups, texts, and metrics of stereotypes to yield a sample that could more directly *quantify* the scope, variability, and predictors of stereotype change in historical language.

Expanding across the wider landscape of social groups

The current work expands methodologically and conceptually to include the longest list of group targets studied to-date: we include 72 diverse groups, with not only many more sociodemographic identities than past research (e.g., gender, race, ethnicity, age, religion, citizenship, and so on) but also other stigmatizing characteristics such as those related to the body (e.g., ability, weight), mental health and illness (e.g., depression, schizophrenia), and occupational status (e.g., laborer, unemployment). This diversity of groups is necessary to more accurately capture the real-world variation in our social fabric (Fiske et al., 2002; Pachankis et al., 2018).

After all, social groupings emerge wherever demarcations can be perceived to separate a set of people according to some shared circumstance, characteristic, status, or identity. Membership in groups can thus vary in whether it is concealed or visible, persistent or transient, and originating from birth or acquired over time (E. E. Jones, 1984). Some groups can be extremely large in the number of members (e.g., women, men, old, young) whereas others are relatively small (e.g., groups with specific disabilities, specific occupational groups). Some groups are perceived to be threats to one’s physical safety, economic viability, or values

(Stephan et al., 2009) whereas other groups carry stigmas that are viewed to be disruptive to daily interactions and aesthetically unappealing (Goffman, 1963; E. Jones et al., 1984; Pachankis et al., 2018). Some groups have been argued to be evolutionarily "old," such as age and gender, meaning that they are found to be present and meaningful group markers in all societies and have existed as groups from the earliest formation of social structure. Other groups are relatively "new," such as race and ethnicity, which emerged later in evolutionary history and are more variable across place and culture (Fiske, 2017; Kurzban et al., 2001; Sidanius & Pratto, 1999). Such wide variation in the types of groups yields similarly wide variation in how groups are *stereotyped* in society, as well as unique patterns of how those stereotypes might change or persist over time. And yet, given past methodological limitations, no comprehensive study has ever examined historical change/stability in text-based stereotypes across a more representative and varied sample space of groups.

Of note, such a comprehensive study is important not only for painting a descriptive portrait of variability in stereotype change, but also for providing the necessary sample size to conduct the first statistical tests that quantify how theoretically relevant correlates might help explain such variability. As we elaborate in the *Methods* below, we explore the prediction of stereotype change from two classes of variables. First, we examine how the *type* of group may help explain variation in patterns of change/stability. We compare the degree of change across four broad clusters encompassing: (1) sociodemographics (e.g., racial, religious, sexuality groups such as *Black, White, Christian, Muslim, Gay, Straight*); (2) mental illnesses and health (e.g., *Autistic, Schizophrenic*); (3) occupational statuses (e.g., *Unemployed, Laborer*); and (4) body-related identities (e.g., *Disabled, Fat, Thin*).

There are many theoretically relevant differences across these four types of groups (E. E. Jones, 1984; Pachankis et al., 2017). As but one example: relative to other sociodemographic stigmas (e.g., race, gender, sexuality), body-related stigmas are more widely and explicitly stereotyped in society. Indeed, body-related stigmas (such as anti-fat prejudices) show higher acceptability than other stigmas (Crandall et al., 2002), are among the slowest changing attitudes in contemporary surveys (Charlesworth & Banaji, 2022a), and continue to be negatively represented in media (Greenberg et al., 2003). Other sociodemographic stigmas, by contrast, are perceived as unacceptable in society (Crandall et al., 2002) and have now even reached a point of neutrality in contemporary explicit attitudes, such that, on average, respondents no longer express explicit preferences about race or sexuality (Charlesworth & Banaji, 2022a).

If patterns of *historical* change/stability from 100 years of text can also be carved along the joints of such group clusters, we would have evidence for the relevance of contemporary understandings of groups for predicting historical change. In contrast, if all group targets reveal similar, parallel change/stability, that might suggest that historical English text reflects a general, target-agnostic shift in how social groups are stereotyped. For instance, the expression of stereotypes in text, towards all group types, may have become increasingly rare as norms against prejudice and harm expanded (Haslam, 2016). A finding of similar patterns across group types would also emphasize the need to consider the multitude of other features that characterize differences across groups, above and beyond how those groups are clustered into broad sets.

In this vein, we also consider a second set of correlates that provide more granular insights into possible differences across groups. Specifically, we consider whether *linguistic features* of how the groups are represented and referred to in language can help explain observed variability in group stereotype change. These linguistic features include the average (1)

polysemy (i.e., multiple meanings), (2) frequency, and (3) semantic drift (i.e., change in dictionary definition) of a group's labels (i.e., the terms used to refer to a group target). Polysemy and frequency have been shown to relate to linguistic change broadly, such that words that are more polysemous and less frequent changed more in their semantic definition over time (Hamilton et al., 2016). It remains an open question, however, whether variability across groups' *stereotype* change (i.e., how a group concept shifts in its relationship to traits) might also be parsimoniously explained by basic linguistic features such as the frequency or polysemy of a group's labels.

Finally, we examine a fourth linguistic variable – the consistency/inconsistency in how a group is represented across diverse text corpora (books, newspapers, Internet text), in its manifest content, as well as latent valence, warmth, and competence. This variable, which we term *corpus-inconsistency*, is discussed in more detail in the following section.

Expanding the number of language corpora for analysis

In addition to expanding in scope across groups and their correlates, the current work also steps beyond previous investigations to consider variability of group representations across multiple text sources – sources that vary in size, format, intended audiences, and other methodological and substantive factors. At the time of the current study, the most widely-used sets of pretrained embeddings for social scientific inquiry encompassed embeddings trained on (1) *Google Books*, a corpus of millions of English-language fiction and non-fiction texts available across 200 years; (2) the *Corpus of Historical American English,* a smaller curated and genre-balanced corpus of English-language books across 200 years; (3) the *New York Times Annotated Corpus,* a more contemporary yearly corpus of *New York Times* articles since 1990;

and (4) although not diachronic or time-stamped, the most widely-used of any corpus, the *Common Crawl,* a massive corpus argued to reflect all Internet text from the 2000s and before.

Even just these four text sources have inherent differences. As but one example, texts may differ in the degree to which they report the "fact" of events as they unfold (newspapers) versus offer more opinions and interpretations of social life (edited books). Here, we propose that, just as social scientists using survey methodologies might compare the presence and potency of stereotypes across slices of society (e.g., across demographic groups; Charlesworth & Banaji, 2021), so too might it be informative for text-based analyses to incorporate and compare the stereotypes revealed from multiple corpora that reflect unique discourses.

To this end, in addition to including multiple corpora to test robustness, we also directly quantify the degree of *corpus-inconsistency* in manifest stereotype structure (i.e., differences across corpora in the top traits) and latent valence, warmth, and competence. Of note, corpus-inconsistency has limitations for interpretation: the chosen set of pretrained embeddings vary not only in substantive ways (e.g., through the content of the text and the intentions of the text authors) but also in methodological ways (e.g., the preprocessing and training decisions of the embedding creators). Therefore, the *source* of corpus-inconsistency as methodological and/or substantive cannot be conclusively identified in the current work using the current text sources. Nevertheless, we argue that the *extent* of corpus-inconsistency in group representations can still be taken as an initial index of social variability versus social consensus (i.e., when there is low versus high agreement in the endorsement or expression of stereotypes; Gardner et al., 1973). In this way, corpus-inconsistency likely has relevance to understanding change/stability. For instance, within many theories of social change (Moscovici, 1976), variability in opinions is an essential precursor to change, since differences in narratives are the best means for disrupting the

majority opinion (Gardikiotis, 2011; Prislin & Crano, 2012). Thus, although admittedly our most exploratory analyses, the introduction of corpus-inconsistency presents, to our knowledge, the first attempt to quantify variation in stereotypes across large-scale text corpora and explore its relevance for historical stereotype change.

Expanding the metrics of stereotype change

So far, the few studies examining stereotype change in text have focused on one or two metrics to quantify change (e.g., the overlap of top trait associates at time *t* and *t+1;* Charlesworth et al., 2022b; Garg et al., 2018). And yet, stereotype change can be operationalized in terms of both *manifest* content (i.e., the actual top ten traits associated with a group) and *latent* structure along multiple axes of meaning (i.e., the average cosine similarity between those top ten traits and vectors of positivity/negativity, warmth/coldness, or competence/incompetence). Because change in one metric (e.g., manifest content) need not imply change in a second metric (e.g., latent valence; Bergsieker et al., 2012) it is necessary to expand investigations and compare across metrics that more readily capture the various ways in which stereotypes may both change and persist.

Here, we contribute improved and diversified metrics for operationalizing stereotype change in text at both the manifest and latent levels. As elaborated below, we begin by improving methods for studying manifest changes in the top trait stereotypes by looking at changes in the associations between the embeddings of traits at time *t* and *t+1*. Said another way, we examine how the complex (distributed) meanings of traits associated with a group at time *t* are related to the distributed meanings of traits associated with that same group at time *t+1*. Although this new metric of stereotype semantic change makes use of the distributed meanings of traits (i.e., associations between trait embeddings), it can still be interpreted as reflecting what

we call *manifest* changes in stereotypes. That is, because each trait has complex and varied meanings, even a change between relatively similar traits (e.g., from "kind" to "thoughtful" or from "lazy" to "helpless") will still be computed as a certain degree of change in the manifest content of stereotypes.

As such, we also address the question of whether such manifest changes are reflective of deeper changes along more *latent* or reduced subdimensions of the traits' meaning – the average valence, warmth, and competence of a group's stereotype at time *t*. To illustrate: a shift from "lazy" to "helpless", would be counted as a change in manifest stereotype content but would not change the latent average valence (both traits are similarly negative), warmth (both traits are similarly cold), or competence (both traits are similarly incompetent).

Although some work using NLP (Charlesworth, Caliskan, et al., 2022) and more traditional survey methods (Bergsieker et al., 2012; Devine & Elliot, 1995) have previously suggested a dissociation between manifest semantic content and latent valence, no comprehensive study has yet considered stereotype change along *multiple* latent subdimensions, including both valence and semantics (warmth and competence), as we do here. If we find that change is greater in manifest content than in *any* latent subdimension, then that would suggest that change in the distributed meaning of traits may be more complex than any one reduced dimension. For instance, change in the manifest content for a particular group (e.g., *Old*) could reflect a movement from referring more to "tradition" (vs. young people's naivete and progressiveness) to referring more to "control" (vs. young people's impulsiveness and activity) – a change in distributed manifest semantic content that is unique to the age stereotype and not easily captured by reducing along axes of warmth, competence, or valence alone. Practically, by adding multiple metrics in this way, the results also shed new light on whether and how stereotypes and their

consequences may endure, even if some metrics suggest change. We return to this implication in the general discussion.

The current project

Ultimately, the recent availability of large and diverse records of text across hundreds of years, coupled with advances in NLP methods to systematically quantify the content of such records, allows new examinations of group stereotypes: (1) using the largest number of groups to-date, including 72 groups that encompass a diversity of physical and mental qualities, occupational status, and sociodemographic identities; (2) contrasted across diverse corpora that span 115 years of published English-language text with both methodological and substantive variation; and (2) operationalized through multiple metrics of change, including those that reflect more manifest content versus latent structure. Using this unique combination of varied groups, texts, metrics, and time periods, we advance the methods and conceptual insights into the variability and predictors of historical stereotype change.

Results answer three guiding questions: (1) How, and to what extent, have group stereotypes changed, on average, in both manifest content and latent structure of valence, warmth, and competence? (2) How, and to what extent, do the 72 groups *vary* in their patterns of change across such metrics of stereotype structure? (3) What are the *correlates* of variability across groups in patterns of change? Specifically, do some types of groups (e.g., sociodemographic groups versus physical and mental qualities) change more than others over time? And are there other features of groups – polysemy, semantic drift, frequency, and corpus-inconsistency of how we refer to and represent groups – that help explain which groups change or remain stable across historical texts?

Methods

Transparency and Openness

In the following subsections we report how we determined the sample of text corpora, the sample of group stereotypes, any data exclusions, and all analyses including supplemental and exploratory analyses. All data and analysis code are available at OSF (https://osf.io/gzuy4/). Data were analyzed using *R*, version 4.2.2 (R Core Team, 2022) with all packages listed and cited in the R scripts available at OSF. The study was not pre-registered. As the data and analyses reported below constitute secondary analyses of archival data, the study was exempt from ethics review.

Analysis Procedure

Overview. The analysis procedure is summarized in six steps, each elaborated in greater detail below. In the first step we selected diverse text corpora with pretrained word embeddings. Second, we chose a large sample of groups and represented each group using a set of synonyms. Third, we computed the associations between groups and a list of over 600 traits (from Peabody, 1987) to identify the bottom-up manifest content of group stereotypes in text (i.e., the traits associated with each group). Fourth, we transformed this manifest content of group stereotypes into scores on latent subdimensions by examining the top trait stereotypes' average cosine similarity to vectors representing meaning on (a) valence, (b) warmth, and (c) competence. Fifth, we computed *change* in these manifest and latent representations across time within each corpus as well as averaged across all corpora. Sixth and finally, we explored possible explanatory variables of change in group stereotypes including the role of different *types* of groups (e.g., sociodemographic vs. body-related groups), linguistic features of groups (frequency, polysemy, semantic drift), and the variability in representations across corpora (corpus-inconsistency).

In the supplementary materials (SM) we also elaborate on detailed procedures including: (1) preprocessing and training information from the pretrained embeddings; (2) methodologies for selecting the set of 72 groups; and (3) methodologies for generating labels to represent groups. Additionally, in the *SM* we compute various exploratory analyses to illustrate the potential breadth of new directions spurred by the methods and variety of group targets. Additional analyses include, among others: (1) alternative operationalizations of latent structure using the relative norm distance (RND) from Garg and colleagues (2018); (2) analyses using raw values (rather than absolute values) of change in latent valence, to illustrate differences in the direction of change; (3) changes in bottom-up discovered clusters of groups across time; and (4) time-lagged relationships between corpus-inconsistency and manifest semantic change.

Step One: Select the Text Corpora. Group stereotypes were extracted from four sets of pretrained embeddings – *Google Books (Books), Corpus of Historical American English (COHA), New York Times (NYT),* and *Common Crawl (CC)* (Table 1). The chosen embeddings were the most widely used and validated sets of pretrained diachronic (i.e., time-stamped) and contemporary embeddings available at the time of research. The corpora were chosen not only to vary across time but also in format (e.g., relatively edited, and controlled books versus more spontaneous Internet text) and embedding algorithm (e.g., *GloVe, word2vec, PPMI*). Such diversity of texts captures not only robustness of results but also potentially informative variability in which groups are (in)consistently represented across corpora.

Table 1.
Details on text corpora used for analyses

	Estimated size (all word occurrences)	Estimated vocabulary (unique words)	Timespan used	Training method and embedding dimension
Google Books (Books)	850 billion	41,000-71,000	1900-1999	*word2vec*, 300-dimensions
Corpus of Historical American English (COHA)	410 million	11,600-15,100	1900-1999	*word2vec*, 300-dimensions
New York Times (NYT)	62 million	20,936	1990-2015	*PPMI,*

Common Crawl (CC)	42 billion	1.9 million	2014	100-dimensions *GloVe,* 300-dimensions

Google Books English-All (Books) Embeddings. The *Google Books* English-All dataset (hereafter referred to simply as *Books*) is taken from the *Google Books n-grams dataset* (second version; Lin et al., 2012), with approximately 850 billion words of all English books archived over 200 years from 1800-1999. Although not all books are included in the *Books* dataset, the coverage is estimated to be approximately 4-6% of all books ever published from 1800-1999 (Michel et al., 2011), providing a wide and diverse coverage of book-based text. We used pretrained word embeddings from the *Books* data provided by Hamilton and colleagues (Hamilton et al., 2016b), which were 300-dimensional embeddings trained using the *word2vec* algorithm (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013); see *SM* for additional model and preprocessing specifications. Only embeddings after 1900 were used because many of the groups examined in this manuscript were in common reference only from the 1900s onwards (e.g., *Schizophrenia, Gay*).

Corpus of Historical American English (COHA) Embeddings. A primary concern raised against the *Books* dataset is that, despite its massive size, it is not *genre-balanced* across time – that is, it varies in the proportion of fiction to non-fiction texts across decades which may confound examinations of change across time. To address this concern, we use a second historical corpus – the *Corpus of Historical American English* (*COHA;* Davies, 2010) – which is a substantially smaller set of English-language books (approximately 0.05% the size of the main *Books* corpus, with about 410 million words) but is carefully selected to ensure genre-balance and representativeness. *COHA* embeddings were also obtained from Hamilton and colleagues (Hamilton et al., 2016b) using the same specifications as the *Books* embeddings.

New York Times (NYT) Embeddings. To offer a complementary and more contemporary perspective on change, we also use pretrained embeddings from Yao and colleagues (2018) created from 99,872 *New York Times* articles published between January 1990 and July 2016, yielding 26 full years of data (using 1990-2015). Given that the average *NYT* article is approximately 620 words long, we estimate the total size of the *NYT* embeddings is derived from approximately 62 million words (making it the smallest corpus in the current paper). Yet, to our knowledge, these *NYT* embeddings are the only pretrained data with a timespan that is at once contemporary, of a relatively long duration, and of sufficient temporal granularity (i.e., by year rather than by decade).

Common Crawl (CC) Embeddings. Finally, given our interest in testing corpus inconsistency (i.e., differences across corpora from multiple formats and perspectives of society) we also included one of the largest and most widely-used corpora from a single timepoint – the pretrained embeddings from the *Common Crawl*. The CC is a large database of text pulled from across the Internet in 2014, with 42 billion words trained using the GloVe algorithm (Pennington et al, 2014) to create 300-dimensional embeddings. In this way, the corpus captures a slice of societal discourse that reflects the relatively more spontaneous or uncontrolled text produced by all users of the Internet. Additionally, the CC embeddings have an expansive vocabulary of nearly 2 million words, meaning that they capture even relatively rare words in English. Because it is from a single timepoint, the CC is only used to examine overall stereotype content and average ratings of valence, warmth, and competence of groups, and identify corpus-inconsistency.

Step Two: Choose groups and represent groups in labels. To capture a more representative sample of the true variation of social groups, we used a list of 93 stigmatized

statuses, characteristics and identities created by crowdsourcing from the general population as well as from experts in stigma research (Pachankis et al., 2018). Thirty-six of these group targets were ultimately able to be used in the current research because they could be accurately represented in single words with multiple synonyms (e.g., *Teen parent* could not be retained), and were not redundant with other identities (e.g., *Breast cancer* and *Pancreatic cancer* were collapsed; see SM for more details on the group selection process). For each of the chosen 36 groups, we generated a non-stigmatized comparison group (e.g., *Abled* vs. *Disabled; Young* vs. *Elderly;* or *Sober* vs. *Alcoholic*), resulting in a final list of 72 groups. This list provides, to our knowledge, the longest list of groups studied using word embeddings and includes a diversity of identities not routinely studied in social psychology (e.g., mental and physical health).

Nevertheless, the sample still sets an upper limit on the power to detect significant effects. Because we use summary values (i.e., the slope value of change in latent valence across timepoints), the *N* for regression analyses is 72 – or the number of groups. A sensitivity power analysis for simple correlations with this sample size suggests we are adequately powered at 80% power, with alpha = 0.05, to accurately detect significant correlations of $r = .32$, or small-to-moderate effects. For models using multiple regression, we are adequately powered at 80% power, with alpha = 0.05, to detect model effect sizes (variance explained) of $f^2 = 0.24$, or small-to-moderate variance explained. We also note that, given the diversity of analyses in the current manuscript, as well as the aim to provide a descriptive portrait across a comprehensive sample of social groups, a single power analysis is not an appropriate metric for evaluation of the quality of the sample. Indeed, although the sample size of groups is 72, we replicate all analyses across multiple metrics of stereotype change, multiple corpora, and multiple timepoints, yielding more accurate insights into the robustness and scope of conclusions.

Having established the list of target groups we next developed lists of labels to represent each group. Notably, in any research on group stereotypes, the choice of how to represent a given group – that is, the choice of *which labels* to use – will affect the resulting stereotypes of that group (e.g., Sigelman et al., 2005). In the current work, we sought to balance two goals – comprehensiveness and specificity – in the label synonyms. Comprehensiveness means that we capture all words used to refer to the target groups, at the risk of including some words that may be related to the general *word* concept but may not have a clear and specific link to the *group* concept (e.g., words such as "pink" and "purple" emerge as related to the group label "black" but only because of their links through referring to colors and not to groups). Specificity, therefore, means that we intended to use those words, and only those words, that can best represent the central and *group-specific* meaning of a representation.

Of course, this process will be imperfect and prone to researcher decisions. Thus, we caution that the current results are reflections of the chosen words used to represent each group. At the same time, the results are likely robust to small deviations in the group label lists, with previous work showing that the inclusion or exclusion of one or two group labels does not significantly affect results (Charlesworth, Caliskan, et al., 2022).

With these caveats in mind, our process for generating group labels began by using thesaurus searches from historical and contemporary thesauruses to develop comprehensive lists of all group-related words (see SM). Additionally, given the concern that some groups may have emerged or changed in their labels over time (e.g., *Schizophrenia* was not in use until 1910), we intentionally sought labels that also reflected the historical terms used to refer to the same group concepts (e.g., *dementia praecox* and *psychosis* were commonly used to refer to the same set of symptoms as would be labelled, today, as schizophrenia). Of note, readers will also see that we

occasionally elected to include some group-related slurs (e.g., for groups *Black, Gay, Disabled*), when those slurs emerged as top synonyms in the historical and contemporary thesauruses, and were frequently used across history to describe the given group (e.g., the *N*-word has a long history and frequent usage; Rahman, 2012).[1] Future research may explore how removing these slurs could alter specific groups' patterns of change. However, given our focus on the overarching patterns across groups, removing the few slurs is unlikely to substantively alter key conclusions.

In the end, each group was represented by approximately 7.21 labels ($SD_{Nlabels}$ = 3.95; see Table 2 for labels). Using *lists* of multiple words helps guard against concerns that results are shaped by the inclusion or exclusion of any one term (e.g., inclusion/exclusion of a single slur word).

Table 2.
Label synonyms used to represent 72 groups, and assigned cluster membership.

Group	Labels	Cluster[b]
Gay	homosexual homosexuals gay gays lesbian lesbians bisexual bisexuals queer queers lgbt transgender transgendered homo f*got f* transvestite tribads tribades sodomy sodomite sodomites homophile homophiles	Sociodemographic
Straight	heterosexual heterosexuals straight hetero	Sociodemographic
Old	old elderly elder elders older aged seniors grandparent grandparents grandmother grandmothers grandfather grandfathers	Sociodemographic
Young	young youngster youngsters youth youths teenager teenagers child children grandchild grandchildren granddaughter grandson granddaughters grandsons	Sociodemographic
Laborer	laborer laborers labourer labourers bluecollar craftsman craftsmen mechanic mechanics peasant farmer builder bricklayer	Occupation
Manager [a]	manager managers management whitecollar supervisor supervisors director directors executive executives	Occupation
Immigrant	immigrant immigrants migrant migrants newcomer newcomers asylum residency resident noncitizen noncitizens	Sociodemographic
Citizen [a]	citizen citizens citizenship citizenships denizen denizens inhabitant inhabitants native	Sociodemographic
Aboriginal	aboriginal aboriginals native aborigine inuit indigenous natives eskimo navajo pueblo apache sioux cherokee hopi comache algonquin shawnee pawnee lakota pima alaskan cheyenne	Sociodemographic
White [a]	white whites european europeans british english american americans caucasian caucasians englishman englishmen englishwoman englishwomen	Sociodemographic
Divorced	divorced divorces unmarried unhitched separated alimony estranged single	Sociodemographic

[1] Not all groups were chosen to be represented with slurs. As one example, we chose not to include slurs about White people (e.g., "*hokey*") because those slurs were not in common usage across history and were not the main way of referring to the group. In a crowd-sourced database of race-related slurs (*Racial Slur Database*), Black/African origin related slurs are the most frequent, while White-related slurs are relatively rare. Thus, the few chosen slurs represent words that were frequently-used, and often operated as stand-ins for referring to the group, even in a relatively neutral descriptive way (Popa-Wyatt & Wyatt, 2018).

Married	married marriage marriages spousal connubial wedded matrimony matrimonial espoused marital coupled mated	Sociodemographic
Alien	alien immigrant immigrants emmigrant emmigrants deported smuggled undocumented noncitizen noncitizens	Sociodemographic
Atheist	atheist atheists infidel infidels secular agnostic godless atheism ungodly heathen heathens	Sociodemographic
Religious	religious religion theist theists theistic faithful pious spiritual theological	Sociodemographic
Arabian	arabian arab arabs arabians oriental orientals israeli israelis palestinian palestinians iraqi iraqis syrian syrains iranian iranians muslim muslims egyptian egyptians	Sociodemographic
Schizophrenic	schizophrenic schizophrenics schizophrenia demented praecox psychotic psychotics psycho psychos mental	Mental health
Sane [a]	sane normal soundminded wellminded balanced rational reasonable levelheaded sensible lucid fairminded	Mental health
Christian [a]	christian christians fundamentalist fundamentalists catholic catholics evangelical evangelicals baptist baptists christianity protestant protestants lutheran lutherans methodist methodists	Sociodemographic
Other religion	jewish jew jews judaism gentile gentiles zionist zionists synagogue torah bethelem hebrew semitic yiddish kosher orthodox muslim islamic muslims moslem moslems arab arabs sunni sunnis shia shias islamist	Sociodemographic
Short	short midget midgets dwarf dwarfs dwarfism shorter	Body-related
Tall	tall lanky gangly towering statuesque 6ft sixfoot 7ft sevenfoot longlimbed	Body-related
Latino	hispanic hispanics latino latinos latina latinas cuban cubans mexican mexicans spanish guatemalan honduran nicaraguan panamanian argentinian colombian brazilian venezuelan caribbean	Sociodemographic
Unemployed	unemployed jobless unemployment poverty impoverished inactive welfare	Occupation
Employed	employed employees job working work onduty jobholding hired	Occupation
Muslim	muslim islamic muslims moslem moslems arab arabs sunni sunnis shia shias islamist	Sociodemographic
Autistic	autistic autistics autism asd asperger aspergers handicapped neurological mentally	Mental health
Server	server servers tradesperson tradespeople tradesman mechanic mechanics plumber plumbers waiter waitress waiters bartender bartenders cleaner cleaners maid maids	Occupation
Asian	asian asians japanese chinese korean taiwanese tibetan phillipino mongolian tibetan bangladeshi bhutanese indian nepalese pakistani burmese cambodian filipino indonesian malaysian thai	Sociodemographic
Deaf	deaf deafened hearingimpaired earless impaired handicapped disabled	Body-related
Abled [a]	abled able ability capable skilled nimble agile ablebodied nondisabled ambulant	Body-related
Bipolar	bipolar manic manics mania psychiatric depressed depressive depression psychotic psycho psychos	Mental health
Black	black african africans blacks colored coloreds negro negros n*er n*ers ni*a ni*as afro afros	Sociodemographic
Fat	fat fatty weight chubby obese plump overweight tubby stout chunky heavy heavyset hefty potbelly potbellied pudgy	Sociodemographic
Thin	skinny bony thinness slender slenderness rake slim slimness	Sociodemographic
Disabled	disabled cripple cripples crip crips disable wheelchair paralyzed paralysis crutch handicapped disability	Body-related
Uneducated	uneducated ignorant layman inexperienced illiterate illiterates unskilled untutored unknowledgeable untaught uninformed unread unlettered unschooled inerudite	Occupation
Educated	educated intelligent intelligence erudite informed learned wellread scholarly lettered enlightened	Occupation
Poor	poor beggar beggars needy wretch wretches impoverished destitute penniless unaffluent underprivileged	Sociodemographic
Rich	rich affluent wealthy moneyed wealth aristocrat aristocrats aristocracy prosperous privileged bourgeoisie bourgeois noble nobles nobility nobleman noblemen elite elites benefactor benefactors philanthropist philanthropists	Sociodemographic
Drug addict	drugaddict drugaddicts druggie druggies addict addicts addicted crackhead junkie dopehead cocaine overdose overdosed coke dope narcotic meth heroin cannabis weed marijuana	Mental health
Sober [a]	sober abstaining abstinent temperance temperate sobriety soberness teetotalism abstemiousness abstemious teetotal clean	Mental health
Infertile	infertile childless sterile infertility barren impotent unfertile infecund unbearing	Body-related
Fertile	fertile pregnant pregnancy fecund virile withchild progenitive fertility impregnation fecunditiy impregnate	Body-related
Unattractive	unattractive ugly awkward deformed hideous grotesque unappealing unbeautiful unpretty unsightly	Body-related
Attractive	attractive attractiveness beautiful beauty handsome handsomeness goodlooking appealing elegant flattering gorgeous	Body-related
Wheelchair bound	handicapped crip crips wheelchair paralysis paralyzed cripple handicap crippled disabled	Body-related
Smoker	smoker pothead smokers potheads cigarettes tobacco cigarette smoking	Mental health

Nonsmoker	nonsmoker nonsmokers exsmoker exsmokers nonsmoking cessation smokefree abstaining abstemious sober clean	Mental health
Depressed	depressed sad depression suicidal sadness gloomy hopeless unhappy	Mental health
Happy	happy cheerful joyful glad delighted joyous merry cheery contented vivacious lively	Mental health
Ret*	re* stupid dumb retard retards handicapped mental institutionalized disabled impaired mentally	Mental health
Indian	indian indians pakistani pakistanis bangladeshi bangladeshis bengali gendalis hindu hindus gujarati punjabi nepalese nepali kashmiri tibetan gujaratis punjabis nepalis kashmiris tibetans	Sociodemographic
Mute	mute dumb muted aphasia aphonia mutism broca wernicke paraphasia silent muffled	Body-related
Blind	blind visionless blindness blinded impaired handicapped disabled	Body-related
Alcoholic	alcoholic drinker alcoholism intoxicated alcohol drinking intoxication drunk drunkard	Mental health

Note. [a] Some groups serve as comparison groups for multiple identities. Specifically: *Citizen* is used as a comparison group for both *Immigrant* and *Alien*; *Christian* is used as a comparison group for both *Other religion* and *Muslim*; *White* is used as a comparison group for *Aboriginal, Arabian, Latino, Black, Asian*, and *Indian; Sane* is used as a comparison group for *Schizophrenic, Ret*, Bipolar*, and *Autistic; Manager* is used as a comparison group for both *Laborer* and *Server; Sober* is used as a comparison group for *Alcoholic* and *Drug addict;* and *Abled* is used as a comparison group for *Blind, Mute, Disabled, Wheelchair bound,* and *Deaf.* [b] The listed cluster – from the possible clusters of sociodemographic, body-related, mental health, or occupation – was assigned by the authors based on expert knowledge of groups as examined in the social sciences. As described in the Methods, some groups were also assigned using bottom-up clustering from previous research as well as from new agglomerative hierarchical clustering approaches; results from multiple clustering approaches are reported in the Supplementary Materials to assess robustness of all conclusions.

Step Three: Compute trait associations to groups. A primary advantage of studying group stereotypes through text is the opportunity to capture the rich qualitative semantic content associated with each group. We capitalize on these advantages by using the Mean Average Cosine (MAC; Charlesworth et al., 2022a; Manzini et al., 2019) that provides a flexible formula for examining *which* traits emerge, bottom-up, as top traits associated with the group (i.e., the manifest stereotype content). The MAC computation follows four steps. First, we calculate the pairwise cosine similarities between a trait (e.g., *strong*) and all labels used to represent a group (e.g., *strong*-gay, *strong*-lesbian, and so on for all labels representing the group *Gay*). Second, we average the pairwise cosine similarities to get the *strong*-*Gay* MAC score for each trait. We then repeat this process for the comparison group (e.g., to obtain a *strong*-*Straight* MAC score).

Third, because we are interested in the *unique* stereotypic associations to a given target group rather than the more general traits that may be shared across all groups, we calculate the *difference* for each trait's MAC score to group A versus B (e.g., *strong*-*Gay vs. strong*-*Straight*). We then rank traits according to how uniquely associated they are with group A versus B; that is,

we identify the words that have a very strong positive association with group A and a very strong negative association with group B. Fourth and finally, we set a threshold of the top-*N* traits that are taken to indicate the unique group representation – in the main analyses we use the top-10 traits but robustness analyses have shown similar conclusions from top-50 traits as well (Charlesworth et al., 2022a).

Step Four: Compute latent averages of valence, warmth, and competence.[2] Using MAC in this way will provide new insight into the *qualitative* manifest content of group representations. However, the approach is limited in its ability to succinctly *quantify* the representations and shed light on latent (i.e., reduced) subdimensions of stereotype meaning. As such, we also introduce three additional metrics of group representations by computing the average valence, warmth, and competence ratings taken from the top-*N* trait associates.

Consider, first, the valence computation. We begin by calculating historically-contextualized ratings of each trait on a positive-negative continuum for each decade or year. As elaborated in the *SM,* the historically-contextualized valence method follows by computing the relative association (using MAC) between a trait and words representing strong positivity/negativity in each time point, with valence scores ranging from -1 (the most negative) to +1 (the most positive). Each trait therefore gets a timeseries vector of its valence scores: for example, the trait "able" was always positively-valenced but with slight variation from a score of +0.13 in 1900 to +0.11 in 2000. In general, traits varied little in their valence ratings across

[2] Here, in the main text, we describe the methodologies to extract the subdimensions of valence, warmth, and competence in a second step from the top-N qualitative trait content. However, in the Supplemental Materials, we also elaborate on a second method that more directly computes the group representations along axes of valence, warmth, and competence using the Relative Norm Difference (RND; Garg et al., 2018). Briefly, similar in principle to a single-category IAT, the RND directly computes the association between a group (e.g., Gay) and two sets of words (e.g., positive words versus negative words), with no intervening step of computing trait content. Crucially, results are consistent and correlated across approaches, indicating robustness to methodological choices.

timepoints, with correlations of valence across 1900 and 2000 sitting around $r = .60$ to .78. Nevertheless, providing such historically contextualized valence ratings of traits means that we no longer rely on the assumption that the valence of all traits is necessarily stable.

Having calculated each trait's valence in each decade, we then compute the group's average valence score at time *t* by replacing each trait in the top-*N* list for time *t* with its corresponding valence rating and then taking the average across these *N* valence ratings. To make this concrete, let us continue with the example of studying the representation of *Gay.* Imagine that the top-5 unique words in decade *t* are identified using MAC as *[artistic, kind, sexy, friendly, bashful]* with corresponding valence in time *t* of *[0.05 (artistic), 0.04 (kind), -0.06 (sexy), 0.17 (friendly), -0.06 (bashful)]*. Taking the average across these valence ratings we get *valence* = +0.03 for the representation of *Gay* in decade *t*.

A similar approach is used to calculate the average ratings of warmth and competence in the group's representation. Here, we again begin by finding each trait's relative association (using MAC) in each time point *t* to a set of seed words capturing warmth vs. coldness or competence vs. incompetence, with scores ranging from -1 (very cold, very incompetent) to +1 (very warm, very competent). Next, the top-*N* traits in time *t* are replaced with their corresponding scores of warmth or competence and all *N* scores are averaged. In the example above of the top-5 words associated with *Gay* in time *t,* we have corresponding warmth scores of *[0.05 (artistic), 0.08 (kind), -0.01 (sexy), 0.14 (friendly), -0.05 (bashful)]*, with an average across these warmth ratings of *warmth* = +0.04. Similarly, for competence, we have corresponding competence scores of *[0.07 (artistic), 0.03 (kind), -0.05 (sexy), 0.05 (friendly), -0.05 (bashful)]*, with an average across these ratings of *competence* = +0.01. Ultimately, each of the 72 groups ends with a timeseries not only of the top-*N* qualitative traits (manifest content) in each decade

but also three timeseries reflecting the average ratings of valence, warmth, and competence (latent structure).

Step Five: Compute change in manifest content and latent structure.

Manifest content. As discussed in the *Introduction,* we offer methodological advances beyond our own and others' past work on changes in manifest stereotype content (or top trait associates) across historical texts[3]. Specifically, we introduce an analysis of change in manifest semantic content that looks at the average association (cosine similarity) between the distributed *embeddings* of the top-*N* traits in time *t* and the embeddings of the top-*N* traits in time *t+1*.[4] For example, imagine that the top-10 traits for the representation of *Straight* in 1900 were *[stern, direct, stable, upright, able, hard, strong, defensive, deep, steady]* and in 1910 were *[steady, defensive, upright, direct, stable, able, silent, stern, rigid, hard].* We would compute all pairwise cosine similarities between the traits across decades (e.g., *stern-steady, stern-defensive, stern-upright*, and so on), and then take the average. Higher average pairwise cosine similarities indicate higher similarity between the embeddings (i.e., distributed meanings) of traits at time *t* and *t+1*.

The magnitude of results can be interpreted along the same lines as an absolute value of correlation effect sizes, such that perfectly consistent trait representations across timepoints would yield an average cosine similarity of 1, and perfectly inconsistent trait representations would yield an average cosine similarity of 0. To convert this result into an interpretable metric

[3] Previously, we and others have calculated semantic change in stereotypes by counting the number of different traits in the top-*N* lists across successive decades or years (e.g., 1900-1910, 1910-1920, 1920-1930, and so on) and averaging the N different traits across all pairs of successive decades. Such an approach, however, risks missing similarity in the more distributed meanings of traits; a shift between two similar traits (e.g., *lazy* to *helpless*) would be equivalent to the degree of semantic change between two very dissimilar traits (e.g., *lazy* to *active*).

[4] Supplemental tests showed that results from this new metric of semantic content change using distributed meanings from embeddings are significantly correlated with previously used metrics of semantic change computed as simple counts of trait overlaps, implying robustness of our conclusions to methodological specifications.

of *change* rather than consistency we take the inverse (1 – consistency), such that higher scores (closer to 1) now indicate the greatest change or inconsistency in trait representations across timepoints.

Latent structure. We next examined change in the latent structure or average ratings of group representations along their valence, warmth, and competence scores. Such average scores have a natural metric range, and we can therefore compute change as the Spearman's *rho* between the timeseries of average valence scores (or warmth or competence scores) and a vector indicating the timestamp. For illustration, imagine the valence scores for the representation of *Atheist* within *Books* were *[-0.33, -0.20, -0.30, -0.17, -0.23, -0.22, -0.19, -0.22, -0.14, -0.08]* for the timeseries of 10 decades, thus yielding a Spearman's effect size with time of $rho_{valchange}$ = *.48*. In other words, the average valence of the stereotype associated with *Atheist* moderately increased across time, becoming more positive. To align with the range of [0, 1] for change in manifest content, described above, we take the absolute value, or $|rho_{valchange}|$. Results from raw *rho* scores are provided in the SM for comparison and illustrated in Figure 7. An identical process is followed using the timeseries of warmth scores and competence scores to calculate change in latent warmth and competence.

Step Six: Exploring predictors of *which* groups change or remain stable. As will be seen in the results below, the 72 group representations varied substantially in their degree of change for both manifest content and latent structure of stereotypes. While some groups (e.g., *Abled*) were generally stable across time, other groups (e.g., *Gay*) revealed large changes in manifest semantic content and latent structure. Confronted with such variability we explore correlates of change across: (1) the *types* of groups; and (2) linguistic features of polysemy, frequency, semantic drift, and corpus-inconsistency of group representations.

Predicting change from group types. The list of 72 groups can be divided into a variety of group sets defined either top-down (e.g., based on inspection of the groups and expert decisions of how they align with each other) or bottom-up (e.g., derived from human participants' ratings of the groups along dimensions of stigma). Our primary analysis focuses on the more interpretable top-down approach, with two bottom-up approaches reported in the SM. For the top-down approach, before any analyses were performed, the first and final authors inspected the list of 72 groups and classified each group into one of four clusters: (1) "body" groups (e.g., *Fat, Thin, Abled, Disabled*); (2) "mental health" groups (e.g., *Autistic, Bipolar*); (3) "occupation" groups (e.g., *Unemployed, Laborer*); and, the largest set, (4) "sociodemographic" groups (e.g., *Black, White, Christian, Old;* see Table 1 for each group's top-down cluster membership). Change in manifest content and latent valence, warmth, and competence were then predicted from the factor variable of cluster membership, with body-related groups dummy-coded as the baseline group.

Predicting change from linguistic features of group representations. A second class of predictors looks beyond social psychology theories and expectations about group types to also consider linguistic dynamics (Hamilton et al., 2016b, 2016a). The driving question here is: to what extent are the observed changes in manifest and latent stereotype structure correlated with other processes implicated in broader *linguistic* change? We begin with three variables from Hamilton and colleagues (2016), measuring each word's: (1) semantic drift, or how much a word changed in meaning (its shifting placement among neighboring words); (2) polysemy, or how much a word has multiple meanings (distinct or overlapping neighborhoods of words); and/or (3) frequency, or how frequently a group's labels are used in text. For each group target, we calculate an average drift, polysemy, and frequency score by averaging across the scores for all

labels used to represent the group. For example, the group *Abled* is represented by group labels including *able, capable,* and *ability,* which have corresponding semantic drift scores of 0.23, 0.43, and 0.25, averaging out to a semantic drift of 0.30, with higher scores indicating that the group labels have, on average, changed more in meaning. Average group target drift, polysemy, and frequency scores are then used as predictors in regression models predicting groups' change in semantic content and latent subdimensions.

In addition to these three available metrics, we introduce and calculate a fourth linguistic feature: corpus-inconsistency. This metric sheds light on potential differences across groups in the degree to which stereotypes are seen to be consensual and widely shared (e.g., for stereotypes of groups like *Fat*) versus debated and varied (e.g., for stereotypes of groups like *Gay* or *Black*). To calculate a single metric of corpus-inconsistency in manifest content for each group we used a similar approach to the analyses of semantic change but focused only on the overlapping decade shared by all corpora: 1990-2000. That is, we computed the average cosine similarity between the embeddings (distributed meanings) of the top-*N* traits associated with a group in corpus A (e.g., *NYT* in 1990-2000) and the embeddings (distributed meanings) of the top-*N* traits associated with a group in corpus B (e.g., *Books* in 1990-2000). If the two corpora had perfectly overlapping lists of traits with very similar distributed meanings, we would find high average cosine similarity (i.e., cosine of 1). We then repeat the process for all pairs of corpora (i.e., *NYT-Books, NYT-COHA, NYT-CC, Books-COHA, Books-CC, COHA-CC*) and take the average, resulting in a final average score of corpus-consistency in semantic representations across corpora. To transform this metric into a score of *in*consistency (theoretically expected to yield significant positive correlations with our outcomes of change) we simply take the inverse, or 1 – average consistency across corpora.

Corpus-inconsistency in latent *valence* was computed as the median absolute deviation (MAD) of valence scores across all corpora in decade *t*. Median absolute deviation is computed by first computing the median across all scores, then taking the difference between each score and the median, and finally taking the median of the absolute values of those differences. For example, imagine the valence of *Abled* across the *NYT, Books, COHA,* and *CC* embeddings had scores of *[0.12, -0.01, -0.10, 0.07],* respectively; the median would be 0.03, and the MAD would be 0.10. An identical approach is applied to calculate corpus-inconsistency for latent warmth and competence.

Results

Change in manifest content. On average, across 72 groups and across all three time-stamped corpora, there was a high degree of change in manifest trait content (i.e., the distributed meanings of the top-ten traits) between decades or years (M = .71, SD = .05) (Figure 1A). As elaborated above, this score can be interpreted along the lines of correlation magnitudes, since it reflects essentially the inverse of a correlation[5], such that an effect size of .71 corresponds to a large effect. To aid interpretation, we can also compare this result to past methods quantifying trait content change through simple trait overlap, which indicated that the majority of traits (~60%) turned over across successive timepoints, on average across groups.

The three time-stamped corpora differed in the average degree of change in manifest content (Figures 1B-D). Representations in the *NYT* showed less change in manifest content than either *Books*, $t(141.61) = -18.00$, $p < .001$, $d = 3.00$, or *COHA*, $t(102.82) = -2.94$, $p = .004$, $d = 0.49$,

[5] That is, across corpora and groups, the average cosine *similarity* between traits across successive time was small-to-moderate in magnitude, average cosine across t and t+1 = .29. Therefore, the inverse (which reflects *change* in the manifest content) was moderate-to-large in magnitude, the $M_{semchange}$ = .71 reported in the main text.

and *COHA* also indicated less change in manifest content than *Books*, $t(99.94) = -8.00$, $p < .001$, $d = 1.33$. Such differences across corpora align with the fact that the *New York Times* corpus covers a smaller historical period (26 years) compared to the book-based corpora (which cover 100 years) and, as such, may capture less turnover in the manifest content of group stereotypes.

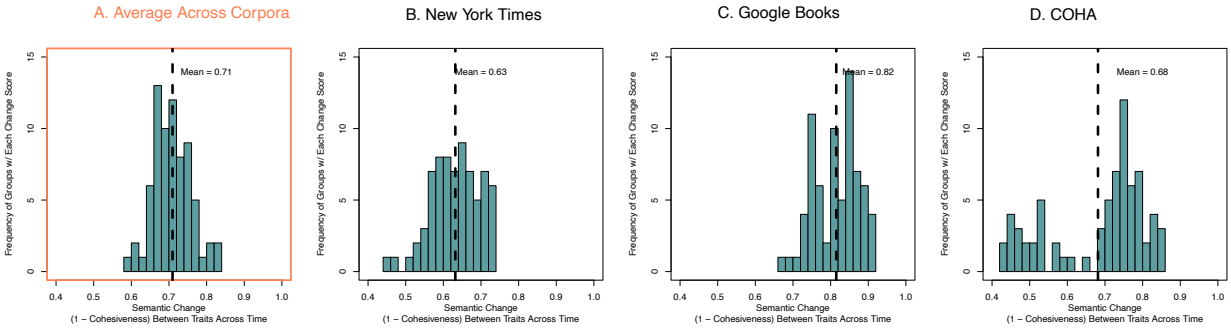


Figure 1. **Distributions of change in manifest stereotype content across (A) all corpora, averaged, (B) *New York Times*, (C) *Google Books,* and (D) *Corpus of Historical American English*.** Y-axis indicates frequency of observations (groups) at each score of manifest change. X-axis indicates change in manifest content indexed as the inverse of consistency between the trait representations at time *t* and *t*+1; higher scores indicate more change. Vertical dashed black line indicates the mean change score for manifest content.

Most relevant to the current manuscript, we also identified substantial *variability* in the extent of change in manifest content across groups (Figure 2), with a range of [.59, .83]. Interpreted alongside the alternative metric of trait overlap, results show that, for the groups with the least change, the majority of stereotype content (~80%) was stable over history; for the groups with the most change, however, only a small portion (~10-20%) of the top trait associates remained consistent across time. Below we discuss the correlates of such variability across groups to shed light on the features and types of groups that correspond to more change in manifest content.

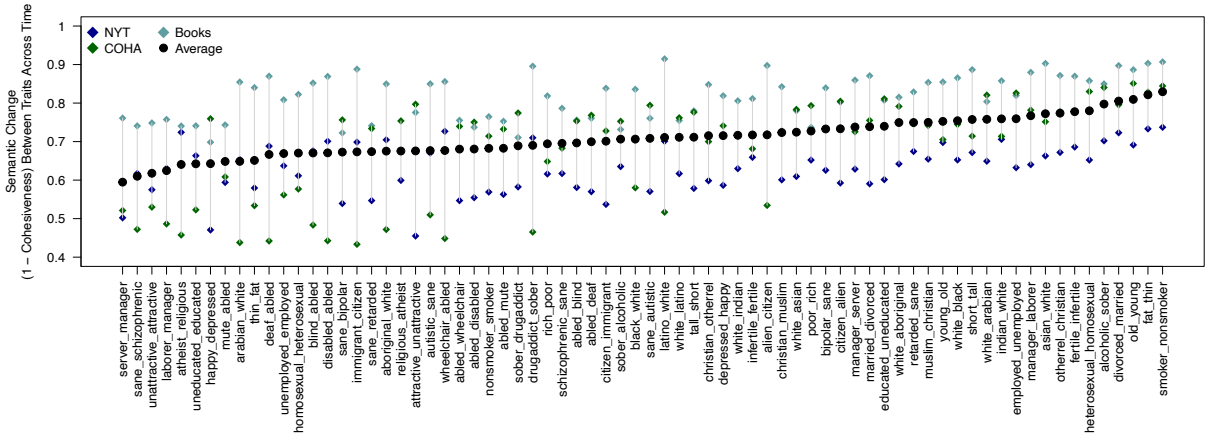


Figure 2. **Change in manifest content across groups.** Y-axis indicates the average score for change in manifest content indexed as the inverse of consistency between the trait representations at time *t* and *t*+1; higher scores indicate more change. X-axis indicates groups ordered from the least changing (*Server*) to the most changing (*Smoker*); the target group is listed first (before the underscore), with the comparison identity listed second. Black circles indicate mean change in manifest content across decades, collapsing across all corpora. Dark blue diamonds indicate change in manifest content in the *New York Times,* dark green diamonds indicate change in manifest content in *COHA,* and light blue diamonds indicate change in manifest content in *Google Books.* Vertical gray bars indicate the range of change scores across corpora (connecting the minimum to maximum).

Change in latent structure: warmth. Given that the manifest content of stereotypes (i.e., the distributed meaning of traits) changed by a large degree over time, the next question arises: did change also trickle into shifts along latent semantic (warmth and competence) and valence dimensions? Or did those particular underlying dimensions remain relatively stable, despite changes in manifest content?

Turning first to latent warmth: across corpora and all groups, there was change in warmth, with the absolute value of Spearman's *rho* for warmth trajectories revealing a moderate effect size ($M_{|rho|}$ = .38, $SD_{|rho|}$ = .14). Results again differed in the expected ways across the three corpora, with the *NYT* indicating less change in latent warmth ($M_{|rho|}$ = .28) than *Books* ($M_{|rho|}$ = .49), $t(125.01) = -5.10$, $p < .001$, $d = 0.85$, or *COHA* ($M_{|rho|}$ = .37), $t(131.13) = -2.21$, $p = .03$, $d = 0.37$, and *COHA* also changing significantly less than *Books*, $t(140.85) = -2.69$, $p < .001$, $d = 0.45$. Most crucial, we found wide variation in the degree of change in latent warmth across

groups (Figure 3), varying from groups that were persistently cold (*Drug addict*) and persistently warm (*White, Abled*) to groups that changed with decreasing (*Christian*) and increasing warmth (*Heterosexual*); Figure 7A presents results colored by the direction of change with additional details in the SM. In sum, when it comes to this initial subdimension of warmth, results show (1) relatively more stability, on average, than the observed change in manifest content, as well as (2) variability across groups that warrants further exploration – below we consider those correlates that help shed light on which groups change more than others along the warmth axis.

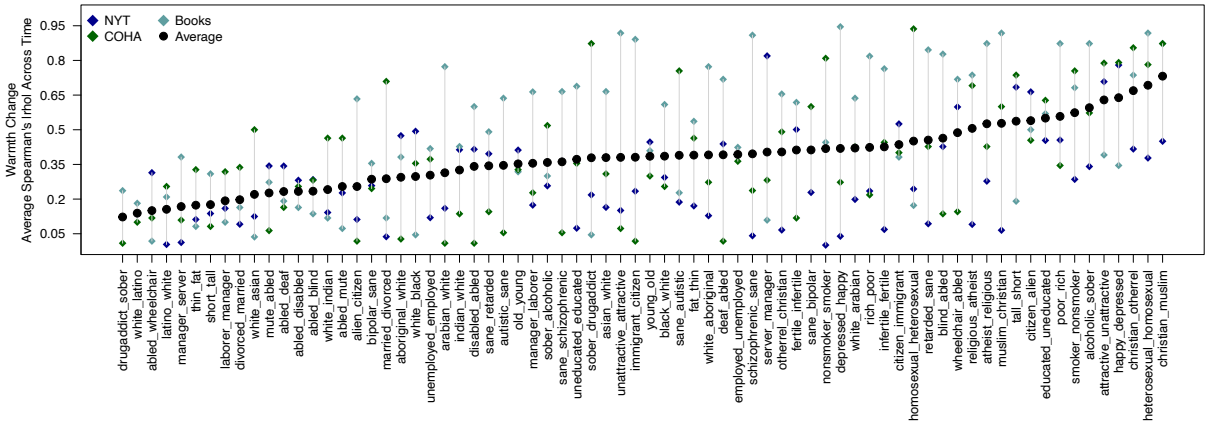


Figure 3. **Change in latent warmth across groups.** Y-axis indicates the slope of change in the top trait representations' average warmth, indexed as Spearman's |rho| for the timeseries of warmth scores. Higher scores indicate more change in warmth. X-axis indicates group labels ordered from the least changing (*Drug addict*) to the most changing (*Christian*), regardless of direction of change; the target group is listed first (before the underscore), with the comparison group listed second. Black circles indicate mean |rho|, collapsing across all corpora. Dark blue diamonds indicate warmth change |rho| from the *New York Times,* dark green diamonds indicate warmth change |rho| from *COHA,* and light blue diamonds indicate warmth change |rho| from *Google Books.* Vertical gray bars indicate the range of warmth change scores across corpora.

Change in latent structure: competence. Results were similar for change in latent competence, with average change that was moderate in magnitude ($M_{|rho|}$ = .44, $SD_{|rho|}$ = .16) similar to latent warmth, and lower than for change in manifest content. The corpora also differed in degree of latent competence change: here, *COHA* ($M_{|rho|}$ = .55), changed the most, compared to either the *NYT* ($M_{|rho|}$ = .35), $t(128.09) = -4.54$, $p < .001$, $d = 0.76$, or *Books* ($M_{|rho|}$ =

.41), $t(136.66) = 2.96$, $p = .004$, $d = 0.49$, with no difference between *Books* and *NYT,* $t(139.23) = -1.60$, $p = .11$, $d = 0.27$.

Again, and most central, we found that group representations varied substantially in their degree of latent competence change (Figure 4), ranging from groups that had very little change (*Sober, White, Other religion,* and *Smoker;* all with neutral or slightly positive competence across all time) to groups with consistent changes in competence, including a set of disability related groups, *Wheelchair-bound, Blind,* and *Deaf,* which generally reflected *decreases* in latent competence (Figure 7B). Following from the data on latent warmth, the same overarching conclusions persist for changes in latent competence: (1) there is relatively more stability, on average, than the changes in manifest content, although similar stability to the latent warmth; and (2) there is potentially meaningful variability across groups in the degree of change.

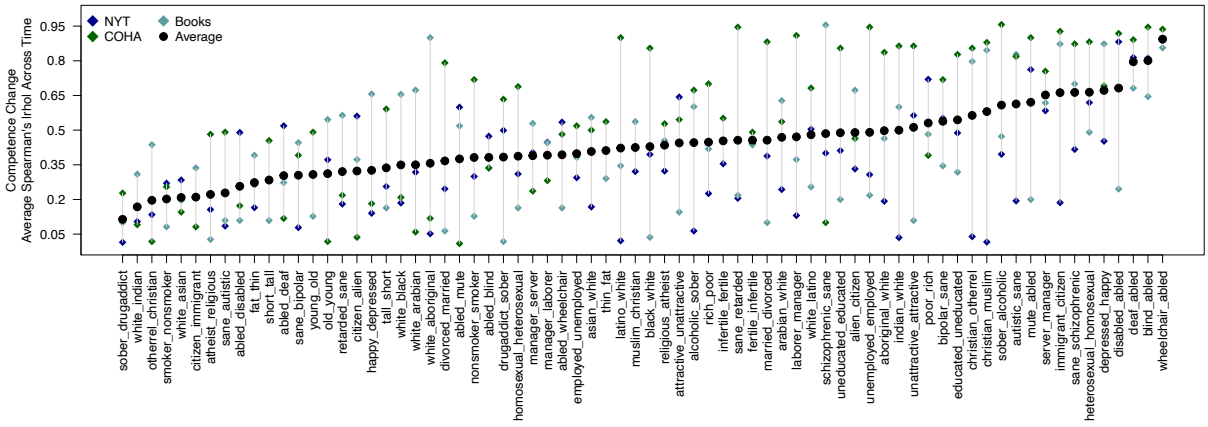


Figure 4. **Change in latent competence across groups.** Y-axis indicates the slope of change in the top trait representations' average competence indexed as Spearman's |rho| for the timeseries of competence scores, with higher scores indicating more change in competence. X-axis indicates group labels ordered from the least changing (*Sober*) to the most changing (*Wheelchair)*; the target group is listed first (before the underscore), with the comparison group listed second. Black circles indicate mean |rho|, collapsing across all corpora. Dark blue diamonds indicate competence change |rho| from the *New York Times,* dark green diamonds indicate competence change |rho| from *COHA,* and light blue diamonds indicate competence change |rho| from *Google Books.* Vertical gray bars indicate the range of competence change scores across corpora.

Change in latent structure: valence. Finally, results showed a similar degree of underlying stability in latent valence of trait representations across 115 years of English-language text, again with moderate effect sizes, $M_{|rho|} = .39$, $SD_{|rho|} = .13$ (Figure 5A). Results differed across the three time-stamped corpora (Figures 5B-D): as with manifest content and latent warmth, the *NYT* indicated the least valence change ($M_{|rho|} = .30$), significantly lower than *Books* ($M_{|rho|} = .53$), $t(132.62) = -6.61$, $p < .001$, $d = 1.10$, although only descriptively lower than *COHA*, $t(136.42) = -1.58$, $p = .12$, $d = 0.26$. *COHA* also changed significantly less than *Books*, $t(141.36) = -4.67$, $p < .001$, $d = 0.78$.

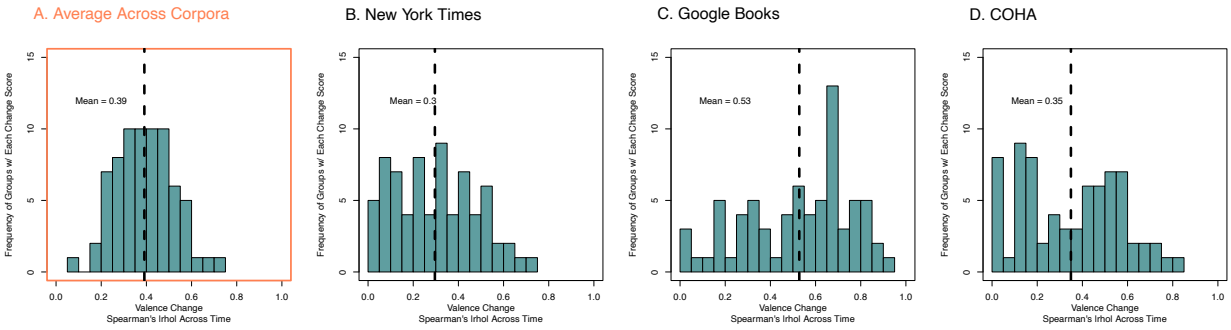


Figure 5. **Distributions of latent valence change across (A) all corpora, averaged; (B) *New York Times*; (C) *Google Books; and* (D) *Corpus of Historical American English*.** Y-axis indicates frequency of groups at each score of change in latent valence. X-axis indicates valence change scores indexed as the |rho| of the valence timeseries across successive decades or years (larger |rho| indicates more valence change). Vertical dashed gray line indicates the mean valence change score (i.e., mean |rho| across successive decades or years).

Finally, we observed variation in the degree of latent valence change across groups (Figure 6), with a similar range in values to that seen for both latent warmth and competence. Groups ranged from relative stability (always in negative valence; *Latino, Uneducated, Thin*) to changing both in the direction of increasing positivity (*Heterosexual, White*) and increasing negativity (*Christian*); see Figure 7C and additional details in the SM for raw *rho* slopes. Thus, as above, we emphasize the two primary take-aways: latent valence, like latent warmth and competence, is relatively more stable on average; but also, the variability in which groups reveal stability versus change demands further exploration and, if possible, explanation.

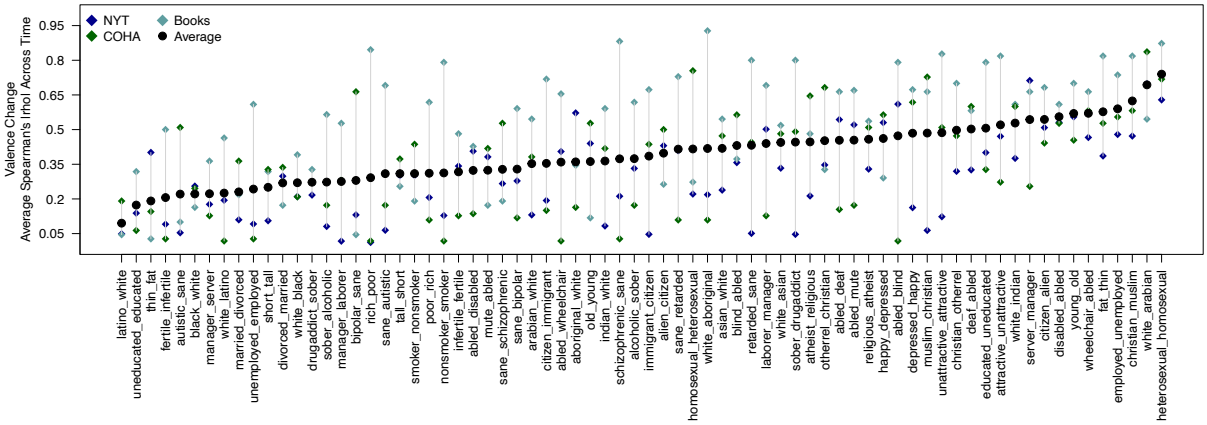
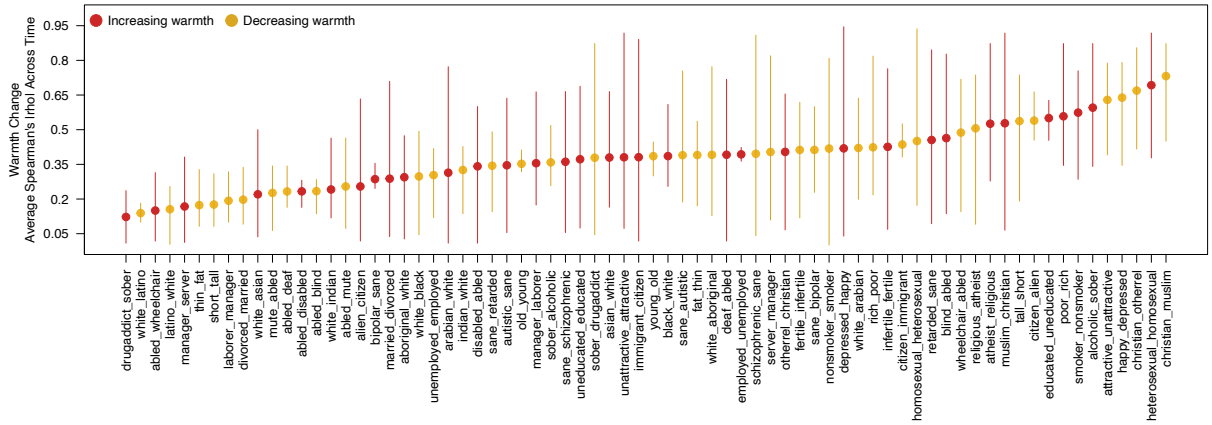
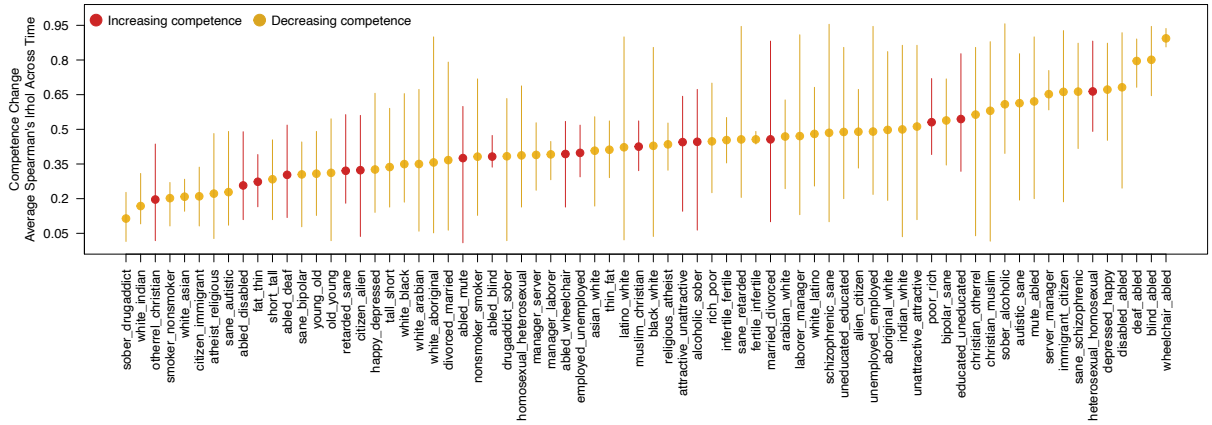


Figure 6. **Valence change across groups.** Y-axis indicates the slope of valence change indexed as Spearman's |rho| for the timeseries of valence scores, with higher scores indicating more valence change. X-axis indicates group labels ordered from the least changing (*Latino*) to the most changing (*Heterosexual*), regardless of direction of change; the target group is listed first (before the underscore), with the comparison group listed second. Black circles indicate mean |rho|, collapsing across all corpora. Dark blue diamonds indicate valence change |rho| from the *New York Times,* dark green diamonds indicate valence change |rho| from *COHA,* and light blue diamonds indicate valence change |rho| from *Google Books.* Vertical gray bars indicate the range of valence change scores across corpora (connecting the minimum valence change to the maximum valence change).

A.



B.



C.

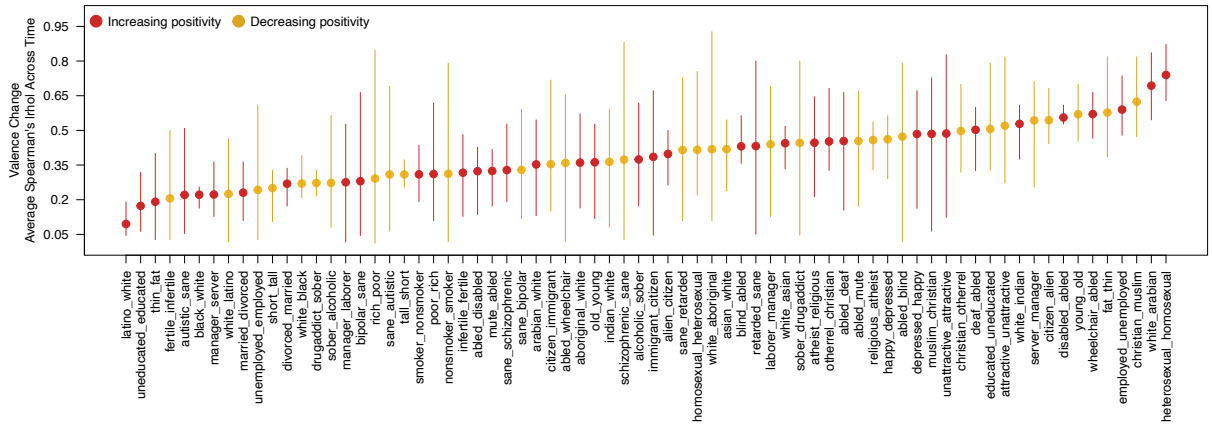


Figure 7. **Direction of change in latent dimensions of warmth (A), competence (B), and valence (C) across groups.** Y-axis indicates the slope of valence change indexed as Spearman's |rho| for the timeseries of valence scores, with higher scores indicating more change along the specified latent dimension. X-axis indicates group labels ordered from the least to the most changing, regardless of direction of change; the target group is listed first (before the underscore), with the comparison group listed second. Red circles indicate that the change, on average collapsing across all corpora, is increasing (i.e., greater warmth, competence, or positivity); yellow circles indicate that the change, on average collapsing across all corpora, is decreasing (i.e., lesser warmth, competence, or positivity). Vertical red or yellow bars indicate the range of valence change scores across corpora (connecting the minimum valence change to the maximum valence change).

Examining variability in group stereotype change. Confronted with such variability in stereotype change across the 72 groups, both in terms of manifest content and in latent warmth, competence, and valence, we next seek to understand the correlates of such variability as a first,

exploratory step towards an explanation of *why* some groups may change more than others along specific metrics. As discussed in the *Introduction*, we focus on two broad classes of variables: (1) the types (clusters) of groups; and (2) linguistic features of how groups are referred to in language, namely polysemy, frequency, semantic drift, and inconsistency of group representations across text sources. Again, when considering the type of groups, we examined both top-down researcher defined groups, and bottom-up empirically defined clusters (see methods above). Results from averages across corpora using the top-down clusters are presented in Tables 2-5; results from corpus-specific models and bottom-up clusters are in the SM.

Correlates of change in manifest content. The first and broadest question is whether these sets of theoretically relevant variables make sufficient traction in explaining the variability of change across groups. Indeed, the R^2 (and adjusted R^2) of the final regression models indicate that 40% (30% using adjusted R^2) of the variance in change of manifest content can be explained by the specified correlates – a combination of the type of group, linguistic features of the groups, and change along latent dimensions. Optimistically, this suggests that the chosen model – and the social psychological expectations it reflects – capture a substantial portion of how group representations vary across 115 years of historical text corpora.

Which of these variables were most meaningfully related to semantic change (i.e., the distributed meanings of top trait content)? First, results presented in Table 2 showed a marginally significant difference between groups in the body-related cluster and those in the sociodemographic cluster, such that stereotypes of sociodemographic groups (M = .72) changed descriptively more on average than did stereotypes of groups in the body-related cluster (M = .69). Of note, sequential models (see SM) showed that this difference moved from significant to marginal after including corpus-inconsistency as a predictor, suggesting that differences between

sociodemographic groups and body-related groups in terms of corpus-inconsistency may be a mediator of differences between these clusters in their patterns of change; such a result suggests (weak) support for the expectation that sociodemographic groups may be more debated and variable across subcultures and, as one consequence, more inclined to change.

More generally, however, the lack of clear significant differences across clusters of groups suggests that the dividing lines of change in manifest stereotype content may be better found along more nuanced boundaries than those captured by four broad sets of groups. As such, we move beyond the *type* of group to also consider relevant linguistic features of how groups are referred to in text.

Polysemy and frequency of group labels emerged as significant predictors of greater change in manifest content, with groups represented using more polysemous and more frequent labels showing more shift in content across time (Table 2). The finding for polysemy is broadly in line with research on linguistic change showing that lexical semantics change more for those words that are used in more diverse ways (Hamilton et al., 2016b). However, the finding for frequency is in the opposite direction to those of Hamilton and colleagues (2016), perhaps reflecting the differences in the outcome variables (i.e., here we focus on stereotypes, whereas Hamilton focused on general lexical definitions). Stereotypes, which are cultural constructions, may be more labile and open to reconstruction as it gets more attention, basic semantics of words may, by contrast, be solidified and taken as "factual" definitions when they are frequently used; such speculations represent new avenues for future research. Nevertheless, at the broadest level, the fact that linguistic processes have any relevance to *stereotype* change newly shows that higher-level social psychology (group representations in text) are interwoven with basic lexical phenomena.

Finally, when looking at relationships *among* the various metrics of change, we found that the degree of change in manifest content was significantly and positively related to change in latent warmth: groups that changed more in their top traits (e.g., *Smoker, Fat, Heterosexual, Alcoholic*) also changed more in their underlying warmth. Counterintuitively, we also found that change in manifest content was significantly but *negatively* related to change in latent average competence: the more a group stereotype changed in trait content, the more stable it was along axis of competence.

An example helps to illustrate these relationships. Consider the stereotypes of *Smoker* in *Books* in 1900, with top trait associates including *bland*, *sarcastic, sly, theatrical, grim,* reflecting a stereotype of *Smoker* as "mysterious" or "bad guys" e.g., the Marlboro Man ads of the 1950s (Gilman & Zhou, 2004). In 2000, by contrast, the top trait associates included *obnoxious, cowardly, dominant, soft, passive,* among others, now reflecting an entirely new semantic representation of the more contemporary stereotype of *Smoker* that evokes a perceived imposition placed on others (e.g., obnoxious because they smoke in public), but also a perceived cowardice to quit and passivity in controlling habits. Despite these changes in manifest content, the average competence in 1900 was -0.02 and in 2000 was -0.03, both times reflecting the similar, mixed competence in representations. In short, the manifest representations of groups may change widely over time and yet *resist* change in latent competence. Indeed, the negative relationship between trait content change and competence change suggest that changes in content might instead happen along a complementary latent dimension of warmth or in other semantic dimensions.

Table 2
Regressions predicting change in manifest content across groups, averaged across corpora

	b	*SE*	*t*	*p*
(Intercept)	-0.26	0.48	-0.54	.59
Cluster: Mental health	0.26	0.33	0.79	.44
Cluster: Occupation	-0.36	0.42	-0.85	.40
Cluster: Sociodemographics	*0.57*	*0.30*	*1.87*	*.07*
Semantic drift	0.16	0.12	1.27	.21
Polysemy	**0.47**	**0.15**	**3.17**	**.002**
Frequency	**0.59**	**0.17**	**3.44**	**.001**
Stigmatized	-0.002	0.30	-0.007	.99
Valence change	-0.17	0.13	-1.30	.20
Warmth change	**0.42**	**0.15**	**2.77**	**.007**
Competence change	**-0.30**	**0.12**	**2.45**	**.02**
Semantic corpus-inconsistency	0.13	0.17	0.80	.43
	R^2 = 0.41, Adjusted R^2 = 0.30, AIC = 186.62, BIC = 215.67			

Note. All metric variables were standardized (centered and scaled) before model fitting; slope estimates thus reflect standardized effect sizes. The "body-related" cluster was used as the dummy-coded baseline for group clusters, and the "stigmatized" groups were used as the dummy-coded baseline for prediction from the contrast of stigmatized/not stigmatized.

Correlates of change in latent warmth. Next, we examine the same correlates but using the dependent variable of change in latent warmth (i.e., the absolute value of the warmth trajectories). Again, the model R^2 (and adjusted R^2) indicated that 49% (41% using adjusted R^2) of variance across groups can be explained by the combination of chosen variables, indicating the relevance of social and linguistic variables to understanding change. Here, however, no significant differences emerged across clusters of groups, suggesting that these four group types may be similarly stable in average warmth across time. Additionally, the linguistic feature of frequency again emerged as significant predictor, although this result did not persist with robustness tests that excluded outliers of group label frequency (*Abled, Short,* and *Employed* were > 1 SD more frequent than other groups), suggesting caution in interpreting this counterintuitive result. No other linguistic features (drift, polysemy, or corpus-inconsistency) emerged as significant predictors of change in latent warmth.

Thus, the most meaningful correlates for latent warmth were the other metrics of stereotype change. Specifically, greater change in latent warmth was also related to greater change in manifest content and greater change in latent valence. This latter result lends confidence in the novel methodologies since warmth and valence are often highly correlated dimensions of meaning (Kurdi et al., 2019) and, as such, should reveal similar patterns of change.

Table 3
Regressions predicting change in latent warmth across groups, averaged across corpora

	b	*SE*	*t*	*p*
(Intercept)	-0.38	0.27	-1.02	.31
Cluster: Mental health	0.29	0.28	1.04	.30
Cluster: Occupation	-0.06	0.38	-0.15	.88
Cluster: Sociodemographics	0.04	0.25	0.17	.86
Semantic drift	-0.17	0.10	-1.68	.10
Polysemy	-0.18	0.14	-1.33	.19
Frequency	**-0.49**	**0.14**	**-3.63**	**<.001**
Stigmatized	0.16	0.22	0.75	.45
Change in manifest content	**0.28**	**0.11**	**2.64**	**.01**
Valence change	**0.47**	**0.09**	**5.03**	**<.001**
Competence change	0.10	0.11	0.98	.33
Warmth corpus-inconsistency	-0.01	0.10	-0.06	.95
	R^2 = 0.49, Adjusted R^2 = 0.39, AIC = 163.69, BIC = 192.73			

Note. All metric variables were standardized (centered and scaled) before model fitting; slope estimates thus reflect standardized effect sizes. The "body-related" cluster was used as the dummy-coded baseline for group clusters, and the "stigmatized" groups were used as the dummy-coded baseline for prediction from the contrast of stigmatized/not stigmatized.

Correlates of change in latent competence. As above, the regression model exploring correlates of latent competence indicated that 35% of variance (24% with adjusted R^2) could be explained through the selected combination of variables. Note, however, that this is the lowest R^2 across our four models predicting the four outcome-metrics of change. Indeed, the differences across groups in latent competence were not predicted by group clusters nor by any linguistic features (frequency, polysemy, drift, or corpus-inconsistency). The only significant predictor was

the negative relationship with change in semantic trait content (as discussed in detail above). Thus, of all manifest or latent metrics of stereotype change, it appears that change in *competence* may be the least explicable through past hypotheses of social and linguistic correlates of change, and thus warrants more targeted theory development.

Table 4
Regressions predicting change in latent competence across groups, averaged across corpora

	b	*SE*	*t*	*p*
(Intercept)	*0.86*	*0.45*	*1.90*	*.06*
Cluster: Mental health	-0.43	0.34	-1.27	.21
Cluster: Occupation	-0.11	0.46	-0.23	.82
Cluster: Sociodemographics	-0.21	0.31	-0.68	.50
Semantic drift	0.11	0.13	0.87	.39
Polysemy	0.11	0.17	0.68	.50
Frequency	-0.13	0.18	-0.72	.47
Stigmatized	*-0.45*	*0.26*	*-1.71*	*.09*
Change in manifest content	**-0.36**	**0.13**	**-2.69**	**.01**
Valence change	0.06	0.14	0.46	.65
Competence change	0.16	0.16	0.99	.33
Corpus-inconsistency	-0.10	0.12	-0.83	.41
	R^2 = 0.36, Adjusted R^2 = 0.24, AIC = 191.32, BIC = 220.36			

Note. All metric variables were standardized (centered and scaled) before model fitting; slope estimates thus reflect standardized effect sizes. The "body-related" cluster was used as the dummy-coded baseline for group clusters, and the "stigmatized" groups were used as the dummy-coded baseline for prediction from the contrast of stigmatized/not stigmatized.

Correlates of change in latent valence. Finally, turning to cross-group differences in change along latent valence, the model accounted for 38% of variance (27% with adjusted R^2), somewhat lower than models explaining change in manifest content or latent warmth, but nonetheless capturing a substantive fraction of variance in how groups have changed along the axis of positivity/negativity. This lower explained variance was also reflected in the fact that no significant differences were found across clusters of groups nor as a function of the linguistic features of polysemy, drift, or corpus-inconsistency.

However, like with change in manifest content, the frequency of group labels significantly and positively predicted the degree of valence change across groups. We interpret this result as suggesting that group stereotypes that are more frequently represented in texts may have more opportunities to be targets of interventions and revisions. In turn, the more those group stereotypes are targets of interventions, the more they will be discussed and debated, perhaps generating a bidirectional feedback cycle between frequency and change for both semantics and valence. Finally, as discussed in detail above, greater change in latent valence was positively related to greater change in latent warmth.

Table 5
Regressions predicting change in latent valence across groups, averaged across corpora

	b	*SE*	*t*	*p*
(Intercept)	-0.13	0.44	0.30	0.76
Cluster: Mental health	-0.42	0.35	-1.20	0.23
Cluster: Occupation	-0.26	0.44	-0.60	0.55
Cluster: Sociodemographics	-0.03	0.30	-0.11	0.91
Semantic drift	0.01	0.12	0.10	0.92
Polysemy	0.25	0.16	1.62	0.11
Frequency	**0.50**	**0.16**	**3.18**	**0.002**
Stigmatized	0.00	0.25	0.00	>.99
Change in manifest content	-0.16	0.13	-1.23	0.22
Warmth change	**0.64**	**0.13**	**5.02**	**<.001**
Competence change	0.06	0.13	0.44	0.66
Valence corpus-inconsistency	0.12	0.11	1.10	0.28
	R^2 = 0.39, Adjusted R^2 = 0.28, AIC = 184.54, BIC = 213.58			

Note. All metric variables were standardized (centered and scaled) before model fitting; slope estimates thus reflect standardized effect sizes. The "body-related" cluster was used as the dummy-coded baseline for group clusters, and the "stigmatized" groups were used as the dummy-coded baseline for prediction from the contrast of stigmatized/not stigmatized.

General Discussion

Stereotypes of groups are often thought to have persistence and resistance to change (Lippmann, 1922), casting long shadows throughout history that uphold hierarchies and differences in treatment between groups (Fiske, 2018). And yet, compared to portrayals of just a

few decades ago, the stereotypes and representations of gay and lesbian people (Charlesworth & Banaji, 2022a; McCarthy, 2020), Black Americans (Bobo et al., 2012), and people with schizophrenia (Pescosolido et al., 2010, 2021) are noticeably different today. When such evidence of change is set against a backdrop of assumed stereotype persistence, questions naturally arise: is change only possible for the handful of groups that we have had the methods and archival surveys to study? What of the vast array of conceptually distinct group stereotypes (e.g., about disability, other body-related identities, other mental illnesses, occupations, or employment status)? How has history unfolded for this wider sample space of group targets? Moreover, what of the different metrics (manifest versus latent structure) of group stereotypes? Have stereotypes shifted only in their manifest content (associated traits) or does change also extend along latent dimensions of valence and semantic meaning?

To date, methodological limitations have kept the study of stereotype change focused on one or a few groups in isolation, for relatively short timescales, and for a single metric of stereotypes. To overcome such limitations, we leveraged innovations in natural language processing (diachronic word embeddings) to provide the first comprehensive quantitative portrait of both manifest and latent stereotype change across 72 varied group targets tracked through 4 corpora of contemporary and historical English-language text spanning 115 years.

The results emphasize two overarching conclusions for advancing our understanding of historical stereotype change, with each conclusion elaborated below. First, on average, across groups and corpora, group stereotypes appear to change more in their manifest content than in latent dimensions of valence, warmth, or competence. Second, the 72 groups varied substantially in their degree of change versus stability, and such variance was well-explained by a combination of relevant correlates including the type of group and linguistic features of how the

group is referred to in text. Alongside such conceptual contributions, we also discuss the methodological advances that help move our field towards a new frontier of quantifying and explaining historical changes in stereotypes at unprecedented scales.

Change in manifest content is greater than in latent valence, warmth, competence

Since the early studies of Katz and Braly (1933), the default method for studying the content of group stereotypes has been to have participants select trait adjectives (usually from a predefined list of ~100 traits) that they believe to be most associated with a given target group. Change in stereotypes is then quantified in terms of how those top-associated traits shift over time, such as when a trait like "lazy" goes from being a top-associate of *Black American* to no longer being explicitly endorsed (Bergsieker et al., 2012). In the current work, we too began by quantifying stereotype change in terms of how the top-associated traits (and their distributed meaning across embedding space) have shifted over time. Results showed that such manifest content indeed changed by a moderate-to-large effect size, showing that the meaning of many group stereotypes has meaningfully shifted over time. Thus, if we look exclusively at how groups have transformed in their trait associates, we would conclude that societal representations are malleable and responsive to the many changes that unfolded across 115 years of history.

However, the stereotypes of groups can also be organized along underlying axes (latent dimensions) of meaning, with three of the most dominant axes captured through latent valence (Osgood et al., 1967), as well as latent warmth and competence (Fiske et al., 2002). Such reduced axes provide a first means to *quantify* underlying stereotype meaning and thereby identify whether the complex changes in manifest content can be reduced to change occurring specifically along these latent dimensions. Of course, collapsing the rich complexity of qualitative stereotype content along reduced dimensions may not be sufficient to capture *all* the

ways in which group stereotypes may vary and change over time. Indeed, results showed that changes in the latent dimensions of valence, warmth, and competence were only small-to-moderate in magnitude, with all latent subdimensions showing similar relative stability.

Such a contrast between greater changes in manifest content and lesser changes in latent valence aligns with both recent (Charlesworth et al., 2022a) and classic work (Bergsieker et al., 2012; Devine & Elliot, 1995; Gilbert, 1951). Crucially, however, the inclusion of further latent dimensions of warmth and competence advances a new, more nuanced interpretation of such findings. That is, the differences in change observed previously do not appear to be divided simply between semantics (i.e., meaning of group stereotypes) and valence (i.e., positivity/negativity). Rather, the more appropriate distinction may lie between the *level of analysis* of a stereotype – dividing manifest content from latent structure. The complex historical change in manifest stereotype content is not reducible to a general trend in which all groups on average are represented with more or less warmth, competence, or valence; there has been no simple shift along one dimension of meaning. Instead, historical change in stereotype content may reflect much greater group-specific complexity (e.g., axes of meaning that are unique to a group target, such as dimensions of "impulsivity"/"control" or "traditional"/"innovative" for age stereotypes).

For translational work, the observed dissociation between manifest change and latent stability may also deepen understanding of how and why stereotypes and their consequences (e.g., status hierarchies or discrimination) appear to be so persistent across time. On the surface, there can be convincing empirical demonstrations of *change* in respondents' self-reported and implicit measures of stereotypes (e.g., Charlesworth & Banaji, 2021) reflecting the true and complex ways in which group stereotypes are evolving along their group-specific dimensions of meaning.

And yet, groups may continue to experience differences in status and opportunity (e.g., Chetty et al., 2020) because they also continue to be divided along underlying axes of valence, warmth, and/or competence – axes that are known to shape group discrimination (Cuddy et al., 2008). Perhaps, then, for an intervention to effectively disrupt a hierarchy, the focus will need to go beyond only changing the content of representations (e.g., to change the representation of *Disabled* from *incapable* to a seemingly more equitable associate of *disadvantaged*). Instead, attention will also be needed to shift the latent positivity/negativity, warmth/coldness, or competence/incompetence of group representations.

Change varies in predictable ways across 72 group stereotypes

In addition to expanding the metrics of stereotype change, this project also uses what is, to-date, the largest number of group targets investigated using word embeddings. Expanding the scope of groups has numerous advantages, most importantly that the added diversity and sample size of group targets facilitates the first quantitative tests of the correlates that help explain *which* groups change. Two overarching findings are notable from such tests. First, groups did indeed vary substantially in the degree of change, with all four metrics of stereotypes showing ranges that varied from groups with largely stable and consistent representations across time (e.g., representations of *Abled* or *Sane* consistently included trait words like “capable,” “reasonable,” “independent” and reflected stable latent meanings) to groups that exhibited almost entirely new stereotypes in manifest content and latent structure (e.g., representations of *Smoker* went from referring to relatively innocuous, even warm associates of “aloof,” “relaxed,” “moody” to negative, cold associates of “severe,” “bitter,” “harsh”).

Second, it is equally notable that the models predicting such wide variability of stereotype change across 72 groups were sufficient to explain between at least 24%, and up to

49%, of variance across groups. The overarching explanatory power of such models is perhaps surprising. In past work, we and others have offered rough interpretations of features that corresponded to group differences in change: for instance, observing that racial/ethnic groups show greater change and variation than non-racial/ethnic groups (Appiah, 2018; Charlesworth et al., 2022a; Fiske, 2017); or that groups with greater frequency of discussion (e.g., race, sexuality) show greater change than groups that are rarely mentioned (e.g., age, disability) (Charlesworth & Banaji, 2019). Such qualitative interpretations were offered with the caveat that such features of group type or frequency likely only described a small portion of variability across group targets. After all, theoretical interpretations developed to explain a small set of data from a handful of groups may not have strong predictive relevance when extended to a wider sample space of 72 groups that vary along many more undefined axes. And yet, the finding that the current regression models can indeed explain meaningful variability across groups lends, to our knowledge, the first quantitative evidence in support of past hypotheses. We hope that such expansive methods and samples of groups will continue to spur both theory development and theory testing to identify additional group features that might capture the remaining variance.

Correlates of stereotype change across groups: the role of linguistic features

Among the correlates explored, significant predictors emerged in *linguistic features* of how the groups were referred to in text. For instance, greater polysemy (multiple meanings) of group labels was related, as would be expected, to greater change in a group's manifest content; and, most consistently, greater frequency of group label mentions was related to greater change in a group's manifest content and latent valence. At the broadest level, we take such results to emphasize how higher-order social processes of stereotype change may be interwoven with basic linguistic processes of word usage (Hamilton et al., 2016a, 2016b). It has been shown before that

the word used to describe a group will influence how that group is stereotyped in a given moment (e.g., *Black American vs. African American;* Hall et al., 2021); however, the current results are novel in showing that other, more general and basic features of groups' labels – their polysemy, drift, and frequency – may also be tied into how group stereotypes change. Future work is poised to consider the time-lagged relationships between stereotype change and linguistic features, identifying, for example, how increases in frequency of referring to a group may increase the rate of stereotype change at a subsequent time-step (or vice versa).

Correlates of stereotype change across groups: the role of corpus-inconsistency

In addition to investigating three well-established linguistic features of group representations, we also made use of four varied text sources to introduce a fourth linguistic feature – *corpus-inconsistency* – that quantifies how a group stereotype varies across texts. As discussed in the *Introduction,* such a metric is helpful for operationalizing the theoretically relevant variable of social consensus (Gardner et al., 1973), with groups that have higher social consensus and consistency across texts expected to be taken more as "fact" and therefore less likely to change. In contrast, groups that have lower social consensus and more inconsistency across texts may be more open to debate, minority influence and, ultimately, to change (Gardikiotis, 2011).

While the metric of corpus-inconsistency has some interpretational ambiguity in terms of the *source* of inconsistency (since the four sets of pretrained embeddings vary in a range of unspecified ways), the face-validity of results (elaborated in the SM) suggests it can still be a useful tool for operationalizing consensus or variability. For instance, dominant and non-stigmatized groups like *Abled* and *Sane* show high consistency in their representations, regardless of the text source; by contrast, stigmatized social groups (e.g., *Latino, Aboriginal,*

Black) more prone social desirability (Devine et al., 2002) showed high *inconsistency* in representations. Having said this, differences across groups in corpus-inconsistency did not predict change in manifest content nor in any latent dimension (although there was suggestive evidence it may play a mediator role for change in manifest content). As such, although the metric holds interest in its own right for future work, the current data suggest it may not play a central explanatory role for group stereotype change. We hope that the metric will be implemented in future work using larger variation of text sources (e.g., blogs, conversations, newspapers from different political leanings) trained using similar methods and preprocessing, such that the sources of inconsistency might be more carefully controlled and understood.

Correlates of stereotype change across groups: the role of group type

Finally, in the same regression models discussed above, we also found that the *type* of group (e.g., clusters of groups as sociodemographic or body-related) was related to the degree of change in manifest content, with sociodemographic groups changing descriptively more than other clusters. However, the type of group was not significantly predictive of change along latent valence, warmth, and competence.

Given the many conceptual and empirical distinctions between these clusters of groups (Pachankis et al., 2018), the absence of evidence for differences in change across clusters might seem unexpected. We interpret such a result as suggesting that the current clustering of groups do not provide the best dividing lines for how change differs across groups. It remains possible that a different clustering of these 72 groups may identify underlying patterns in which certain theoretically defined types of groups reveal more or less change. For instance, the 72 groups also differ in terms of how their stigma "functions" in society: e.g., some stigmas are thought to serve a function of pathogen avoidance, others serve a function of dominance and resource

exploitation, and yet others serve a function of norm conformity (Hatzenbuehler et al., 2013; Link & Phelan, 2001; Phelan et al., 2008). Future work may reveal that clusters according to stigma function are better predictors of the variation in patterns of change. Additionally, it is possible that adding more groups, such as more "threatening" groups related to deviant criminal behaviors, may provide the variance necessary to uncover unique clusters of change. Ultimately, the current findings emphasize conceptual advantages coming from a comprehensive comparative approach across groups, while also making accessible the methods for even more expansive studies going forward.

Methodological implications for studying group stereotypes through text.

Beyond the conceptual implications, the results also hold methodological innovations for research using natural language processing to study group stereotypes and stereotype change. First, a major methodological contribution is the introduction of new approaches to test shifts in the manifest stereotype content across time using the relationship between the distributed meaning (i.e., embeddings) of top trait associates across timepoints rather than, as previous work has done, counting overlap of trait associates (Charlesworth et al., 2022a). This development is helpful not only for improving the study of stereotype content change in text but may also be implemented when using traditional survey methods in which participants select traits over time (Katz & Braly, 1933). There too, using distributed meanings of the top-associated traits can quantitatively distinguish the *degree* of change between two related traits ("lazy" to "helpless") versus unrelated traits ("lazy" to "dirty") thus adding nuance to understanding the amount and content of stereotype change among human participants.

Second, we also provide methodological advances for identifying changes in *latent* dimensions of stereotype meaning. That is, we show how to extract not only latent valence

(through the average scores of top traits in terms of their positivity/negativity) but also latent warmth and competence, by projecting the traits along dimensions newly specified using dictionaries of warmth and competence words (Nicolas et al., 2021, 2022). Such methods are generalizable to any other dimensions of meaning of interest to researchers (e.g., agency, communion, etc.). Additionally, given concerns that traits themselves may change in their degree of positivity/negativity (as well as their degree of competence/incompetence, or warmth/coldness), we developed new methods to extract the scores of traits along a specified dimension of valence, warmth, or competence *within each timepoint* of text. That is, we were able to extract historically contextualized ratings of traits along dimensions of meaning, an advance that we hope can assist not only in more accurately representing stereotype change but also in understanding linguistic change more broadly.

Ultimately, the current manuscript provides a template for expansive quantitative and qualitative comparative studies of group stereotypes and stereotype change. While past conclusions have been limited in studying only a few selected groups, across a few decades, and within only small subpopulations of society, the NLP methods introduced here are widely flexible to study multiple identities, timespans, metrics of stereotypes and stereotype meaning, and any narrative that has been recorded through text. We note again that all data, processing, and analysis code are openly available to researchers through the Open Science Framework (https://osf.io/gzuy4/) to motivate continued discoveries and investigations.

Limitations and conclusions.

Despite the advances made towards understanding variability in group stereotypes, the chosen corpora have both shared and unique limitations. For instance, all four corpora are subject to the potential editing and selection biases inherent in archived texts. Additionally, the *Books*

embeddings are not balanced between fiction and non-fiction; the precise sources of text in the *Common Crawl* embeddings are unknown since it is a representative scrape of all Internet text; and the *COHA* and *NYT* embeddings are relatively smaller in size and therefore have less coverage of low frequency words. Although triangulating stereotypes across multiple corpora helps to guard against idiosyncrasies resulting any corpus alone, future work is poised to make use of the increasing availability of pretrained embeddings corpora to address these limitations.

One notable limitation shared across corpora is that all texts were in English and had a Western focus. Given that stereotypes of groups vary across cultures and languages (Major & O'Brien, 2005; Stangor & Crandall, 2000), it will be important to extend the current study to available embeddings trained on corpora from different languages as well (e.g., Grave et al., 2018). It will be of particular interest to consider whether the manifest stereotype content of groups may vary across cultures, even if the latent subdimensions of valence, warmth, and competence may be similar across societies (Cuddy et al., 2009; Fiske, 2018). Relatedly, the selection of groups used in the current manuscript were drawn from a taxonomy created by experts set in an English-speaking and Western-centric context (Pachankis et al., 2018). Consequently, extending to other languages and cultures may also help introduce new groups that are uniquely stigmatized in other contexts (e.g., certain religious, ethnic subgroups, physical differences).

Additionally, using *words* to study group representations carries inherent limitations. Most notably, in the current project we made use of single (static) word embeddings, in which each word has only one vector to represent its meaning. Static word embeddings have substantial advantages including that they are relatively computationally inexpensive, and are more directly related to typical methods of assessing stereotypes (e.g., single trait generation; Bergsieker et al.,

2012; Katz & Braly, 1933). However, single word embeddings prevent us from investigating groups most appropriately represented with multiple words (e.g., *Teen parent*) or from distinguishing between related groups (e.g., *Lung cancer* versus *Breast cancer*). We are encouraged by ongoing work adapting contextualized embedding approaches (e.g., BERT; (Devlin et al., 2018) diachronically across time (Hofmann et al., 2021).

Finally, the results reported here reflect only the first pass of analyses that can be performed on the rich data of trait stereotypes across dozens of groups and hundreds of years. We have not examined the changes of any one group in detail (e.g., what are the exact traits that changed or persisted for the representations of *Schizophrenic, Mute, Muslim,* and so on), leaving largely unexplored the qualitative content changes of many socially-relevant stereotypes. The goal is that research building from the current empirical data may help shed light on where change has succeeded in transforming our representations of groups and then use those lessons to expand change across other group targets as well.

References

Appiah, A. (2018). *The lies that bind: rethinking identity, creed, country, color, class, culture*. Liveright, W. W. Norton.

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, *102*(6), 1214–1238. https://doi.org/10.1037/a0027717

Bhatia, N., & Bhatia, S. (2021). Changes in Gender Stereotypes Over Time: A Computational Analysis. *Psychology of Women Quarterly*, *45*(1), 106–125. https://doi.org/10.1177/0361684320977178

Bobo, L. D., Charles, C. Z., Krysan, M., & Simmons, A. D. (2012). The Real Record on Racial Attitudes. In P. V. Marsden (Ed.), *Social Trends in American Life: Findings from the General Social Survey since 1972* (pp. 38–83). Princeton University Press.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., & Banaji, M. R. (2021). Patterns of Implicit and Explicit Attitudes II. Long-Term Change and Stability, Regardless of Group Membership. *American Psychologist*, *76*(6), 851–869. https://doi.org/10.1037/amp0000810

Charlesworth, T. E. S., & Banaji, M. R. (2022a). Patterns of Implicit and Explicit Attitudes IV. Long-Term Change and Stability From 2007 to 2020. *Psychological Science*, *33*(9), 1347–

1371. https://doi.org/https://doi.org/10.1177/0956797622108425

Charlesworth, T. E. S., & Banaji, M. R. (2022b). Word embeddings reveal social group attitudes and stereotypes in large language corpora. In M. Dehghani & R. L. Boyd (Eds.), *Handbook of Language Analysis in Psychology* (pp. 594–508). Guilford Publications Inc.

Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022a). Historical Representations of Social Groups Across 200 Years of Word Embeddings from Google Books. *Proceedings of the National Academy of Sciences*, *119*(28). https://doi.org/https://doi.org/10.1073/pnas.2121798119

Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022b). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(28), e2121798119. https://doi.org/10.1073/PNAS.2121798119/SUPPL_FILE/PNAS.2121798119.SAPP.PDF

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and Economic Opportunity in the United States: an Intergenerational Perspective*. *The Quarterly Journal of Economics*, *135*(2), 711–783. https://doi.org/10.1093/qje/qjz042

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M.

H., Croizet, J.-C., Ellemers, N., Sleebos, E., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailó N 13, R., Morales, E., Moya, M., Palacios, M., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*, 1–33. https://doi.org/10.1348/014466608X314935

Davies, M. (2010). *Corpus of Historical American English*.

Devine, P. G., & Elliot, A. J. (1995). Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139–1150. https://doi.org/10.1177/01461672952111002

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. https://doi.org/10.1037/0022-3514.82.5.835

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of North American Chapter of the Association for Computational Linguistics-Human Language Technologies 2019*, 4171–4186. http://arxiv.org/abs/1810.04805

Fiske, S. T. (2017). Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) Versus Variable Stereotypes (Race, Ethnicity, Religion). *Perspectives on Psychological Science*, *12*(5), 791–799. https://doi.org/10.1177/1745691617708204

Fiske, S. T. (2018). Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, 1–7. https://doi.org/10.1111/1467-8721.ep10771786

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. https://doi.org/10.1037//0022-3514.82.6.878

Gardikiotis, A. (2011). Minority Influence. *Social and Personality Psychology Compass*, *5*(9), 679–693. https://doi.org/10.1111/j.1751-9004.2011.00377.x

Gardner, R. C., Kirby, D. M., & Finlay, J. C. (1973). Ethnic stereotypes: The significance of consensus. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *5*(1), 4–12. https://doi.org/10.1037/h0082327

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 of the United States of America*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gilbert, G. M. (1951). Stereotype persistance and change among college students. *Journal of Abnormal and Social Psychology*, *46*(2), 245–254. https://doi.org/10.1037/h0053696

Gilman, S. L., & Zhou, X. (2004). *Smoke : a global history of smoking* (S. L. Gilman & X. Zhou (eds.)). University of Chicago Press. https://press-uchicago-edu.ezp-prod1.hul.harvard.edu/ucp/books/book/distributed/S/bo3535915.html

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Prentice-Hall. https://doi.org/10.2307/2091442

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning Word Vectors for 157 Languages*. https://fasttext.cc/

Greenberg, B. S., Eastin, M., Hofschire, L., Lachlan, K., & Brownell, K. D. (2003). Portrayals of Overweight and Obese Individuals on Commercial Television. *American Journal of Public Health*, *93*(8), 1342–1348. https://doi.org/10.2105/AJPH.93.8.1342

Hall, E. V., Townsend, S. S. M., & Carter, J. T. (2021). What's in a Name? The Hidden Historical Ideologies Embedded in the Black and African American Racial Labels. *Psychological Science*, *32*(11), 1720–1730. https://doi.org/10.1177/09567976211018435

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2116–2121. https://doi.org/10.18653/v1/d16-1229

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1489–1501. https://doi.org/10.18653/v1/p16-1141

Haslam, N. (2016). Concept Creep: Psychology's Expanding Concepts of Harm and Pathology. *Psychological Inquiry*, *27*(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

Hatzenbuehler, M. L., Phelan, J. C., & Link, B. G. (2013). Stigma as a fundamental cause of population health inequalities. *American Journal of Public Health*, *103*(5), 813–821. https://doi.org/10.2105/AJPH.2012.301069

Hofmann, V., Pierrehumbert, J., & Schütze, H. (2021). *Dynamic Contextualized Word Embeddings*. 6970–6984. https://doi.org/10.18653/v1/2021.acl-long.542

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Jones, E. E. (1984). *Social stigma : the psychology of marked relationships*. W.H. Freeman.

Jones, E., Farina, A., Hastord, A., Markus, H., Miller, D., & Scott, R. (1984). *Social stigma: The psychology of marked relationships*. Freeman.

Jones, J. J., Amin, M. R., Kim, J., & Skiena, S. (2020). Stereotypical gender associations in language have decreased over time. *Sociological Science*, *7*, 1–35. https://doi.org/10.15195/v7.a1

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, *28*(3), 280–290. https://doi.org/10.1037/h0074049

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387–15392. https://doi.org/10.1073/pnas.251541498

Lin, Y., Michel, J.-B., Aiden, L., Orwant, J., Brockman, W., & Petrov, S. (2012). *Syntactic Annotations for the Google Books Ngram Corpus*. Association for Computational Linguistics. http://books.google.com/ngrams.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing Stigma. *Annual Review of Sociology*, *27*(2001), 363–385.

Lippmann, W. (1922). *Public Opinion*. MacMillan Press.

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. In *Annual Review of Psychology* (Vol. 56, pp. 393–421). Annual Reviews. https://doi.org/10.1146/annurev.psych.56.091103.070137

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian

is to police: Detecting and removing multiclass bias in word embeddings. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 615–621. https://doi.org/10.18653/v1/n19-1062

McCarthy, J. (2020). *U.S. support for same-sex marriage matches record high*. Gallup News. https://news.gallup.com/poll/311672/support-sex-marriage-matches-record-high.aspx

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint*. http://arxiv.org/abs/1301.3781

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. *Proceedings of the International Conference Language Resources and Evaluation*. http://arxiv.org/abs/1712.09405

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv Preprint*. http://arxiv.org/abs/1310.4546

Moscovici, S. (1976). *Social influence and social change*. Published in cooperation with European Association of Experimental Social Psychology by Academic Press.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *Eur J Soc Psychol*, *51*, 178–196. https://doi.org/10.1002/ejsp.2724

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. https://doi.org/10.1037/pspa0000312

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1967). *The measurement of meaning*. University of Illinois Press.

Pachankis, J. E., Hatzenbuehler, M. L., Wang, K., Burton, C. L., Crawford, F. W., Phelan, J. C., & Link, B. G. (2017). The Burden of Stigma on Health and Well-Being: A Taxonomy of Concealment, Course, Disruptiveness, Aesthetics, Origin, and Peril Across 93 Stigmas. *Personality and Social Psychology Bulletin*, 014616721774131. https://doi.org/10.1177/0146167217741313

Pachankis, J. E., Hatzenbuehler, M. L., Wang, K., Burton, C. L., Crawford, F. W., Phelan, J. C., & Link, B. G. (2018). The Burden of Stigma on Health and Well-Being: A Taxonomy of Concealment, Course, Disruptiveness, Aesthetics, Origin, and Peril Across 93 Stigmas. *Personality and Social Psychology Bulletin*, *44*(4), 451–474. https://doi.org/10.1177/0146167217741313

Peabody, D. (1987). Selecting Representative Trait Adjectives. In *Journal of Personality and Social Psychology* (Vol. 52, Issue 1). http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=1987-15626-001&S=L&D=pdh&EbscoContent=dGJyMNLe80SeqLU4yOvsOLCmr1Gep7NSsqi4TK6WxWXS&ContentCustomer=dGJyMPGuslGwqrFIuePfgeyx44Dt6fIA

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

https://doi.org/10.3115/v1/d14-1162

Pescosolido, B. A., Halpern-Manners, A., Luo, L., & Perry, B. (2021). Trends in Public Stigma of Mental Illness in the US, 1996-2018. *JAMA Network Open*, *4*(12), e2140202–e2140202. https://doi.org/10.1001/jamanetworkopen.2021.40202

Pescosolido, B. A., Martin, J. K., Long, J. S., Medina, T. R., Phelan, J. C., & Link, B. G. (2010). "A disease like any other"? a decade of change in public reactions to schizophrenia, depression, and alcohol dependence. *American Journal of Psychiatry*, *167*(11), 1321–1330. https://doi.org/10.1176/appi.ajp.2010.09121743

Phelan, J. C., Link, B. G., & Dovidio, J. F. (2008). Stigma and prejudice: One animal or two? *Social Science and Medicine*, *67*(3), 358–367. https://doi.org/10.1016/j.socscimed.2008.03.022

Popa-Wyatt, M., & Wyatt, J. L. (2018). Slurs, roles and power. *Philosophical Studies*, *175*(11), 2879–2906. https://doi.org/10.1007/s11098-017-0986-2

Prislin, R., & Crano, W. D. (2012). A history of social influence research. In A. W. Kruglanski & W. Stroebe (Eds.), *Handbook of the History of Social Psychology* (pp. 321–339). Psychology Press. https://doi.org/10.4324/9780203808498-24

Rahman, J. (2012). The N Word: Its History and Use in the African American Community. *Journal of English Linguistics*, *40*(2), 137–171. https://doi.org/10.1177/0075424211414807

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* Cambridge University Press. https://doi.org/10.2307/2655372

Sigelman, L., Tuch, S. A., & Martin, J. K. (2005). What's in a name?: Preference for "black" versus "African-American" among Americans of African descent. *Public Opinion Quarterly*, *69*(3), 429–438. https://doi.org/10.1093/poq/nfi026

Stangor, C., & Crandall, C. S. (2000). Threat and the social construction of stigma. *The Social Psychology of Stigma*, 62–87. https://psycnet.apa.org/record/2000-05051-003

Stephan, W. G., Ybarra, O., & Morrison, K. R. (2009). Intergroup threat theory. In T. D. Nelson (Ed.), *Handbook of prejudice* (pp. 43–59). Psychology Press. https://psycnet.apa.org/record/2008-09974-003